A COMPARISON BETWEEN FIXED AND MULTIRESOLUTION ANALYSIS FOR ONSET DETECTION IN MUSICAL SIGNALS

Chris Duxbury, Juan Pablo Bello, Mark Sandler and Mike Davies

Centre for Digital Music
Queen Mary University of London, London, UK
{chris.duxbury|juan.bello-correa}@elec.qmul.ac.uk

ABSTRACT

A study is presented for the use of multiresolution analysis-based onset detection in the complex domain. It shows that using variable time-resolution across frequency bands generates sharper detection functions for higher bands and more accurate detection functions for lower bands. The resulting method improves the localisation of onsets on fixed-resolution schemes, by favouring the increased time precision of higher subbands during the combination of results.

1. INTRODUCTION

In [1] we showed that onset detection in the complex domain can offer significant improvements on basic energy-based onset detection methods. In this paper, we consider the effects of using a fixed resolution spectral analysis when compared with a multiresolution subband approach.

Subband schemes, such as those discussed in [2, 3], were proposed because different onsets may be stronger in different subbands, as well as the argument that spectral analysis in subbands more closely represents the non-linearities of human hearing. These previous subband based onset detection schemes were essentially energy based. As such, they were effective at selecting strong percussive transients, but were not as effective in detecting softer onsets, particularly at low frequencies. In [4], we presented a hybrid scheme which went some way to solving this issue by using energy in the upper subbands, and the spectral distance measure in lower subbands. This idea was extended in [5], where a Kullback-Lieber function was used in the lower subbands, which is similar in principle to the spectral distance measure. However, this work preceded the development of the complex detection function, which effectively combines the energy, and frequency (now measured using phase information) approaches. In previous results, we have shown how robust the complex domain detection function is to onset type. We shall now discuss the effects of extending the complex detection approach to a subband structure.

2. MULTIRESOLUTION MUSIC SIGNAL ANALYSIS

Window length selection is a key consideration when using an FFT based analysis. Good time resolution, implying a short window length, is needed for detecting fast changes within the signal. On the other hand, good frequency resolution, and therefore a long window length, is required for accurate location of sinusoidal components, particularly at low frequencies where small changes in frequency are heard by the listener.

There are two main approaches to solving this fixed-resolution problem. The first we refer to as multiresolution-in-time based on window switching, and the second multiresolution-in-frequency based on subband analysis. Naturally, resolutions in time and frequency are inversely proportional. Hence, in reality, both approaches are multiresolution in both time and frequency.

In time-varying multiresolution signal analysis, as described in [6], window switching techniques are used such that short analysis windows and hop sizes are used at transient frames whilst longer windows are used at more steady state regions. This is used in many audio coding schemes, as well as for transient preservation in sinusoidal modelling. The key problem with this approach is that the sharp change in frequency resolution at window boundaries makes the matching of sinusoidal components across frames problematical, suggesting it is impractical for many applications.

In the frequency-varying multi-resolution scheme, the frequency spectrum is split into subbands, and processing is performed independently on each. This allows shorter analysis windows to be used at higher frequencies whilst lower frequencies can still have the required frequency resolution to separate closely spaced sinusoids. The advantages of the basic principles are described in [7]. In the following we will briefly discuss a number of proposed methods for achieving multiresolution in frequency.

2.1. Multiresolution in frequency

In [8], an approach is proposed that uses multiple different length FFT analyses on the signal. The longer window analysis bins are used for low frequency component values, whilst the short window FFT bins are used for high frequencies. Complexity in an algorithm such as this increases linearly with the number of different length FFT analyses performed. In essence, this is a redundant analysis, although certain bins are ignored at the processing stage.

Another approach, proposed in [9], proposes the use of a twice-oversampled Laplacian pyramid structure, which is traditionally used for image compression. This has the advantage that the subbands are approximately alias-free. However, the oversampling required still leads to an increase in computation.

