

ADAPTIVE WHITENING FOR IMPROVED REAL-TIME AUDIO ONSET DETECTION

Dan Stowell, Mark Plumbley
Centre for Digital Music
Queen Mary, University of London

ABSTRACT

We describe a new method for preprocessing STFT phase-vocoder frames for improved performance in real-time onset detection, which we term “adaptive whitening”. The procedure involves normalising the magnitude of each bin according to a recent maximum value for that bin, with the aim of allowing each bin to achieve a similar dynamic range over time, which helps to mitigate against the influence of spectral roll-off and strongly-varying dynamics. Adaptive whitening requires no training, is relatively lightweight to compute, and can run in real-time. Yet it can improve onset detector performance by more than ten percentage points (peak F-measure) in some cases, and improves the performance of most of the onset detectors tested.

We present results demonstrating that adaptive whitening can significantly improve the performance of various STFT-based onset detection functions, including functions based on the power, spectral flux, phase deviation, and complex deviation measures. Our results find the process to be especially beneficial for certain types of audio signal (e.g. complex mixtures such as pop music).

1. INTRODUCTION

Audio onset detection (or audio segmentation), the process of detecting the beginning of “events” such as musical notes in an audio stream, is a fundamental component of machine listening [8]. It is very often used in tasks such as tempo estimation, beat tracking, and automatic transcription. It can be useful in offline processes (e.g. to analyse a database of music recordings) but also in online and real-time processes such as interactive music systems [8] or on-the-fly melodic transcription [5]. Not all onset detectors are suitable for real-time applications, because they may require look-ahead or a significant amount of computation.

1.1. Current onset detectors

In recent years a number of approaches to onset detection have been investigated (see [3] for a useful overview). The procedure typically involves a data reduction step, converting the audio rate signal to an *onset detection function* (ODF) which is at a much lower sampling rate, followed by a step to identify onsets in this ODF.

Time-domain methods for producing the ODF are possible but most current techniques convert the signal to the frequency- or complex-domain. This is typically achieved using a phase-vocoder in which the audio signal is converted to a stream of STFT frames. The subsampled ODF is then produced, which may for example amount to one numerical value per STFT frame. Onsets can then be selected via identifiable phenomena in the ODF signal, such as exceeding a threshold.

Within this basic recipe many variants have been studied. The algorithm for producing the ODF is an important consideration: within the literature many are discussed, common algorithms including spectral power, spectral flux, high-frequency content (HFC), phase deviation, weighted phase deviation, and complex deviation. See for example [3] and [10] for discussion of these methods. [4, section 2.3] makes use of the Kullback-Leibler divergence statistic, and variations thereupon, as an ODF. [14] trains a neural net to produce a suitable ODF from frequency-domain information.

Variations on the basic recipe exist. For example, the signal can be broken down into separate frequency bands, and each subband treated separately, the results later being combined into a single onset-detection output [11].

There are a variety of approaches to onset detection that do not fit the template just described. For example, instead of STFT, filter banks [13], wavelet decomposition [9], probabilistic modelling [2], or pitch tracking [7] can be the basis of the analysis.

1.2. Onset selection

Selecting onsets from the ODF signal can be accomplished in different ways. (Most ODFs are designed such that they reach a high value when onsets occur; this behaviour will be assumed for this discussion.)

Perhaps the most basic selection method is to use a threshold-based trigger, meaning that an onset is said to be detected whenever the ODF signal transitions from below to above a given threshold. This method is computationally very simple and can easily run in real time. It can work well with some ODFs, but may work less well if the ODF signal is prone to variations in range (e.g. if the heights of its peaks vary according to the general intensity of the music). As discussed by Brossier [4, section 2.4], DC removal and normalisation can at least partly mitigate against such problems, although these are not causal pro-

cesses and so are unsuitable for real-time use; for real-time application Brossier instead uses a “dynamic threshold” computed using the median (and optionally also the mean) calculated over a short buffer around the current frame. This can be equivalently stated as subtracting a proportion of the median from the signal and then using a static threshold. An alternative way to regularise the ODF (and also compatible with real-time use) is low-pass filtering, but compared against the median-based method this can be more vulnerable to undue influence by outliers [3].

Various studies use peak-picking to select onsets (e.g. [3], [6], [10]). Note that peak-picking algorithms used in onset detectors typically include some thresholding, to avoid detecting minor peaks which might otherwise cause spurious detections.

