Onset Detection in Musical Audio Signals

Stephen Hainsworth*† and Malcolm Macleod ‡
†Cambridge University Engineering Department, Cambridge, CB2 1PZ, UK
‡QinetiQ, Malvern, WR14 3PS, UK
swh21@eng.cam.ac.uk, m.macleod@signal.qinetiq.com

Abstract

This paper presents work on changepoint detection in musical audio signals, focusing on the case where there are note changes with low associated energy variation. Several methods are described and results of the best are presented.

1 Introduction

The issue of onset detection in sound signals is one which has applications in speech processing, audio manipulation (coding, for instance) as well as the plethora of applications in music processing. Until recently, only the issue of abrupt changes in signal power, usually called transient detection, has been considered. This detects either consonants in speech or sharp onsets such as percussion sounds in music. It does not, however, address the change of harmonic content, without an associated strong power transient, which is a common occurrence in some genres (e.g. choral music, string quartets, solo flute). We will address this problem.

Specifically, the aim is to extract change points in musical signals which are equivalent to the human perception of a new note (or notes) starting. This may be accompanied by a change in amplitude but the current paper concentrates on the case in which there is not an associated power change. The eventual application in mind is beat tracking in musical signals and hence importance is assigned to minimising false detections rather than maximising true detections. If one were considering signal manipulation, the reverse would probably be true. [1]

Work to date in the area of musical change detection mainly centres around energy change. The typical approach (e.g. (Dixon 2001)) is to take the sound signal, create a power evolution function, E_n (usually via smoothing) and find onsets in this, often by searching for peaks in the difference function, $D_n = E_n - E_{n-1}$. This method will obviously miss changepoints where there is little energy change. More recently, several approaches have been proposed which address this and look for harmonic change. Duxbury (Duxbury, Sandler, and Davies 2002) described a system which used energy change detection in three bands above 1.1kHz but a method which looked for spectral change below this. His spectral change measure was a modified Euclidean distance measure between successive STFT frames on a bin by bin basis, where only positive power change was considered in order to ignore decreases in energy. Davy (Davy and Godsill 2002) proposed a method based upon bilinear time-frequency representations and support vector machine classification of resulting spectra to classify the novelty of a new time slice. It is, however, unclear how much of the change detection is due to harmonic change and how much is due to amplitude variation. Abdallah (Abdallah and Plumbley 2003) has also recently investigated the problem and used a combination of independent component analysis (ICA) to evaluate the "surprise" measure of a new frame given the recent past and HMMs to detect the onsets from this measure. Klapuri (Klapuri 1998) also investigated note onset detection and developed a method for refining the location of an onset by examining the relative energy difference function in various frequency bands.

2 Methods Using Harmonic Models

In this section, several methods for change detection will be discussed and reasons presented why they have not been pursued in the current study.

2.1 Traditional Change Detection

The area of change detection is well researched and there are many books on the subject (e.g. (Gustafsson 2000)). Several important assumptions are however made in these methods: firstly, it is assumed that the signal is generated by one of a small number of models and that these models are well defined. Secondly, it is generally assumed that the process is then corrupted by noise which is not correlated with the process. From these assumptions, the usual approach is to use a likelihood function to evaluate the probability of the data being generated by each of the models and to use Bayes theorem to incorporate any prior knowledge, such as minimum length

* Supported by the George and Lillian Schiff Foundation

[1] For the overall beat-tracking code, the power transient onsets are also considered but the method of extracting them is not novel and is hence not reported here.

of time between change points or the probability of changing from one model to another.

The reason why it is not practical to use these methods is because in music, the assumptions given above do not hold - there is an infinity of models to consider because there is an infinite number of combinations of notes that could be played simultaneously. To solve the problem would require an extra stage of model selection in order to determine how many sinusoidal components are present at any one time and define hyperparameters for this. Punskaya (Punskaya, Andrieu, Doucet, and Fitzgerald 2002) has investigated this for speech signals, using MCMC estimation methods, but this proves to be extremely computationally expensive.

2.2 Harmonic detection approaches

As we are interested in the changing of sinusoidal components, a sensible approach might be to perform sinusoidal detection and evaluate the change in these measures. Serra (Serra 1997) gives a simple sinusoidal detector while a slightly more rigorous and accurate but more computationally demanding algorithm is used here (Macleod 1998).

Once the parameters of detected sinusoids have been estimated for all frames, it remains to find whether there is any change from one frame to the next. A list of sinusoidal frequencies and their associated amplitudes is not a helpful representation here. Two methods for further processing were investigated: the first involved producing a pseudo spectrum for each time frame by convolving the delta functions representing the sinusoidal frequencies with a suitable window function (e.g. a Gaussian) and then applying distance measures to these, as described below. The motivation for this is the hope that these pseudo spectra will be cleaned-up, denoised versions of the original spectrogram frames with only the harmonic information present.

