Audio Engineering Society

Convention Paper

Presented at the 118th Convention
2005 May 28–31 Barcelona, Spain



# A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychoacoustically Motivated Detection Functions

Nick Collins[1]

[1] *Centre for Music and Science, Faculty of Music, University of Cambridge, 11 West Road, Cambridge, CB3 9DP, UK*

Correspondence should be addressed to Nick Collins (nc272@cam.ac.uk)

**ABSTRACT**

Whilst many onset detection algorithms for musical events in audio signals have been proposed, comparative studies of their efficacy for segmentation tasks are much rarer. This paper follows the lead of Bello et al. 04, using the same hand marked test database as a benchmark for comparison. That previous paper did not include in the comparison a psychoacoustically motivated algorithm originally proposed by Klapuri in 1999, an oversight which is corrected herein with respect to a number of variants of that model. Primary test domains are formed of non-pitched percussive (NPP) and pitched non-percussive (PNP) sound events. 16 detection functions are investigated, including a number of novel and recently published models. Different detection functions are seen to perform well in each case, with substantially worse onset detection overall for the PNP case. It is contended that the NPP case is effectively solved by fast intensity change discrimination processes, but that stable pitch cues may provide a better tactic for the latter.
Keywords: Onset Detection, Detection Functions, Peak Picking, Audio Analysis

## 1. INTRODUCTION

With many possible algorithms for the detection of musical events in an audio signal now published[1, 10, 15, 12] , research questions are turning to the comparative evaluation of such methods[14, 4, 1]. This article seeks to extend the results and review of Bello et al.[1] to explore the potential of psychoacoustically motivated models such as those of Klapuri[12] and Jehan[9]. These onset detection methods can be related to psychoacoustic models of loudness[21, 17, 20, 18].

An issue immediately arises as to the application area. For some applications it may be desirous to seek a close correspondence with the decisions of an experienced human auditor at a concert as music unfolds. This would be the appropriate case for a perceptually motivated segmentation geared to the event classifications of a musical idiom, for computer assisted improvisation with sensitive audio processing. On the other hand, there are applications where the aim is that of reverse engineering, from a given audio signal, all distinct sound producing events. In this situation, the resolution of a human listener's segmentation of events could potentially be exceeded by a computer algorithm, for instance, in marking all strikes of a snare roll. For such cases, it would seem most appropriate to take the benchmark as being the non real-time mark-up of segments in an audio editor program, by a human expert[1]. For evaluation, this can be made a known quantity in a generative procedure for producing test audio; Duxbury et al.[4] utilise MIDI piano renders, where the exact onset time is known.

Subjectivity is a factor in the former situation, for multiple interpretations (possibly as a result of attentional mechanisms) are available to human auditors. This should not provide too much controversy though for monophonic signals where any overlap (due to decaying resonation of an instrument body or reverberation, for instance) is negligible compared to the signal power of a new event.

The physical onset of a sound is separated from the perceptual onset[6]. Especially for slow attacks on stringed instruments, the start of sound output does not necessarily match the moment our attention registers an attack. Such effects may be related to processes of temporal integration in the auditory system[5]. This issue will be avoided herein by considering the physical onset alone as the target. Reaction time to that onset may vary between algorithms, and the nature of a signal will, of course, provide an important factor for consideration.

The case of polyphonic audio is more contentious yet, for here there are competing streams: some events may be promoted at the expense of others. Potentially, there is a stronger subjective element in the choice of important events amongst the more substantially overlapping aggregate. For this reason, complex audio mixes are not considered in this article. In practical applications the onset detection algorithms discussed below may be applied with the proviso that they may not deal comfortably with near simultaneous events with distinct spectral signatures. A simple solution might see onset detectors restricted to certain filter bands.

Onset detection algorithms are frequently split into two components: the detection function, a signal representing the changing state of a musical signal, typically at a lower sampling rate, and a second stage of peak picking within the detection function to find onset times[1]. There may be detection functions at multiple frequency bands and at multiple rates which are recombined in some special way in peak picking[12, 4]. Those detection functions treated In this article are separable in a straight forward way from the final peak picking stage. In the comparison experiments, following the lead of the initial study of Bello et al. [1], the different detection functions are computed, with onsets output from an adaptive peak picking stage common to all functions.

To set the scene for the experiments the next

few sections will introduce some of the detection functions to be compared (section 2), and in particular models of onset detection inspired by psychoacoustics (section 2.1). A novel detection function following the essential idea of Klapuri is outlined in more detail in 2.2, before peak pickers are briefly mentioned in 2.3. The evaluation strategy for onset detection algorithms is discussed in 3; the comparison experiments follow. As a result of the qualifications about polyphonic audio above, the initial experiment (section 4) shall deal with the least contentious case of non-pitched percussive (NPP) events, linking this to the results of Bello et al.[1]. A second experiment (section 5) considers pitched non-percussive (PNP) events. Both experiments are discussed in terms of the successful detection functions and possible explanations for their success. The paper is completed by a summary including a brief mention of possible application areas.