One consideration for real-time use is that peak-picking requires a delay of at least one STFT frame, because in order to identify a peak we must determine that the values both before and after the peak are lower. In the setup to be used in this paper, for example, the separation between frames is 5.8 ms. A delay of 5.8 ms may seem minor, especially considering the limits on human ability to resolve events in time, since events separated by less than around 30 ms are generally perceived as simultaneous by humans [18]. However, for real-time applications such as automatic accompaniment or resynthesis we may wish the total system to be able to respond with perceived simultaneity. In this context, taking into account delays introduced by analogue-to-digital and digital-to-analogue conversion and other processes that may occur upstream or downstream, avoiding a delay of 5.8 ms may be desirable in order to allow the total system response to be fast enough for perceived simultaneity.

Furthermore, imperceptibly short delay times can affect the sound quality even if they do not affect the perception of simultaneity, such as introducing a “flam” effect to the sound. (“Flam” is a drummer’s technique of playing a quiet grace note immediately before a note, with only a very small separation between the two. Reference [20] provides MIDI-based examples; the separations in the flam sounds range from about 17 to 45 ms.) The reduction of delay is therefore an important aim for real-time onset detectors.

An alternative to threshold-based triggering or peak-picking is pattern-recognition. For example, Kapanci and Pfeiffer [12] use Support Vector Machines trained to identify onsets on a multi-resolution ODF. That study was not targeted towards real-time application, although machine learning approaches could in theory be used for rapid real-time onset selection.

1.3. Choosing an onset detector

The proliferation of onset detection methods arises because no one method can be shown to be generally optimal. Onset detection performance can differ according to the type of audio data used for testing: for example, whether the music used is largely percussive or not, and

the degree of polyphony. Some onset detectors are specialised to a particular domain (e.g. pitch-tracker based onset detectors, typically intended for use with pitched non-percussive music), and some (e.g. neural networks) can be specialised by training them with a particular class of input signal. Even for these specialised onset detectors, the results can fall some way short of ideal [7].

As well as the domain of application, practical considerations may also influence the relative appeal of the different onset detection methods. Methods involving machine learning techniques will typically require training on an appropriate annotated dataset, which may be impractical. For real-time applications, the onset detector must be causal (it cannot look into the future) and must be relatively efficient, so that the onset decisions can be produced quickly enough to be useful for downstream processes, and in some applications quickly enough for the output to be perceived by a human listener as simultaneous, as discussed above. These demands exclude some approaches, and they also tend to favour vocoder-based approaches since these can be implemented using the very efficient FFT algorithm [4, section 2.2].

Our research is primarily targeted towards real-time applications, so we are interested in developments that may improve STFT-based onset detectors while remaining causal and relatively efficient.

1.4. Problems with onset detectors

We cannot expect onset detectors ever to reach 100% accuracy relative to the “ground truth” annotations provided by a human observer, because the ground truth is not completely stable: annotations produced by different human observers typically exhibit some variation [15]. However, there is still considerable room for improvement, especially in the domains that have thus far proved “difficult” for onset detectors, such as music containing strong variations in dynamics, or polyphonic/polytimbral mixtures. The differing performances of onset detectors against different types of audio signal in the recent MIREX audio onset detection contest [1] are testament to this.

As an example of strongly-varying dynamics, consider the first fifteen seconds of Beethoven’s Fifth Symphony (Figure 1). A sequence of very strong notes is followed immediately by a series of very quiet notes. The spectral power measurement (Figure 1(c)) exhibits peaks of dramatically-varying size. This poses a challenge for purely power-based onset detectors, but also to others where magnitude is a factor, including spectral flux and complex deviation. It is even relevant to phase-based ODFs: to attain good results “phase deviation” ODFs often include magnitude information either explicitly [10] or implicitly via magnitude-thresholding of bins ([3], as discussed in [10]).

Normalising each STFT frame to a fixed total magnitude may benefit some ODFs, but clearly not the power ODF, where the resulting output would be a fixed value. Klapuri [13] pursues an alternative route starting from psychoacoustic principles, using the first difference of the log