The second method involved tracking the sinusoidal components over time to form harmonic tracks. A multiple hypothesis tracking algorithm was developed from the theory presented by Blackman (Blackman and Popoli 1999). Detection was performed by examining where tracks started and clustering together tracks which had close onsets.

It turned out that neither method was very satisfactory. This is because the sinusoidal detection function breaks down at the crucial point of interest - the start of notes. Here, the parameters are often masked by transient noise and take a few tens of milliseconds to attain reasonably stable characteristics. During this period, which is the time we are attempting to estimate, the sinusoidal detector will often not produce a detection and hence subsequent processing will be fundamentally flawed. This can be seen in Fig. 1 at all change points for a choral piece with very little transient power; the effect is even more pronounced when there are significant power transients in the signal.

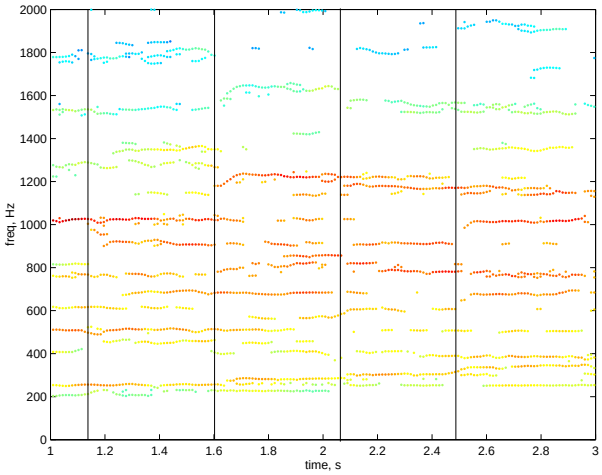


Figure 1: Plot of the output from the harmonic extraction algorithm with hand labelled change points superimposed. The example used is shown in Fig 2a, as a spectrogram.

3 Fourier Transform approaches

Attention is thus turned to approaches based on more complete signal representations. The signal, when plotted as a spectrogram (eg Fig. 2a) produces a time-frequency representation in which the eye can pick out sinusoidal components and transients. It is therefore reasonable to assume that a computer can also be programmed to detect the changepoints from this representation.

The first step is to produce a measure of novelty, or difference for the signal. Various measures exist to evaluate the "distance" between two vectors, the simplest of these being the Euclidean distance

$$D_{EUC}(n) = \sum_{k=0}^{\frac{N}{2}-1} (|X_n(k)| - |X_{n-1}(k)|)^2 \qquad (1)$$

where X_n is the STFT of the n^{th} frame of data, length N bins. Others such as the Kullback-Liebler (K-L) distance

$$D_{K-L}(n) = \sum_{k=0}^{\frac{N}{2}-1} |X_n(k)| \log_2 \left(\frac{|X_n(k)|}{|X_{n-1}(k)|} \right) \qquad (2)$$

are more complex. Foote (Foote and Uchihashi 2001) also proposes a measure specifically for spectral differencing

$$D_{Foote}(n) = 1 - \frac{\langle |\mathbf{X}_n|, |\mathbf{X}_{n-1}| \rangle}{||\mathbf{X}_n|| \times ||\mathbf{X}_{n-1}||} \qquad (3)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner or scalar product and the denominator allows the measure to be normalised to the range [0,1].

The aim is to detect increases in energy for given bins of the spectrogram while ignoring decreases of energy which are

associated with the end of notes[2]. The K-L measure is the one which accentuates the change of amplitude most, but weights it with the bin amplitude of the second frame. The following measure, denoted the MKL or modified Kullback-Liebler distance, is therefore proposed which simply reflects the rate of positive amplitude evolution between two successive frames.

$$d(k) = \log_2 \left(\frac{|X_n(k)|}{|X_{n-1}(k)|} \right) \quad (4)$$

$$D_{MKL}(n) = \sum_{k=0, d(k)>0}^{N/2-1} d(k) \quad (5)$$

Results for these distance measures are shown in Fig. 2b. From this, one can see that the MKL measure is performing significantly better than the other measures, with clear peaks where there are spectral changes. This example is an excerpt from Byrd's 4 Part Mass and has almost no amplitude change, except when voice parts add or leave the texture. However, the measures in Fig. 2b are still not very clear and reliably detecting onsets would be fairly hard. This is due to the fact that on a frame by frame basis, there is a fair degree of variation. To overcome this, multiple frames were histogrammed together. A bilinearly decreasing weighting function was applied to give preference to data immediately around the point under consideration. Also, a low FFT hop rate (1/8 of a frame, corresponding to an 87.5% overlap) was used to decrease the frame by frame variation and increase the time resolution.