## 2. ONSET DETECTION METHODS

It is helpful to define a few of the detection functions that will be encountered. The detection functions in this paper can almost all be expressed as causal operations on FFT bin values. $|X_n(k)|$ is the magnitude of the $k^{th}$ bin for the $n^{th}$ frame of spectral data.

The Queen Mary University of London (QMUL henceforth) signal processing group have proposed a number of onset detection methods[1, 3] which are defined clearly in their papers and are used without alteration from their original definitions herein. QMUL researchers kindly made their code available for testing purposes. This paper treats the phase deviation, being a measure of instantaneous frequency agreement over frames, a more general complex domain onset detection method which acts on the complex numbers rather than just the phases, and the spectral difference, an energy comparison over successive FFT frames.

Other author's detection functions have been reimplemented for this work and this section makes explicit which definitions have been taken. Masri and Bateman[16] define the high frequency content (HFC) as a weighted sum of spectral powers:

$$HFC(n) = \sum_{k=2}^{k=N/2} |X_n(k)|^2 k \qquad (1)$$

and calculate a detection function from considering a ratio of the HFC over consecutive frames (where the denominator is a minimum of 1).

$$DF(n) = \frac{HFC(n)}{HFC(n-1)} \frac{HFC(n)}{\sum_{k=2}^{k=N/2+1} |X_n(k)|^2} \qquad (2)$$

Jensen and Andersen[11] rewrite equation (1) with a squared weighting and sum over magnitudes, not powers.

$$HFC2(n) = \sum_{k=1}^{k=N/2} |X_n(k)| k^2 \qquad (3)$$

They take the (linear) first order difference to form the detection function:

$$DF(n) = HFC2(n) - HFC2(n-1) \qquad (4)$$

Many variants are possible that utilise various exponents and combine the bands before or after taking differences or ratios over frames.

### 2.1. Psychoacoustically Motivated Models

Anssi Klapuri[12] propounds the difference of the log spectral power in bands as a more psychoacoustically relevant feature related to the discrimination of intensity. This relative difference function can be viewed as an approximate differential of loudness (ignoring spectral and temporal masking effects on the excitation summands). Klapuri originally proposed an onset detection model combining detection in multiple bands where the salience of onsets is rated by a loudness summation based on the Moore, Glasberg and Baer loudness model[17].

His most recent onset detection scheme generalises the logarithmic compression, using the same analysis frontend as a recent beat induction model[13]. Because spectral change is the target quantity, negative differences are ignored. Steven Hainsworth has presented an equivalent formulation in the context of spotting harmonic content change, using a 4096 point FFT with a restriction of contributing bands to those in the range 30Hz-5kHz[7].

$$d_n(k) = \log_2(\frac{|X_n(k)|}{|X_{n-1}(k)|}) \tag{5}$$

$$DF(n) = \sum_{k=\alpha}^{\beta} max(d_n(k), 0) \tag{6}$$

where $\alpha$ and $\beta$ define lower and upper limits for a particular subset of bands.

Further schemes in this vein may take advantage of existing psychoacoustic models of loudness of greater complexity[21, 17]. The detection function may be formed from the direct output of a loudness model, or a first order difference of one to enhance change detection. A paper by Timoney et al. [20] describes implementations of various psychoacoustic loudness models in MATLAB.

Tristan Jehan[9] forms an event detection function by taking power in Bark bands and applying a spectral masking correction based on spreading functions familiar from the perceptual coding of audio[18], and post masking with half cosine convolution. His applications are in event sensitive segmentation.

Jensen [10] has suggested a detection function inspired from the speech recognition literature which he names the perceptual spectral flux. He rates this above his earlier high frequency content derived model (equation (3)).

$$PSF(n) = \sum_{k=1}^{k=N/2} W\left(|X_n(k)|^3 - |X_{n-1}(k)|^3\right) \tag{7}$$

In implementation, the top 100 phon equal loudness contour from [8] weights the different bands.

This author has experimented with the weighting of powers in ERB scale bands using equal loudness contours. Detection functions are created by the first order difference of the summation of intensities as an approximation of rate of change of loudness, or by a sum of changes similar to equation (6). As an example of how such a feature is engineered in practise, this particular model is described in detail in the next section.