The simplest approach (as discussed in [8]) to multiresolution in frequency, is the use of a critically sampled constant-Q filterbank of quadrature mirror filters (QMF). QMF filters are pairs of filters, G_0 and G_1, given by:

$$G_1(z) = G_0(-z) \quad (1)$$

$$G_1(\omega) = G_0(\omega - \pi) \quad (2)$$

Hence, G_0 is G_1 modulated by π. If we rearrange this we get:

$$G_1(0.5\pi + \omega) = G_0(0.5\pi - \omega) \quad (3)$$

From this we can see that the filters are a mirror image about 0.5π. A constant Q filterbank is obtained by cascading several QMF pairs, and downsampling the outputs of each pair prior to analysis (see Figure 1). If re-synthesis is required, a reconstruction filterbank must be used (due to aliasing between subbands). A variant of this scheme is proposed in [10], where a bandpass constant Q filterbank is used. However, the lack of downsampling in this scheme leads to a highly redundant representation of the signal.

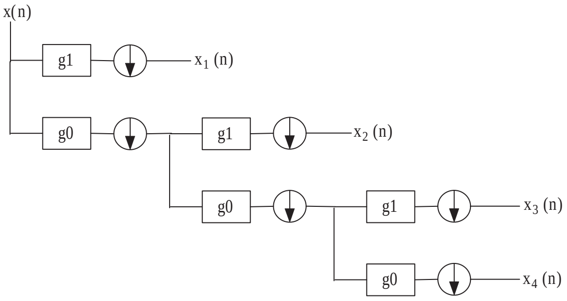

Figure 1: *Constant Q multiresolution filterbank of cascaded QMF filters.*

In this work we favour the use of a simple constant Q filterbank of perfect reconstruction QMF filters. This choice keeps the algorithm simple and efficient. However, any of the above algorithms could be applied equally well (depending on the particular application).

2.2. Subband onset detection

There are several examples in the literature of subband analysis for onset detection.

In [2] an scheme is proposed based on individual energy analysis of subbands. This aims to mimic onset detection by human hearing. After re-scaling, the signal is filtered in a filterbank with 21 overlapping "nearly-critical" bands. The lowest 3 filters are octave spaced, whilst the remaining 18 are third-octave spaced. For each subband, onsets are characterised using the log-amplitude difference detection function. After peak-picking on each subband, the detected onsets are recombined using a psychoacoustic model, based on the loudness model of [11].

In [3], a similar scheme is proposed using the linearly spaced subbands of the FFT. During a short window around the time of an attack, a triangular shape is fitted to the energy profile of each frequency channel using a least-squares method. A detection function is derived from the maximum peak of the triangle, and its mean amplitude before and after the attack. The individual results are then aggregated across frequencies and along an *uncertainty interval* in time.

Both of the above onset schemes are still essentially energy based schemes, and as such, suffer from poor detection of softer note transitions, such as those of bowed strings. Further to this, they tend to under-detect softer low frequency notes, due to the slower increase in energies found at these frequencies. In the following Sections we discuss an alternative method for onset detection in the complex domain and its implementation as a subband scheme.

3. ONSET DETECTION IN THE COMPLEX DOMAIN

There are a number of reasons that justify combining phase and energy information for onset detection: while energy-based approaches favour strong percussive onsets and are more reliable when using high-frequency information (where energy changes are not masked by the overall energy of a sound), phase-based approaches emphasize soft, "tonal" onsets and are more robust in the lower end of the spectrum, where those tonal changes occur. In [1] a fully combined approach in the complex domain is presented and successfully tested. We will briefly explain its theory in the following.

For locally steady state regions in audio signals, we can assume that frequency and amplitude values remain approximately constant. Therefore it is clear that by inspecting changes in either frequency and amplitude, onset transients can be located. Furthermore, by predicting values in the complex domain, the effect of both variables can be considered.

Let us assume that, in its polar form, the target value for the k^{th} bin of the STFT of a signal $s(n)$ is given by:

$$\hat{S}_k(m) = \hat{R}_k(m)e^{j\hat{\phi}_k(m)} \tag{4}$$

where the target amplitude $\hat{R}_k(m)$ corresponds to the magnitude of the previous frame $|S_k(m-1)|$, and the target phase $\hat{\phi}_k(m)$ can be calculated as the sum of the previous phase and the phase difference between preceding frames:

$$\hat{\phi}_k(m) = \text{princarg}[2\tilde{\varphi}_k(m-1) - \tilde{\varphi}_k(m-2)] \tag{5}$$