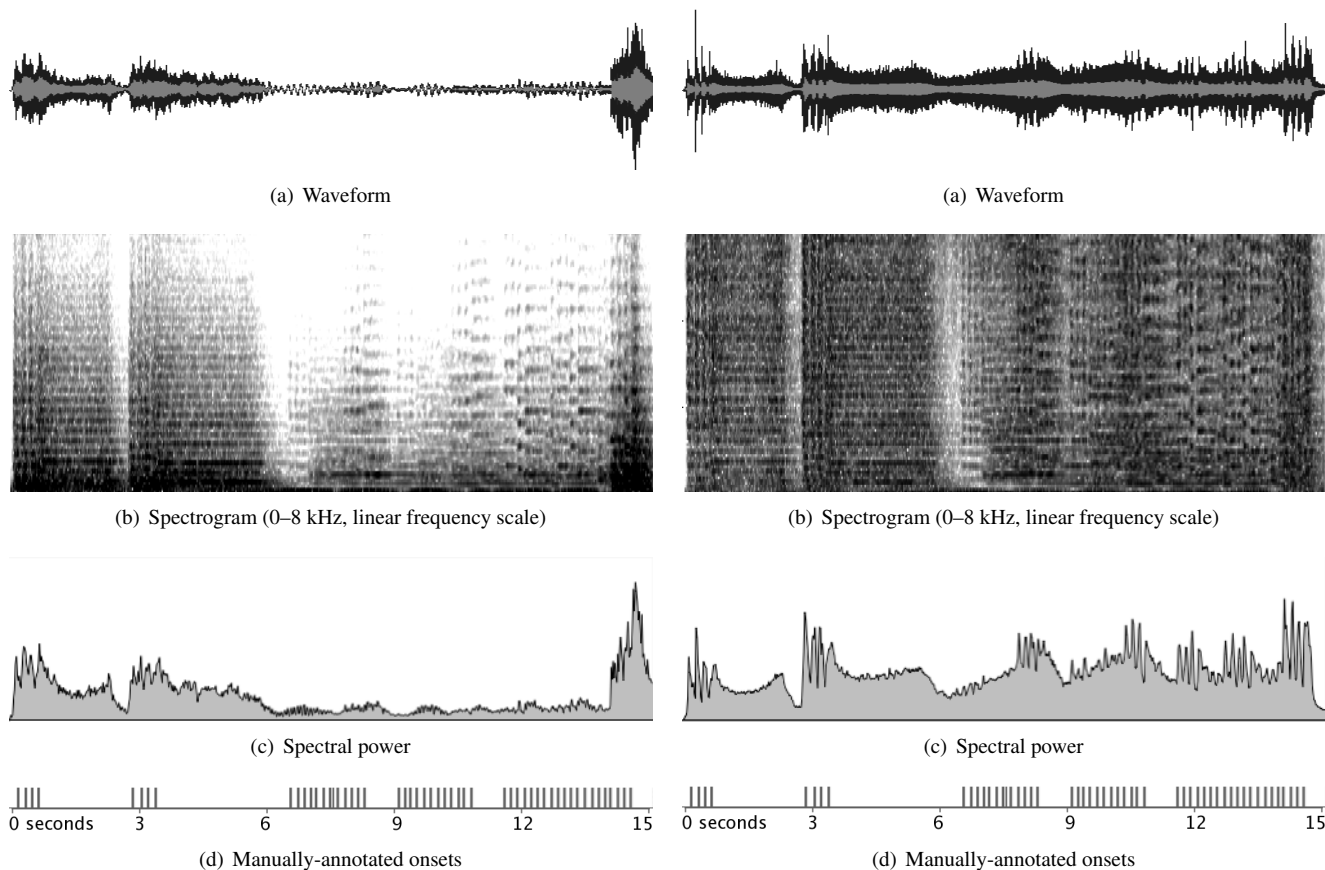


Figure 1. The first fifteen seconds of Beethoven’s Fifth Symphony

Figure 2. The first fifteen seconds of Beethoven’s Fifth Symphony, after adaptive whitening has been applied. Parameters: relaxation time 250 s, floor coefficient 10^{-4} .

amplitude (equivalent to the log of the amplitude ratio between successive frames), which produces an ODF which is independent of overall amplitude scaling.

Besides temporal variability, there is also variability across frequency bands. Musical signals typically exhibit a spectral “roll-off”, with the peak magnitudes reaching lower and lower values towards the higher frequency bands. This means that the lower bands can often “drown out” the higher bands in ODFs such as power or spectral flux, contributing much more strongly to the variation in the ODF signal; information that may be present in the higher bands may be neglected as a result.

The HFC measure has been found useful in onset detection (e.g. [6]), and one reason for this may be that it approximately compensates for the roll-off by emphasising the contributions of higher-frequency bins more strongly than those of lower bins. However, the re-weighting performed by HFC is not derived empirically from the typical spectral slope of music signals but is simply linear in frequency. An alternative way to compensate could be to apply a pre-emphasis filter to the audio signal before processing.

Both HFC and pre-emphasis are relatively naïve in that they re-weight the spectrum in a fixed manner, independent of the characteristics of the individual signal under consideration.

We are interested in finding a procedure that can re-weight the spectrum, as for HFC and pre-emphasis, but in a data-dependent fashion. We propose this can lead to improved accuracy over those techniques, and ideally also in a way which can mitigate against the problems of dramatic dynamic variations which can confound some onset detectors.

2. ADAPTIVE WHITENING

In an initial experiment we analysed an entire music file with a phase vocoder and measured the highest magnitude that occurred in each frequency bin, across the whole file. We call this the *peak spectral profile* (PSP). The PSP could then be used to whiten the signal before onset detection: by dividing each STFT bin’s magnitude by the all-time peak for that bin, we ensure that each frequency band reaches the same maximum over the duration of the recording.