In terms of the two roles for onset detection mentioned in the introduction, whilst perceptual models may abet musical event detection in the manner of a human observer, they may not necessarily give the best solution to match the discovery of transient sound events. However, comparison of such detection functions to others put forward in the literature may provide some interesting results.

### 2.2. A Detection Function Based on Equal Loudness Contours

For 44100 KHz sampling rate audio at 16 bit resolution, a 1024 point FFT with hop size of 512 and Hanning window is taken.

Calibration is a critical issue. As Painter and Spanias suggest[18, page 455], the reference level for the decibel scale can be taken as 1 bit of amplitude. This reference is of course a convenience, since both the pre-recording and playback level of the music are unknown. The equal loudness correction to powers described here is in some sense artificial since the level of the original acoustic stimulus should determine how the contours are applied, and the dynamic range of 16 bit audio is around 90dB, 30dB less than that of human hearing, and 10 dB less than the contour data set. The fit to the 2-100dB contour area must be determined. I choose to place the 1bit level at 15dB, so that the 90dB dynamic range of the audio is spread over the contours' range.

For 15dB at 1 bit amplitude $1/2^{15}$, a multiplier $\zeta$ is obtained by:

$$15 = 20\log_{10}\left(\frac{1}{2^{15}} * \zeta\right) \quad (8)$$

$$\zeta = 10^{15/20} * 2^{15} = 184268 \quad (9)$$

. The bins of the FFT can then be converted to decibels with the following formulation:

$$B_n(k) = 20\log_{10}\left(\zeta * |X_n(k)|\right) \quad (10)$$

Corrections to these decibel levels are calculated using equal loudness contour data; the author's implementation uses ISO226:2003[8]. Linear interpolation is applied where bin values fall between the contours in decibels SPL or centre frequency. Any values outside the 2 and 100dB phon curves are clamped to these curves, an assumption of below minimum field and saturation of excitation respectively. To make processing more efficient, FFT bins are combined (powers averaged) according to an ERB scale before the logarithmic decibel transform and contour correction. 40 ERB scale bands are used, from the formula in [17] where F is frequency in kHz:

$$numberofERBs = 21.4\log_{10}(4.37F + 1) \quad (11)$$

For a spectral difference function the sum of differences, as in the Klapuri/Hainsworth formula above, can be taken in a generalised form:

$$D_n(k) = C_n(k) - \frac{\sum_{m=1}^{M} C_{n-m}(k)}{M} \quad (12)$$

$$DF(n) = \sum_{k=1}^{40} max(D_n(k), 0) \quad (13)$$

Where the generalisation via parameter M promotes smoothing in the calculation. Of course, M=1 is equivalent to the earlier formula. $C_n(k)$ refers to the $k^{th}$ contour corrected ERB scale band signal at time n.

Alternatively, a loudness like summation can be followed and the signal L(n) or its first order difference forms the detection function:

$$L(n) = 10\log_{10}\left(\sum_{k=1}^{40} 10^{0.1C_n(k)}\right) \quad (14)$$

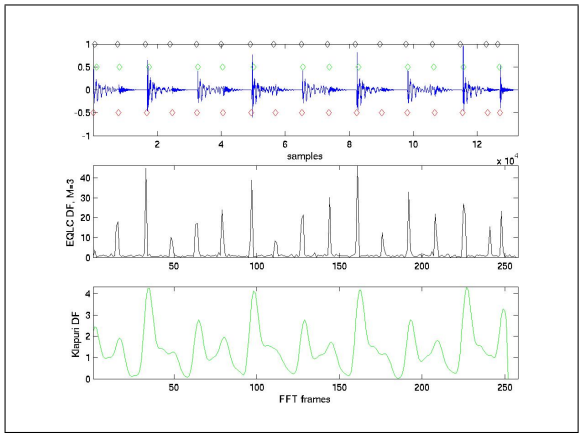


Fig. 1: Detection function (13) for M=3 compared to a recent Klapuri model

$$DF(n) = L(n) - L(n-1) \quad (15)$$

It is understood that the ISO data is gathered from experiments with continuous sinusoidal tones, and that the extension to complex sounds involves some difficulties. Nevertheless, this method provides an approximate and efficient correction for loudness for human hearing.

Figure 1 shows the detection function given by equation (13) for M=3, in comparison with the Klapuri onset detection function from [13], acting on a drum loop signal. The top subplot shows the original sample waveform with the hand marked onsets under the midline, those peak picked from the equal loudness contour detection function on the top and the Klapuri function results inbetween. The sharp definition of the former detection function when compared with the slower integrative process of Klapuri's beat induction frontend is apparent.

### 2.3. Peak Picking

Various detection functions have been described above but the second stage of peak picking remains open. Klapuri has utilised fixed thresholds as a first approximation, but some alternatives have been published including an adaptive threshold peak picker[1] and a formulation

based on a running cumulative average dubbed the note average energy[15].