We may then consider the measured value in the complex domain from the STFT $S_k(m) = R_k(m)e^{\phi_k(m)}$, where R_k and ϕ_k are the magnitude and phase of the current STFT frame. By rotating target and current phasors, such that $\hat{S}_k(m)$ is mapped onto the real axis, and by measuring the Euclidean distance between them, we can quantify the stationarity for the k^{th} bin as:

$$\begin{aligned} \Gamma_k(m) \quad &= \quad \big\{\hat{R}_k(m)^2 + R_k(m)^2 - \\ &\qquad 2\hat{R}_k(m)R_k(m)cos\big(d_{\varphi k}(m)\big)\big\}^{\frac{1}{2}} \end{aligned} \tag{6}$$

where

$$d_{\varphi k}(m) = \text{princarg}[\tilde{\varphi}_k(m) - 2\tilde{\varphi}_k(m-1) + \tilde{\varphi}_k(m-2)] \tag{7}$$

is the phase deviation between target and observed phase values in a given frame. Summing these stationarity measures across all k, we can construct a frame-by-frame detection function as:

$$\eta(m) = \sum_{k=1}^{K} \Gamma_k(m) \tag{8}$$

In [1] it was shown that $\eta(m)$ is an adequate detection function showing sharp peaks at points of low stationarity while returning a smoother profile than those obtained with energy or phase-based methods (consistently outperformed on the experimental results). Also, results showed robustness for a wide range of musical signals.

Subband Number	Frequency Band (kHz)
1	11.05 to 22.1
2	5.525 to 11.05
3	2.7625 to 5.525
4	0 to 2.7625

Table 1: *Complex Detection Frequency Bands*

4. COMPLEX-DOMAIN DETECTION IN SUBBANDS

A simple 3-level constant-Q filterbank of QMF filters is used, splitting the signal into four frequency subbands for individual onset detection analysis (as shown in Table 1).

Figure 2 shows the complex detection function calculated separately for each subband (lower subbands area at the bottom of the figure). In this case, the hop size and window size are set such that they have the same resolution across all subbands. It can be seen that a greater number of onsets are detected in the lowest subband. This is because some of the softer notes have weak energy in the higher subbands. In this case, thresholding of the lowest subband would achieve high accuracy in onsets detected.

However, the localisation of detected onsets is often as important as the number of detections. We know that at higher frequencies, energy changes are sharper, particularly at hard (percussive) onsets. Therefore, the higher subbands are useful for onset localisation, suggesting the use of a shorter analysis window. This is shown in Figure 3, where the multiresolution scheme is such that the hop and window sizes are fixed in terms of samples, but the downsampling in the filterbank leads to a multiresolution analysis in time.

From these examples we can conclude that the lower subbands have a tendency to be robust to noise, and therefore produce accurate results in terms of numbers of detected onsets. Unfortunately, the long windows needed for analysis at these frequencies lead to poor resolution, and therefore poor onset localisation. Conversely, higher subbands tend to be more prone to noise and miss-detection of very low notes, but produce better localised results. In the following we will try use this observations to appropriately combine information from all subbands.

5. COMBINING SUBBAND INFORMATION

The proposed scheme for combining onset information across subbands, is a re-implementation of our previous work in the area [4].

One possible solution to the combination problem, is to generate an overall detection function as the sum of all subband functions, such that:

$$\eta_{all}(m) = \eta_1(m) + \eta_2(t) + \eta_3(t) + \eta_4(t) \qquad (9)$$

where the sub-index corresponds to subband numbers as shown in Table 1. Applying peak-picking to this function does not solve the problem of weaker onsets remaining undetected, even though they may be strong within a certain frequency band (due to masking by stronger onsets in the overall function). An alternative is to peak-pick onsets on each subband, and then combine the results.