This whitening process did improve onset detection in some cases, but it worked less well in recordings with strong variation in content over time (whether timbral variation or dynamic variation). This can be seen in the Beethoven example previously described. The spectral peaks in the loud section would be the ones used for rescal-

ing, but these would not be particularly appropriate for the quiet section. It was also a non-causal process, which made it unsuitable for real-time onset detection scenarios.

Three modifications to this procedure led to the algorithm considered here. Firstly, for causal operation, only the peak magnitudes thus far through the audio file are considered: the PSP is not derived in a separate pass but at the same time as processing the audio for onset detection. This means that the spectral re-weighting can only depend on peak values in the past relative to the frame under consideration. Secondly, in order to cope better with audio in which the dynamics evolve over time, the PSP values decay exponentially over time, meaning that past peak values are gradually “forgotten”. Thirdly, to prevent PSP values falling so low that noise (such as quantisation noise) became overly amplified, a *floor* parameter was added to the algorithm, below which none of the PSP values would be allowed to fall.

The resulting iterative algorithm can be expressed as

$$P_{n,k} = \begin{cases} \max(|S_{n,k}|, r, mP_{n-1,k}) & \text{if } n > 0, \\ \max(|S_{n,k}|, r) & \text{otherwise.} \end{cases} \quad (1)$$

$$S_{n,k} \leftarrow \frac{S_{n,k}}{P_{n,k}} \quad (2)$$

for $n \geq 0$, where m is the memory coefficient, r the floor parameter, and $S_{n,k}$ the value of the complex STFT at frame index n and frequency bin index k .

For convenience, the memory coefficient can be calculated from the STFT frame rate and the desired 60 dB relaxation time, i.e. the amount of time it would take a peak to decay by 60 dB. Values for this *relaxation time* will be referred to in the remainder of this paper, rather than the corresponding memory coefficient.

The algorithm is an adaptive process that aims to whiten the signal in the sense of bringing the magnitude of each frequency band into a similar dynamic range; hence the term *adaptive whitening*.

Figure 2 shows the same signal as in Figure 1 but after adaptive whitening has been applied. The waveform (produced here by using the inverse FFT to convert the whitened signal back to the time domain) exhibits a normalisation effect similar in appearance to the effect of a dynamic range compressor. The spectrogram shows the effect of the adaptive whitening in the frequency domain: as well as reducing the difference between the loud and the quiet notes, the spectral roll-off is quite thoroughly removed from each note, leaving a signal with a generally flat spectral profile. The effect also shows itself in the spectral power measurement, with the quiet and the loud notes exhibiting very similar patterns of activity. Note that this change in the spectral power measurement is not necessarily the same as would be seen with an amplitude-normalisation process such as dynamic range compression, because the relative contribution of the different frequency bands is modified, as well as the overall amplitude.

The adaptive whitening algorithm is relatively lightweight to compute on current hardware, since it uses only

floating-point comparison, addition, multiplication and division operations. These require only a single instruction on typical processors, as compared against trigonometric or logarithmic operations [19, chapter 2]. Its memory requirements are also small, the main requirement being the PSP, an array of floating-point values of the same size as the number of STFT bins.

3. EVALUATION

In order to test the effect of adaptive whitening on real-time onset detection, we investigated an onset detection algorithm in which adaptive whitening could be present or absent, and different algorithms (e.g. power, HFC, spectral flux) could be inserted to produce the ODF. The algorithm is illustrated schematically in Figure 3.

Other than the choice of ODF and the presence/absence of adaptive whitening, all other parts of the onset detection process were held fixed. Audio signals were input as 44.1 kHz mono signals, and FFT was performed using blocks of 512 samples with a 50% overlap between blocks. For onset selection, we first subtracted the median of the ODF values of the previous 11 frames from the current value (our 11-frame window was found in earlier experiments to yield good results), then applied a simple threshold-based triggering.

Initial experiments were performed to explore the effect of parameters r and m (results not shown). For good performance, typical values of r ranged from 10^{-6} to 0.2, and typical values of the relaxation time ranged from 22 to 446 seconds. (The scale of r is related to the overall signal amplitude, which in our experiments reached a maximum value of 1.) The optimum settings exhibited some variation according to the type of music signal.

We then wished to determine the performance of adaptive whitening without varying the adaptive whitening parameters. In the remaining experiments, we therefore held the parameters fixed at 0.1 for r and 25.6 s for the relaxation time, which were generally good settings but not optimised towards any particular type of music signal.