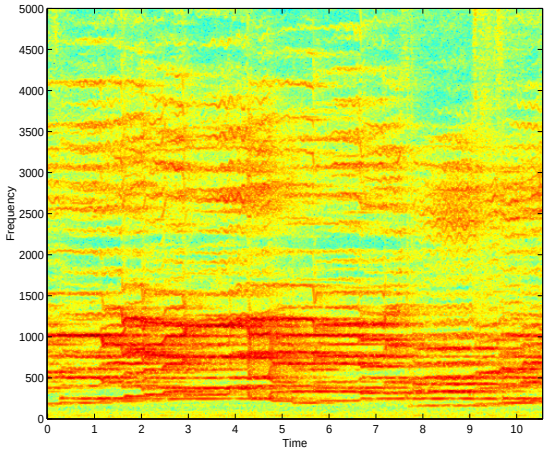
The resulting spectral distance measures can be seen in Fig. 2c. Once again, the MKL measure gives the clearest results and now, peaks are clearly delineated with little noise.

3.1 Detection of changes

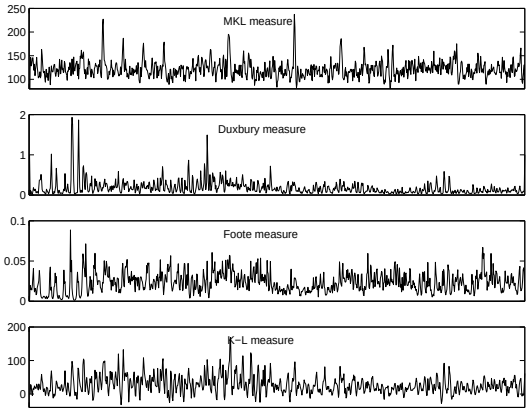
Producing a measure which reflects the change in the signal is only half the problem. Reliably detecting peaks in the measure is a tricky task in itself. A simple two stage strategy was adopted which seemed to work well. Firstly, the measure was smoothed via convolution with a Hanning window of a suitable length. Then, between each crossing of this smoothed function and the mean value of the function, the highest peak in the unsmoothed function was picked as a changepoint. This has the advantage of producing a number of changepoints which is roughly proportional to the complexity of the data.

However, this did tend to over-fit so a second stage was introduced where the spectral difference between two sequential inter-changepoint regions was compared using Foote's measure as in Eqn. 3. Foote's measure was chosen because it is normalised to the range [0,1] and hence allows easy thresholding. If any two sequential regions produced a difference measure of less than 0.1, the changepoint was discarded.

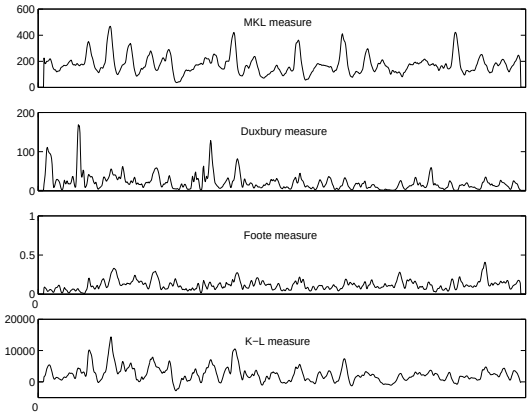
[2]On the whole, the end of notes is not a reliable indication of beat due to sustain, reverb or effects such as staccato.



(a) Raw Spectrogram



(b) Results without frame histogramming



(c) Results with frame histogramming

Figure 2: Plot of the various measures listed in section 3 with and without histogramming for the example shown in a) . The MKL, K-L and Foote measures are described in the text, while the Duxbury measure is a Euclidean distance measure which ignores negative energy change.

4 Results

The eventual algorithm also used band-wise processing to extract harmonic onsets from different bands: 30-300Hz, 300Hz-1kHz and 1-5kHz were chosen to represent bass, low-mid and high-mid frequencies. Above 5kHz, some instruments maintain clear harmonic power (e.g. trumpets) but in many examples, the harmonic information at these frequencies is unclear and hence ignored. Details of the algorithm are as follows: an STFT frame length of 4096 samples was used with histogramming performed over +/- 10 frames; the smoothing kernel for detection was of length 20 frames; and processing of long samples was broken into individual 10s samples (mainly for computational purposes).