QMUL's adaptive peak picker[1, IV] is taken as the common peak picking stage in comparing detection functions below. Detection function signals are normalised and low pass filtered. A median filter calculates the adaptive threshold at any centre point $n$ over points $n-8$ to $n+7$. A single parameter $\delta$ sets the requisite detection level to register an onset for the adaptively corrected detection function. In the comparison tests, $\delta$ was varied between -0.1 and 0.53 in steps of 0.01 to find the best performing peak picker setting.

## 3. EVALUATION

In the comparison of detection functions presented by Bello and colleagues[1] the test set is a database of mono 44.1KHz 16 bit soundfiles, with reference onsets marked up by hand by a single expert. This database is separated into categories of non-pitched percussive (NPP), pitched percussive (PP), pitched non-percussive (PNP) and complex mixture (MIX). For the purposes of a fair comparison, a common peak picker is used across detection functions, being an adaptive threshold picker based on a median filter as described in their paper. A spread of results are obtained for different values of the delta threshold parameter for the peak picker, which are plotted on a graph of percentage onsets detected against percentage of false positive detections as a Receiver Operating Characteristics curve.

In practise, their comparison allowed different filtering coefficients in the peak picker for different detection functions. An algorithm generated onset which fell within a lenient 50mS either side of a reference onset was allowed as a match.

Leveau et al[14] showed that the annotation task involves some variability in decisions between human experts, particularly for complex polyphonic music and instruments with slow attacks. They provide some MATLAB based annotation software and a small test set of their own which has been marked up by three users of their software, with ambiguous onsets removed (`http://www.lam.jussieu.fr/src/Membres/Leveau/SOL/SOL.htm`). Unfortunately, their data files did not work within my version of MATLAB, and their database just had five soundfiles for the PNP case. They do not provide any NPP soundfiles however, on the grounds that such soundfiles are reliably and consistently marked up; they recommend that testing with a 20mS leeway either side is appropriate for such a case.

Evaluations herein are undertaken for the NPP and PNP cases using the QMUL database of soundfiles, with a 25mS tolerance for the NPP case and 50mS for the PNP. These test sets and some MATLAB code for their detection functions and peak picker were kindly provided by the QMUL group, and allows a discussion in relation to results in their earlier paper[1]. Because the QMUL database contains on the order of 106 soundfiles in the NPP category, corresponding to 3094 onsets, it was decided to run the comparison on this larger test set. The original review paper used only 212 onsets to evaluate detections in the non-pitched percussive group. Dependency on any one soundfile is thereby much reduced, increasing confidence in the generality of results. It is difficult, however, for any detection function to score as highly as in the more reduced original study. For the PNP case, 18 soundfiles with 446 onsets formed the test set (containing examples of solo string and vocal lines), where the original review just tested over 93 onsets.

A measure of Correct Detection Ratio (CDR) was proposed in [15] to score results, and is described by the equation:

$$\mathrm{CDR} = \frac{total - missing - spurious}{total} * 100\% \tag{16}$$

This is not constrained, however, to return values between 0-100. An evaluation for-

mula from[2], originally used for the assessment of beat tracking algorithm performance, gave an alternative scoring mechanism, combining matches $m$, false positives $F^+$ (spurious) and false negatives $F^-$ (missing).

$$\text{score} = \frac{m}{m + F^- + F^+} * 100\% \qquad (17)$$

Note that the denominator includes the term for the number of onsets in the trial $n$ as $m + F^-$.

There are many published models of onset detection, and variants are easy to devise, including weighted sums of functions, and whether to take first order derivatives. There are also free parameters in some models that could potentially be optimised. This paper can only hope to explore a representative set, the specific emphasis being on psychoacoustically motivated detection functions.

It is acknowledged that the comparisons rely upon the implementation of algorithms from technical papers, which may or may not be entirely true to the original author's implementations, particularly if those author's have tweaked software to their own specific test databases. I have tried to remain as faithful as possible to the papers but cannot guarantee an absolutely fair comparison. The experiments do establish some sort of comparative baseline however against which any improved implementations can be tested.

## 4. FIRST COMPARISON- NPP

In the first experiment on the NPP test set, 16 detection functions were compared with respect to the detection of 3094 onsets. The trials were run in MATLAB using a combination of the original QMUL test code for the QMUL detection functions and the standard adaptive peak picker second stage, and the author's own implementations of the alternative models. A close comparability to the Bello et al. review paper was thereby maintained. The different detection functions are named according to the descriptions in [1] where possible; that review paper also gives full definitions for the peak picker itself.