The dataset used for evaluation was the set of audio files and hand-annotated onsets used in the MIREX 2005 and 2006 onset detection competitions. This dataset is divided into four main types of audio:

- 30 recordings of monophonic pitched instruments (including voice)
- 30 recordings of solo percussion instruments (including drum kits)
- 10 recordings of polyphonic pitched instruments
- 15 recordings of “complex mixtures” including pop music, classical music and world music

Audio file lengths are between 2 and 36 seconds. Each recording is accompanied by annotations produced by 3 different listeners, except for the complex mixtures which

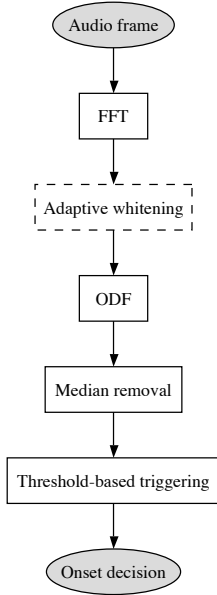


Figure 3. Block diagram of the onset detector used. The dashed line indicates that adaptive whitening may be enabled or disabled; “ODF” can be any of the ODFs used.

have 5 different sets of annotations each. In total there are 9,333 annotated onsets.

Evaluation of the onset detector performance was similar to that in other studies (e.g. [3], [10]). A detected onset was counted as “correct” if it fell less than 50 ms ahead of or behind a hand-annotated onset. Each of the 3 or 5 annotations for each audio file was used separately to assess the onset detector performance, and then the results summed together to determine overall performance. Total numbers of correct detections (true positives), false detections (false positives), and failures to detect annotated onsets (false negatives) were recorded.

For each permutation of the onset detector, totals were recorded at a range of threshold settings, and also at ranges of parameter settings where appropriate. The totals were then used to calculate the Precision and Recall statistics:

$$P = \frac{O_{TP}}{O_{exp}} \quad (3)$$

$$R = \frac{O_{TP}}{O_{orig}} \quad (4)$$

where O_{orig} is the number of onsets hand-annotated, O_{TP} is the number of correct detections, and O_{exp} is the number of onsets produced by the onset detector.

The F-measure, used to summarise Precision/Recall information [21, chapter 5], was also calculated:

$$F = \frac{2PR}{P + R} \quad (5)$$

The F-measure, and plots of Precision against Recall for different threshold settings, were used in the following experiments to evaluate the performance of onset detectors with and without adaptive whitening.

To determine the usefulness of adaptive whitening as a preprocessing step we tested the onset detector using various ODFs, with adaptive whitening both enabled and disabled. The following ODFs were tested:

1. Power (Pow)
2. Phase deviation (PD)
3. Weighted phase deviation (WPD)
4. Rectified spectral flux (SF)
5. Complex deviation (CD)
6. Rectified complex deviation (RCD)
7. High-frequency content (HFC)
8. Modified Kullback-Leibler divergence (MKL)

The power ODF simply uses the instantaneous power measurement, as discussed earlier. For definitions of ODFs 2–6 see [10]; for the HFC see [16]. The Modified Kullback-Leibler divergence algorithm is as presented by Brossier [4, section 2.3]:

$$MKL_n = \sum_{k=0}^K \log \left(1 + \frac{|S_{n,k}|}{|S_{n-1,k}| + \epsilon} \right) \quad (6)$$

where $S_{n,k}$ represents the STFT bin at frequency k for frame n . ϵ is added to the calculation to avoid large variations when very low energy levels are encountered.

For the ϵ coefficient we tested a number of values; the results here use $\epsilon = 0.01$ which we found to work much better than the lower values such as 10^{-6} recommended by Brossier [4]. (As with r above, the scaling of ϵ is related to the overall signal amplitude.)

Our results are summarised in Table 1 and Table 2. The peak F-measures are given, determined over a range of threshold settings in each case. Figure 5 illustrates the Precision/Recall curves for the power ODF, with and without whitening. Note that our results on non-whitened ODFs will differ from other published results (e.g. [1]) because of differences in the experimental design, in particular the method of triggering and the STFT frame size.

4. DISCUSSION

The results show a general pattern, which is that adaptive whitening improves the performance of most ODFs, especially for the complex mixture and monophonic pitched datasets – which tend to be the datasets on which the non-whitened ODFs generally achieve the poorest performance. Adaptive whitening most strongly benefits the power, WPD, SF, CD, and RCD ODFs. The HFC performance is slightly improved except with the polyphonic pitched dataset, while the PD performance is slightly improved or slightly degraded, depending on dataset.