Figure 3 shows the results of the complete algorithm for the choral example used throughout the paper. Only one true change (at 3.9s) out of 20 is completely missed in all bands, while there are no false detections. It should be noted that the missed change was discarded by the 2nd detection stage in this example.

This level of accuracy is maintained over examples from a wide variety of styles with no extra adaption of the algorithm. Often, these examples contain percussive onsets which would be adequately detected using existing techniques. Though this algorithm will sometimes detect these, if the underlying harmonic structure does not change, often the second stage of the detection algorithm will discard these purely percussive changes. This means that the algorithm is a true detector of harmonic change. Further examples can be found at `http://www-sigproc.eng.cam.ac.uk/~swh21`.

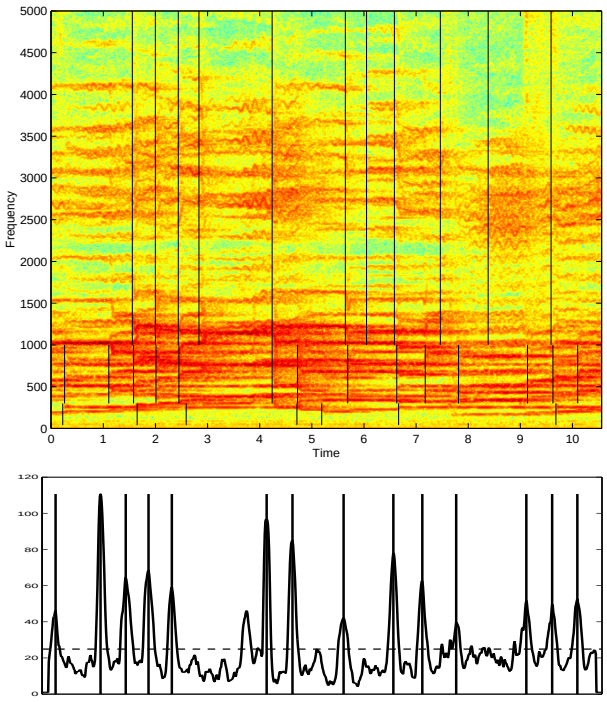


Figure 3: The output from the complete algorithm with changepoints shown as vertical lines. The lower plot shows the MKL measure for the middle frequency band with associated changepoints shown.

5 Conclusions

This paper presents a number of methods for detecting musical changepoints which are mainly harmonic in nature as opposed to power transients. The proposed technique performs very well with low computational cost and algorithmic complexity. The output of this detector is intended to be used alongside data from a transient locator in a beat detection algorithm detailed in (Hainsworth and Macloed 2003).

References

Abdallah, S. and M. Plumbley (2003). Unsupervised onset detection: a probabilistic approach using ICA and a hidden Markov classifier. In *Proc. Cambridge Music Processing Colloquium*.

Blackman, S. and R. Popoli (1999). *Design and Analysis of Modern Tracking Systems*. Artech House.

Davy, M. and S. Godsill (2002). Detection of abrupt spectral changes using support vector machines: an application to audio signal segmentation. In *Proc. ICASSP*.

Dixon, S. (2001). Automatic extraction of tempo and beat from expressive performaces. *J. NMR 30*(1), 39–58.

Duxbury, C., M. Sandler, and M. Davies (2002). A hybrid approach to musical note detection. In *Proc. Digital Audio Effects Workshop (DAFx)*, Hamburg, pp. 33–8.

Foote, J. and S. Uchihashi (2001). The beat spectrum: a new approach to rhythm analysis. In *Proc. Int. Conf. on Multimedia and Expo (ICME)*.

Gustafsson, F. (2000). *Adaptive Filtering and Change Detection*. Wiley.

Hainsworth, S. and M. Macloed (2003). Beat tracking with particle filtering algorithms. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, NY.

Klapuri, A. (1998). Automatic transcription of music. Master's thesis, Audio Research Group, University of Tampere, Finland.

Macleod, M. (1998). High resolution nearly-ML estimation of sinusoids in noise using a fast frequency domain approach. In *Proc. EUSIPCO*, pp. 1849–52.

Punskaya, E., C. Andrieu, A. Doucet, and W. Fitzgerald (2002, March). Bayesian curve fitting using MCMC with application to signal segmentation. *IEEE Trans. Signal Processing 50*(3), 747–68.

Serra, X. (1997). Musical sound modeling with sinusoids plus noise. In C. Roads, S. Pope, A. Picialli, and G. De Poli (Eds.), *Musical Signal Processing*, Chapter 3, pp. 91–122. Swets and Zeitlinger.