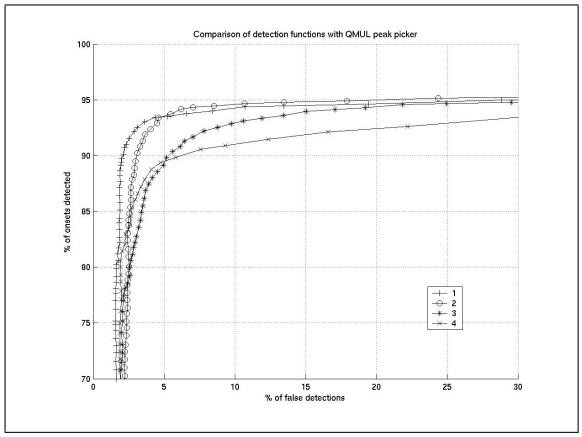


Fig. 2: NPP: Comparison of detection functions 1-4

For each detection function, 64 values of parameter $\delta$ (-0.1 to 0.53 in steps of 0.01) for the adaptive peak picker were explored. Plotting onsets detected against false positives for different values of the delta parameter draws out a Receiver Operating Characteristics (ROC) curve.

In the competition were three of the QMUL detection functions, some variants of the HFC detection function, and various psychoacoustically motivated models. Table 1 shows results, and provides links to the equations for the detection functions where given above; the detection functions will be referred to as DF1 to DF16 as indicated in the table. 1OD stands for 1st order difference. DF7 was tested because the QMUL group had (perhaps mistakenly) been using this alternative definition of Masri's HFC. For DF9, the maximum power was calculated in the time domain within windows of 1024 samples with step size of 512. ROC plots are given in figures 2, 3 and 4 for all the detection functions to show the variation of the onset detector's performance with peak picker parameter $\delta$.

| detection function | score (eqn 17) | CDR | Onsets | False Positives | best $\delta$ |
|---|---|---|---|---|---|
| 1. eqn (13), M=3, contour | 89.5 | 83.9 | 93.4 | 4.4 | 0.1 |
| 2. eqn (13), M=2, no contour | 89.3 | 83.6 | 93.4 | 4.6 | 0.12 |
| 3. PSF eqn (7) Jensen[10] | 85.5 | 77.9 | 92.2 | 7.8 | 0.14 |
| 4. eqn (6) Hainsworth[7] | 85.3 | 75.7 | 89.4 | 4.8 | 0.12 |
| 5. complexsd[3] | 74.5 | 57.9 | 88.9 | 19.3 | 0.03 |
| 6. Klapuri[13] | 74 | 55.8 | 82.6 | 11.6 | 0.03 |
| 7. HFC $\sum \lvert X\rvert k$ 1OD | 74 | 56.8 | 85.3 | 15 | 0.09 |
| 8. spectral difference [1] | 73 | 54.6 | 88.5 | 21.2 | 0.03 |
| 9. log(max power) 1OD | 72.4 | 53.2 | 83.5 | 15.4 | 0.05 |
| 10. eqn (15) contour | 70.4 | 48.7 | 80.1 | 13.8 | 0.21 |
| 11. eqn (4) Jensen[11] | 69.1 | 46.5 | 81.8 | 18.4 | 0.1 |
| 12. HFC $\sum \lvert X\rvert^2 k^2$ | 64.3 | 32.4 | 83.8 | 30.4 | 0.03 |
| 13. Jehan[9] | 59.4 | 26.8 | 68.4 | 14.4 | 0.09 |
| 14. phase deviation[1] | 57.6 | 20.8 | 72.9 | 26.6 | 0.01 |
| 15. eqn (15), no contour | 54 | 14.6 | 62.2 | 15.2 | 0.31 |
| 16. eqn (2) Masri[16] | 42.2 | -12.9 | 55.2 | 30.8 | 0.01 |