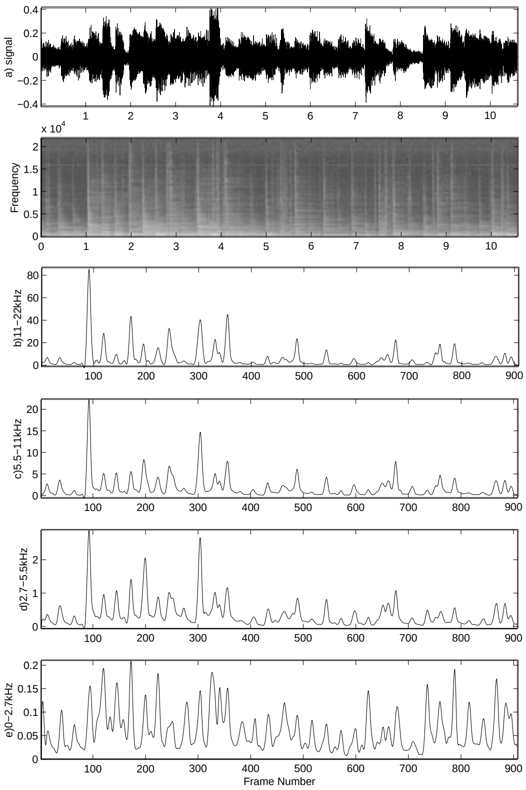

Figure 2: *Fixed-resolution window-length subband complex detection functions. The overlap is 50% and the hop size corresponds to 11.6 ms (256 samples at 44100 kHz sampling frequency) for all subbands*

5.1. Peak-picking

For the detection of onsets, first our algorithm normalises and DC-removes the obtained subband detection functions. This is to facilitate the thresholding of the detection functions by emphasising the characteristics of similar peaks and making them more uniform, not only within a signal, but between a number of different signals.

Then, the median filter is used to obtain an adaptive threshold curve $\delta_t(m)$. This is calculated as the weighted median of an H-length section of the detection function around the corresponding frame, such that:

$$\delta_t(m) = \delta + \mathrm{median}\,\eta(k_m), k_m \in [m - \frac{H}{2}, m + \frac{H}{2}] \qquad (10)$$

δ is a constant value with a large influence on the number of good and false onset detections. [12] demonstrated the effectiveness of the median filter for the thresholding of peaks in detection functions generated from music.

Finally, local maxima above the calculated threshold are simply selected as the onsets of a particular subband.

5.2. Combining detected onsets

After peak-picking, each output of the subband scheme produces a list of onset positions and magnitudes. In many cases, onsets

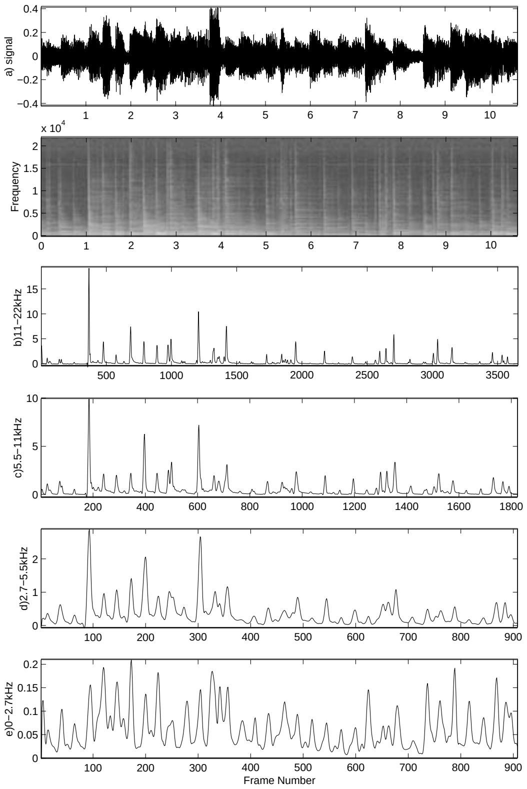

Figure 3: *Multiresolution window-length subband complex detection functions. The hop size corresponds, in (e) and (d), to 11.6 ms (256 samples at 44100 kHz sampling frequency) for the lower subbands - (e) and (d) -, 5.8 ms (128 samples at 44100) for the second subband (c), and 2.9 ms (64 samples at 44100) for the highest subband (b). Window lengths are such that in all subbands a 50% overlap is used.*

are present in multiple subbands. Let us define $P(m)$ as a signal-length vector containing one at onsets and zeros elsewhere, such that:

$$P_{all}(m) = P_{\eta_1}(m) + P_{\eta_2}(m) + P_{\eta_3}(m) + P_{\eta_4}(m) \qquad (11)$$

where the sub-index refers to the subband detection functions or to the combination of them "all". When onsets appear in more than one subband, they will tend to be slightly misaligned (due to the multiresolution nature of the scheme). Thus, a short window k_m of 50ms is taken, such that only one onset is allowed per window. Selection of onsets within this window is done by weighting the output of each subband:

$$P_{all}(k_m) = \alpha P_{\eta_1}(k_m) + \beta P_{\eta_2}(k_m) + \gamma P_{\eta_3}(k_m) + P_{\eta_4}(k_m) \qquad (12)$$

such that $k_m \in [m-25ms, m+25ms]$. α, β, and γ are weighting terms such that:

$$\alpha > \beta > \gamma > 1 \qquad (13)$$

This approach is adopted rather than position averaging to favour the use of onsets from higher-frequency subbands (thus improving time resolution), while increasing the number of correct detections with information from the lower subbands (used when detections are not obtained in the higher bands). A second reason for using a weighting scheme of this nature is that it may be tuned so that only 'hard' (percussive) or 'soft' (tonal) onsets are selected.

6. COMPLEX SUBBAND DETECTION RESULTS

In order to test the effectiveness of the subband based complex domain detection compared to the fixed resolution scheme, detection results were tested on recordings of a MIDI-controlled acoustic piano. This test data was preferred over hand-labelled music files (as used in [1]), as localisation of onsets is at the core of our discussion, and the hand marked data proved to lack the accuracy for this. Let us define a measure of onset detection accuracy as given by:

$$Accuracy(\%) = \frac{Total}{Total - Missed - Bad} * 100 \qquad (14)$$

where $Total$ represents the total number of correct onsets, $Missed$ is the number of missed detections, and Bad is the number of bad detections. We measured the accuracy of onsets detected as the correct detection analysis (CDA) window varies in size. This window is the acceptable distance between the measured onset, and the true onset. Hence, as the window is short, only well localised onsets will be labelled as good detections. As the window increases, less accurate onsets will be accepted as good detections.

From the results in Figures 4 and 5, it can be seen that the multiresolution scheme outperforms the fixed window complex domain onset detection for short CDA window lengths. This is in line with the hypothesis that a multiresolution scheme will improve onset localisation. However, for longer CDA window lengths, the multiresolution scheme under-performs. This is due to a higher number of bad detections, as illustrated by the right side of Figure 6. This increase in bad detections is brought about by the noise introduced by the upper subband detection functions, as can be seen in the upper subbands of Figure 3.

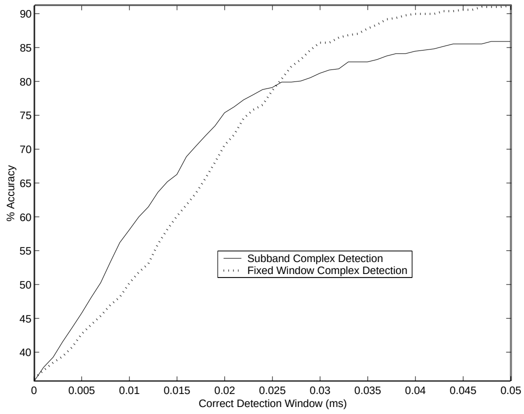

Figure 4: *Percentage accuracy comparison of localisation of detected onsets complex detection and multiresolution complex detection.*

7. CONCLUSIONS

Fixed-scale complex-domain onset detection is a robust and efficient method that successfully incorporates energy and phase in-

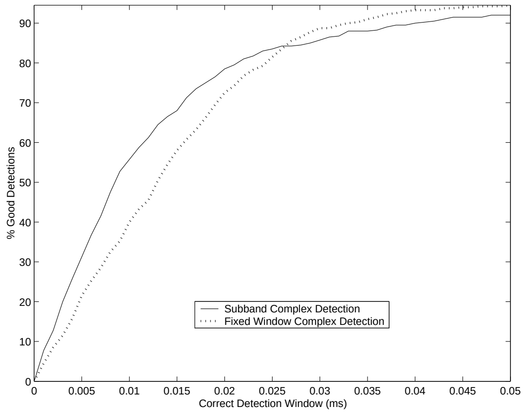

Figure 5: *Percentage good detection comparison of localisation of detected onsets complex detection and multiresolution complex detection.*