Of the non-whitened ODFs, the MKL produces some of the strongest results, and in fact performs consistently

Dataset	ODF	% F-measure	% Precision	% Recall
Complex mixture	Pow	70.2	73.4	67.2
	PD	75.0	77.4	72.8
	WPD	65.5	63.8	67.4
	SF	67.2	81.1	57.3
	CD	72.4	77.2	68.1
	RCD	64.0	61.9	66.2
	HFC	73.5	77.3	70.0
	MKL	78.5	80.9	76.2
Solo drums	Pow	92.8	92.6	93.0
	PD	90.8	91.0	90.7
	WPD	92.2	92.1	92.4
	SF	89.9	94.7	85.5
	CD	93.7	93.9	93.4
	RCD	90.8	88.7	93.0
	HFC	87.7	79.9	97.3
	MKL	94.7	95.0	94.6
Monophonic pitched	Pow	53.7	54.2	53.2
	PD	58.6	66.1	52.6
	WPD	51.7	57.0	47.3
	SF	56.7	57.2	56.2
	CD	57.9	63.1	53.5
	RCD	47.6	50.2	45.2
	HFC	56.8	58.3	55.4
	MKL	64.6	64.9	64.3
Polyphonic pitched	Pow	87.6	90.3	85.1
	PD	81.7	89.6	75.0
	WPD	76.8	72.8	81.3
	SF	75.4	85.2	67.6
	CD	88.6	89.7	87.5
	RCD	68.9	59.8	81.3
	HFC	84.4	85.8	83.0
	MKL	82.4	88.4	77.2

Table 1. Performance at peak F-measure for the onset detector using various ODFs, without adaptive whitening.

worse when combined with adaptive whitening. This is an interesting exception, and it would be useful in future research to explore the reasons why adaptive whitening does not improve the MKL ODF. It may for example be the case that the adaptive whitening process alters the STFT frames in such a way as to reduce the relevance of the bin-by-bin ratio measure taken between successive frames.

The RCD and WPD functions were put forward by Dixon [10] as an improvement to the CD and PD functions respectively. In our results, we do not find RCD to improve on CD. Curiously, although WPD does not improve on PD in our non-whitened tests, it does perform better than PD on all four datasets when adaptive whitening is enabled. This is likely to be connected with the way magnitude information is used in the two ODFs: in PD, magnitude-thresholding is used for inclusion of a given STFT bin in the calculation, and it may be that adaptive whitening reduces the appropriateness of this bin-by-bin magnitude threshold.

The generally best-performing onset detectors in our experiments were the adaptively-whitened CD, the adaptively-whitened power, the adaptively-whitened RCD and the non-whitened MKL. The strong performance of the

Dataset	ODF	% F-measure	% Precision	% Recall
Complex mixture	Pow	79.3	82.8	76.0
	PD	74.9	75.4	74.3
	WPD	80.6	85.3	76.4
	SF	73.7	81.7	67.1
	CD	81.7	82.0	81.4
	RCD	80.3	83.5	77.2
	HFC	76.1	82.5	70.6
	MKL	75.8	77.6	74.0
Solo drums	Pow	93.3	96.0	90.7
	PD	91.8	94.9	88.9
	WPD	93.3	96.5	90.2
	SF	92.5	94.0	91.1
	CD	93.5	96.0	91.1
	RCD	93.5	95.1	92.0
	HFC	90.3	88.0	92.6
	MKL	90.8	90.9	90.7
Monophonic pitched	Pow	66.3	70.8	62.5
	PD	60.6	63.9	57.7
	WPD	60.8	63.0	58.7
	SF	63.3	64.0	62.6
	CD	67.3	73.1	62.4
	RCD	62.0	63.4	60.6
	HFC	62.7	64.9	60.6
	MKL	55.4	51.7	59.6
Polyphonic pitched	Pow	88.0	90.7	85.4
	PD	79.1	80.6	77.6
	WPD	83.8	84.3	83.4
	SF	79.4	85.8	73.9
	CD	87.3	88.1	86.4
	RCD	84.6	83.8	85.5
	HFC	77.4	73.9	81.2
	MKL	62.1	57.0	68.2

Table 2. Performance at peak F-measure for the onset detector using various ODFs, with adaptive whitening activated.

adaptively-whitened power ODF is notable given the computational simplicity of the power algorithm. We concur with Dixon [10] who suggests that given the similar achievements of the strongest-performing ODFs, the choice of ODF for an application could be based on other factors such as simplicity of implementation and speed of execution. Indeed, for real-time applications efficiency of execution is an important factor, and this may speak for the adaptively-whitened power ODF, as indicated by the CPU usage figures presented in Figure 4.

Some investigators (e.g. [6], [11]) have found it beneficial to combine different ODFs together, creating a hybrid onset detector that can outperform its constituent components. It may be interesting in future to investigate the role of adaptive whitening in this context, for example in combining an adaptively-whitened ODF with a non-whitened ODF (e.g. the non-whitened MKL).