Table 1: NPP test set comparison of detection functions with QMUL peak picker

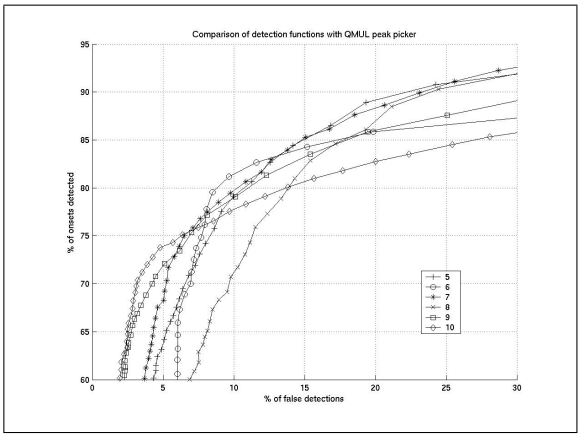


Fig. 3: NPP: Comparison of detection functions 5-10

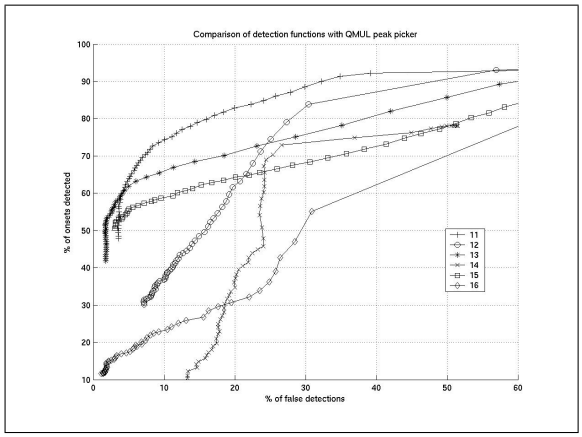


Fig. 4: NPP: Comparison of detection functions 11-16

### 4.1. Discussion

The best performing detection function is seen to be the Klapuri/Hainsworth derived function from equation (13) detailed in section 2.2. The performance of this algorithm was seen to be slightly improved by the equal loudness contour correction (the db to phon transform was not applied for DF2). The best performing values of M are given here for each case. Given the closeness of score between DF1 and DF2, it is arguable whether the contour correction is necessary, but the basic Klapuri principle of difference of logs, within ERB scale bands, shows good promise. The original Hainsworth method (difference of logs on the basic FFT bins) was also successful, and whilst results were relatively comparable for different values of $\alpha$ and $\beta$, the original choices of a range of FFT bins covering 300-5000Hz is the one whose scores are given here. A full range version performed marginally worse (score 83.9, CDR 75.3). That compression by some exponent function is a useful tactic is supported by the Jensen's high scoring DF3, and even a time domain power treated by a first order difference of logs (DF9) achieves a respectable score. Alternative version of this windowed power fared moderately worse, the bare power getting a [score,CDR] of [55,16.5], the 1OD of this [65.5,36.2], the log power without 1OD gaining [68.2,43].

In the course of compiling the table, various variants of the HFC equation were tested, including combinations of values for the exponents of the magnitudes $|X|$ and the weighting factor $k$; none outperformed DF7. Various authors have avoided Masri's original formulation of HFC as a sum over powers $|X|^2$ and instead treated the magnitudes $|X|$: this approach seems justified from the relative performance of DF7 and DF16 in the table.

Purer loudness functions modeling the excitation for a human listener perform less well at the NPP task. This is not wholly unexpected if we consider the applications again- our hearing systems are not necessarily set up to achieve good literal segmentation performance, but to parse events (Scheirer's notion of *understanding without separation*[19] is relevant here). Klapuri's beat induction frontend performs adequately at the segmentation task, but is angled more towards the discovery of useful onset information for the correlation operations required by beat induction. Jehan's masking corrected excitation function is not a great marker of percussive onsets, though it may work well at discovering the same events a human observer (rather than one working with a sound editor) would extract from an audio stream. The loudness summation form of the equal loudness contour detection function (equation (15)) is seen to perform much more poorly, though again this is probably a case of whether modeling a human response is the application. The contour corrected version definitely outperforms the bare log transform version however. A number of loudness models were trialed [20] to see if they could provide competitive performance, but in fact, most likely for the reasons given above, did not score particularly highly. DF9, the log of the windowed max power, performed better and is much more computationally efficient.

Whilst some effort was put into finding a superior performing detection function/peak picker combination, the performance of the adaptive peak picker could not be significantly bettered for the NPP test set, though it could be matched by a slightly simpler smooth-1OD-threshold peak picker (which has an advantage in requiring less delay to operate in real-time conditions). In particular, an implementation of the note average energy (NAE) peak picker[15] degraded performance; for example, DF1 fell to a score of 77.2 and CDR of 62.7 with this peak picker.

With respect to Bello et al's original study[1], the phase deviation performs significantly worse compared to the spectral difference as given in their table 1. Further, the high frequency content no longer performs so well when taken across the much expanded test set.