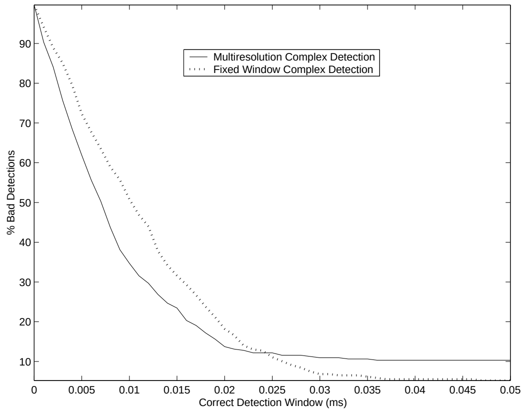

Figure 6: *Percentage bad detection comparison of localisation of detected onsets complex detection and multiresolution complex detection.*

formation, therefore allowing the detection of “hard” and “soft” onsets for a wide range of music signals. On the other hand, previous methods have shown the advantages of using multiresolution analysis for onset detection, given the heterogeneous behaviour of onsets in different subbands.

In this paper we present a simple subband scheme for complex-domain onset detection. The approach uses a constant-Q filterbank of QMF filters followed by complex-domain onset detection on each of the resulting bands. Experiments show that by using variable time resolution across frequency bands we can improve on the localisation of onsets at higher subbands while relying on the high detection rates returned by lower subbands.

The algorithm for the combination of subband onset detections, favours the use of the onset times from higher subbands. Results show that for short comparison windows (CDAs), the proposed scheme improves detections on the fixed resolution complex-domain onset detection. Conversely, for longer comparison windows, the fixed-resolution approach shows more robustness, mostly due to the over-detections introduced by high-frequency noise into the subband scheme. Therefore, the multiresolution method proves an alternative for a range of music applications where the precision of the detection is of great importance.

8. REFERENCES

[1] J. P. Bello, C. Duxbury, M. Davies, and M. Sandler, “On the use of phase and energy for musical onset detection in the complex domain,” *IEEE Signal Processing Letters*, vol. 11, no. 6, June 2004.

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

[3] Xavier Rodet and Florent Jaillet, “Detection and modeling of fast attack transients,” in *Proceedings of the International Computer Music Conference*, 2001.

[4] C. Duxbury, M. Sandler, and M. Davies, “A hybrid approach to musical note onset detection,” in *Proc. Digital Audio Effects Conf. (DAFX,’02)*, 2002.

[5] S. Hainsworth and M. Macleod, “Beat tracking with particle filtering,” in *Proceedings of the IEEE Workshop on Applications of Signal Processing*, 2003.

[6] M. M. Goodwin, *Adaptive Signal Models: Theory, Algorithms, and Audio Applications*, Phd thesis, University of California, Berkeley, 1997.

[7] P. Masri and A. Bateman, “Improved modelling of attack transients in music analysis-resynthesis,” in *Proc. International Computer Music Conference (ICMC’96)*, 1996.

[8] D. Anderson, “Speech analysis and coding using a multiresolution sinusoidal transform,” in *Int. Conf. Acoustics, Speech and Signal Processing*, 1996.

[9] T. S. Verma S. N. Levine and J. O. Smith, “Multiresolution sinusoidal modeling for wideband audio with modifications,” in *Int. Conf. Acoustics, Speech and Sig. Proc., (ICASSP)*, 1998.

[10] D. Ellis and B. Vercoe, “A wavelet-based sinusodial model of sound for auditory signal separation,” in *Proc. Int Computer Music Conference*, 1991.

[11] C. J. Plack B. J. C. Moore, B. R. Glasberg and A. K. Biswas, “The shape of the ear’s temporal window,” *J. Acoustic Soc. Am.*, vol. 83, pp. 1102–1116, 1988.

[12] I. Kauppinen, “Methods for detecting impulsive noise in speech and audio signals,” in *Proc. DSP’2002*, 2002.