Another avenue for future research is to explore fully the effect of the parameters to the adaptive whitening algorithm, and how these vary according to the type of music signal presented.

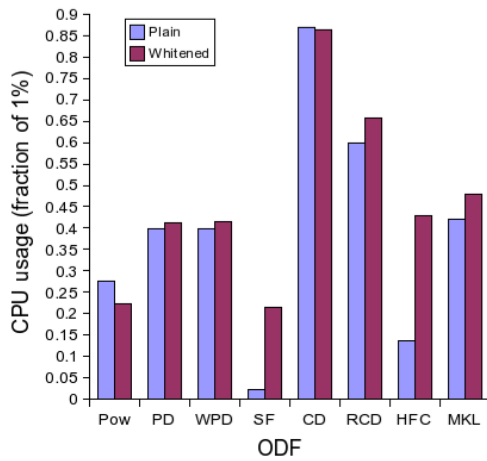


Figure 4. Real-time CPU usage for each ODF, with and without adaptive whitening. The values are derived for each ODF separately, as follows: 60 of the same onset detector are run in parallel, analysing the same 0.25 s audio loop (played with a different phase offset and separate FFT for each ODF instance), and the average CPU usage is recorded. The average CPU usage for the same system without any ODFs is subtracted (i.e. for running the audio playback and FFT), and the resulting value is divided by 60 to give an estimate for the CPU usage of a single ODF. Tests were performed in SuperCollider 3, using the first author’s C++ implementation of the ODFs, on a 2 GHz Mac Intel Core 2 Duo.

5. CONCLUSIONS

In this paper we have described adaptive whitening, a simple and computationally efficient process which improves the performance of a variety of real-time onset detectors. It adaptively modifies the magnitudes of STFT frames in a way which compensates for spectral roll-off and dynamic variation in a musical signal.

We have investigated the effect of adaptive whitening on a number of ODFs from recent literature on musical onset detection. With the exception of the Modified Kullback-Leibler and phase deviation ODFs, we found that adaptive whitening improved onset detector performance generally. In particular, it improved performance on the complex mixture and monophonic pitched datasets most strongly – the datasets least effectively processed by non-whitened ODFs – for which improvements ranged from around 2 to around 16 percentage points. We may say in this respect that adaptive whitening makes the “difficult” cases “easier”.

The strongest performing of the adaptively-whitened ODFs were rectified complex deviation, power, complex deviation, and weighted phase deviation, all of which performed generally well across the different datasets. Given the general similarities in peak performance, for real-time use the requirement of computational efficiency may tend to favour the adaptively-whitened power ODF.

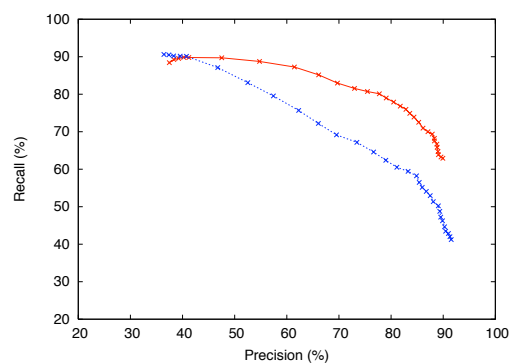
Our experiments are conducted using SuperCollider 3 [17], and to this end we have implemented the adaptive whitening algorithm as a frequency-domain unit generator for SuperCollider. This is freely available online.¹