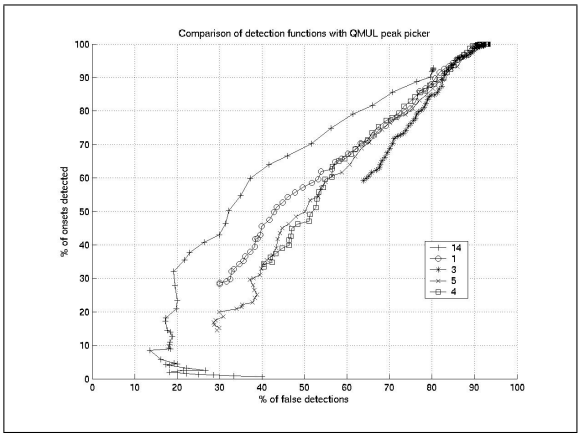


Fig. 5: PNP- Comparison of top five scoring detection functions

## 5. SECOND COMPARISON- PNP

In the second experiment, using the PNP test set, the same 16 detection functions were compared with respect to the detection of 446 onsets. These onsets were in general more widely spaced than in the NPP set, and marked out relevant pitched note events.

Table 2 gives the results, with the same set of DF1 to DF16 above, unordered this time so as to avoid confusing the reader with new labels. A single ROC plot is provided for the best four performing detection functions 5.

### 5.1. Discussion

Performance on the PNP task was markedly worse for all detection functions assessed. High rates of false positives were an inseparable side effect of matching onsets. Pronounced energy based cues for event boundaries were not obvious on examination of the sound files, where note events flowed into each other. Further, low frequency amplitude modulation was a potential detection confound.

It is proposed that the test soundfiles in the PNP case may be segmented on the basis of stability of pitch percept, a task for which the phase deviation detection function (a measure of change in instantaneous frequency) was best suited amongst those considered. Attempts to devise a pitch tracker that can mark out event boundaries by stability of cues are being investigated, though vibrato (frequency modulation) on stringed instruments is another possible tracking confound- something analogous to categorical perception should probably be built in. In general, effective performance may rely upon strategies specific to the recognition of familiar timbres and playing characteristics.

Whereas the NPP set was segmented effectively by many different detection functions as a non-linear editing task potentially superior to human listening, the PNP case is an example where the modelling of human perceptual processes must underlie effective mark-up. None of the models investigated here is a sufficient encapsulation of human segmentation by pitch cues to score as highly as the earlier comparison. Smoothing of detection functions based on energy cues was obviously insufficient to cure the problems.

## 6. CONCLUSIONS

This study has compared a number of published and original detection functions on two contrasting large test sets of hand marked audio files. The first case (NPP) was effectively solved by difference of log power functions derived from Klapuri's work[12]. Relatively simple discrimination functions in this vein performed well, with fuller psychoacoustic models of loudness less effective in application. There are differences between perceptual segmentation (finding event boundaries as a human observer would function in real-time) and physical segmentation (breaking up events as fast and as accurately as possible for digital editing purposes). This difference was further supported in PNP comparison, where a more subjective mark-up of events had taken place in the test data, most likely based on a pitch segmentation strategy and not an intensity discrimination one. All detection functions performed

| detection function | score (eqn 17) | CDR | Onsets | False Positives | best $\delta$ |
|---|---|---|---|---|---|
| 1. eqn (13), M=3, contour | 35.8 | -36.8 | 51.3 | 43.5 | 0.36 |
| 2. eqn (13), M=2, no contour | 27.6 | -49.1 | 38.8 | 40.8 | 0.35 |
| 3. PSF eqn (7) Jensen[10] | 36.1 | -86.3 | 59.2 | 63.9 | 0.53 |
| 4. eqn (6) Hainsworth[7] | 30.5 | -50 | 44.8 | 46.9 | 0.44 |
| 5. complexsd[3] | 31.1 | -46.4 | 45.1 | 44.8 | 0.28 |
| 6. Klapuri[13] | 12.5 | -80.3 | 18.6 | 48.4 | 0.09 |
| 7. HFC $\sum \lvert X\rvert k$ 1OD | 28.5 | -146.0 | 49.8 | 74.5 | 0.53 |
| 8. spectral difference [1] | 10.1 | -84.3 | 15.0 | 48.9 | 0.38 |
| 9. log(max power) 1OD | 7.6 | -94.2 | 12.1 | 60.3 | 0.41 |
| 10. eqn (15) contour | 11.9 | -80.7 | 17.49 | 47.3 | 0.48 |
| 11. eqn (4) Jensen[11] | 20.0 | -129.8 | 34.8 | 74.1 | 0.53 |
| 12. HFC $\sum \lvert X\rvert^2 k^2$ | 0.6 | -115.7 | 1.1 | 94.1 | 0.52 |
| 13. Jehan[9] | 7.4 | -85.4 | 10.1 | 35.7 | 0.36 |
| 14. phase deviation[1] | 43.6 | -15.7 | 59.9 | 37.2 | 0.08 |
| 15. eqn (15), no contour | 9.3 | -87.4 | 14.1 | 52.6 | 0.48 |
| 16. eqn (2) Masri[16] | 9.1 | -90.8 | 14.3 | 57.6 | 0.49 |

Table 2: PNP test set comparison of detection functions with QMUL peak picker

significantly worse and the most successful, the phase deviation, could be related to a measure of instantaneous frequency.