6. REFERENCES

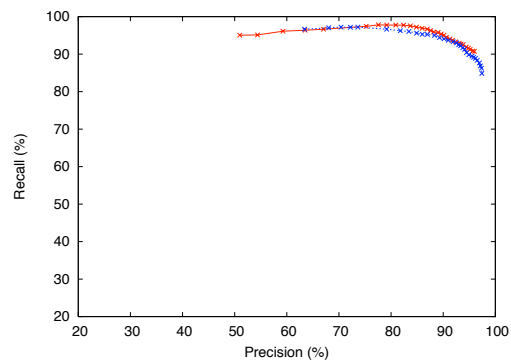
- [1] Mirex 2006 Audio Onset Detection Results. www.music-ir.org/mirex2006/index.php/Audio_Onset_Detection_Results, retrieved 30th March 2007.
- [2] S. A. Abdallah and M. P. Plumbley. Probability as metadata: Event detection in music using ICA as a conditional density model. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 233–238, April 2003.
- [3] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, 2005.
- [4] P. M. Brossier. *Automatic Annotation of Musical Audio for Interactive Applications*. PhD thesis, Queen Mary, University of London, August 2006.
- [5] P. M. Brossier, J. P. Bello, and M. D. Plumbley. Fast labelling of notes in music signals. In *Proc. Int. Symposium on Music Information Retrieval (ISMIR)*, pages 331–336, 2004.
- [6] P. M. Brossier, J. P. Bello, and M. D. Plumbley. Real-time temporal segmentation of note objects in music signals. In *Proc. International Computer Music Conference (ICMC’04)*, pages 458–461, 2004.
- [7] N. Collins. Using a pitch detector for onset detection. In *Proc. Int. Symposium on Music Information Retrieval (ISMIR)*, pages 100–106, 2005.
- [8] N. Collins. *Towards Autonomous Agents for Live Computer Music: Realtime Machine Listening and Interactive Music Systems*. PhD thesis, University of Cambridge, 2006.
- [9] L. Daudet. Transients modeling by pruned wavelet trees. In *Proc. International Computer Music Conference (ICMC’01)*, pages 18–21, 2001.
- [10] S. Dixon. Onset detection revisited. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-06)*, pages 133–137, Montreal, Quebec, Canada, Sept. 18–20, 2006.
- [11] C. Duxbury, M. Sandler, and M. Davies. A hybrid approach to musical note onset detection. In *Proceedings of the DAFX Conference, Hamburg, Germany*, pages 33–38, 2002.

¹ www.elec.qmul.ac.uk/digitalmusic/downloads/adaptivewhitening

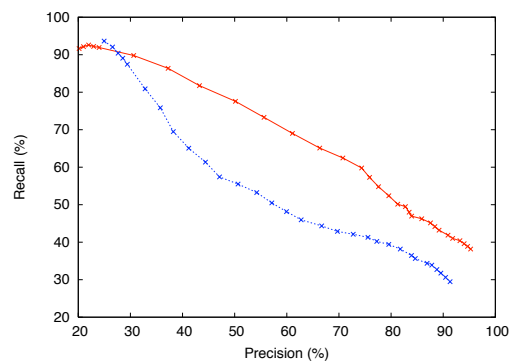
- [12] E. Kapanci and A. Pfeffer. A hierarchical approach to onset detection. In *Proc. International Computer Music Conference (ICMC'04)*, pages 438–441, 2004.
- [13] A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 6, pages 3089–3092, 1999.
- [14] A. Lacoste and D. Eck. A supervised classification algorithm for note onset detection. *EURASIP Journal on Applied Signal Processing*, August 2007.
- [15] P. Leveau, L. Daudet, and G. Richard. Methodology and tools for the evaluation of automatic onset detection algorithms in music. In *Proceedings of 5th International Symposium on Music Information Retrieval*, pages 72–75, Barcelona, Spain, 2004.
- [16] P. Masri. *Computer Modeling of Sound for Transformation and Synthesis of Musical Signals*. PhD thesis, University of Bristol, UK, 1996.
- [17] J. McCartney. Rethinking the computer music language: SuperCollider. *Computer Music Journal*, 26(4):61–68, 2002.
- [18] B. C. J. Moore. *An introduction to the psychology of hearing*. Academic Press London, UK, 2nd edition, 1982.
- [19] D. Patterson, J. Hennessy, D. Goldberg, and K. Asanovic. *Computer Architecture: a quantitative approach*. Morgan Kaufmann, 3rd edition, 2003.
- [20] Vic Firth Education Team and The Percussive Arts Society. 40 essential rudiments: The flam. www.vicfirth.com/education/rudiments/20flam.html, retrieved 30th March 2007.
- [21] I. Witten and E. Frank. *Data Mining: Practical machine learning tools and technique*. Morgan Kaufmann, 2nd edition, 2005.



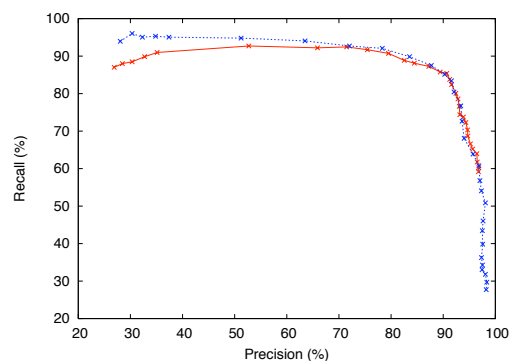
(a) Complex mixture



(b) Solo drums



(c) Monophonic pitched



(d) Polyphonic pitched

Figure 5. Precision/recall plots comparing power-based onset detection with and without adaptive whitening, for each of the four MIREX datasets. The x-axis shows the precision, and the y-axis the recall; the closer to the “top right”, the better the results. For each audio type, the “plain” power ODF results at various threshold settings are displayed as the blue (dashed) line, and the equivalent results for the power ODF with adaptive whitening are displayed as the red (solid) line.