For applications, perceptual segmentation may mimic the event categorisation of human listeners, and has dividends in machine listening for musical improvisation and composition. Such signal understanding, however, is in contrast to as fast as possible onset detection for percussive transients, and requires some delay in operation, typically of the order of 200mS when modeling temporal integration processes. This processing delay may also be commensurate with note/phone event lengths and hence categorically quantised pitch tracks, where events are marked up after they have occurred, giving chance to determine their boundaries. The nature of the sound events to be detected determines the appropriate detection strategy.

## 7. ACKNOWLEDGMENTS

This research is supported by AHRB grant 2003/104481. Many thanks to my supervisor Ian Cross.

Juan Bello at QMUL provided the onset detection test suite and various members of that group were willing to discuss their onset detection research.

Joseph Timoney provided MATLAB code and Brian Glasberg and Michael Stone made C code available for their loudness model.

## 8. REFERENCES

[1] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and S. B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 2004.

[2] Simon Dixon. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1):39–58, 2001.

[3] Chris Duxbury, Juan P. Bello, Mike Davies, and Mark Sandler. Complex domain onset detection for musical signals. In *Proc. Digital Audio Effects Workshop (DAFx)*, 2003.

[4] Chris Duxbury, Juan Pablo Bello, Mark Sandler, and Mike Davies. A comparison between fixed and multiresolution analysis for onset detection in musical signals. In *Proc. Digital Audio Effects Workshop (DAFx)*, 2004.

[5] David A. Eddins and David M. Green. Temporal integration and temporal resolution. In *Hearing*, pages 207–42. Academic Press, San Diego, CA, 1997.

[6] John W. Gordon. The perceptual attack time of musical tones. *J. Acoust. Soc. Am.*, 82(1):88–105, July 1987.

[7] Stephen Hainsworth and Malcolm Macleod. Onset detection in musical audio signals. In *Proc. Int. Computer Music Conference*, pages 163–6, 2003.

[8] ISO. Acoustics: Normal equal-loudness-level contours. Technical Report ISO226:2003, International Organisation for Standardization, 2003.

[9] Tristan Jehan. Event-synchronous music analysis/synthesis. In *Proc. Digital Audio Effects Workshop (DAFx)*, Naples, Italy, October 2004.

[10] Kristoffer Jensen. Causal rhythm grouping. In *Proceedings of the 2nd International Symposium on Computer Music Modeling and Retrieval*, Esbjerg, Denmark, May 2004.

[11] Kristoffer Jensen and Tue Haste Andersen. Real-time beat estimation using feature extraction. In *Proc. Computer Music Modeling and Retrieval Symposium, Lecture Notes in Computer Science*. Springer Verlag, 2003.

[12] Anssi Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, pages 3089–92, 1999.

[13] Anssi P. Klapuri, Antti J. Eronen, and Jaakko T. Astola. Analysis of the meter of acoustic musical signals. *IEEE Trans. Speech and Audio Processing*, forthcoming, 2004.

[14] Pierre Leveau, Laurent Daudet, and Gaël Richard. Methodology and tools for the evaluation of automatic onset detection algorithms in music. In *Proc. Int. Symp. on Music Information Retrieval*, 2004.

[15] Ruolun Liu, Niall Griffth, Jaqueline Walker, and Peter Murphy. Time domain note average energy based music onset detection. In *Proceedings of the Stockholm Music Acoustics Conference*, Stockholm, Sweden, August 2003.

[16] Paul Masri and Andrew Bateman. Improved modelling of attack transients in music analysis-resynthesis. In *Proc. Int. Computer Music Conference*, 1996.

[17] Brian C. J. Moore, Brian R. Glasberg, and Thomas Baer. A model for the prediction of thresholds, loudness, and partial loudness. *J. Audio Eng. Soc.*, 45(4):224–40, April 1997.

[18] Ted Painter and Andreas Spanias. Perceptual coding of digital audio. *Proc. of the IEEE*, 88(4):451–513, 2000.

[19] Eric D. Scheirer. Towards music understanding without separation: Segmenting music with correlogram comodulation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1999.

[20] Joseph Timoney, Thomas Lysaght, Marc Schoenweisner, and Lorcán Mac Manus. Implementing loudness models in MATLAB. In *Proc. Digital Audio Effects Workshop (DAFx)*, 2004.

[21] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models (2nd Edition)*. Springer Verlag, Berlin, Germany, 1999.