Simple Spectrum-Based Onset Detection

Simon Dixon
Austrian Research Institute for Artificial Intelligence
Freyung 6/6, Vienna 1010, Austria
simon.dixon@ofai.at

Abstract

In a recent empirical study, various methods for detecting the onset times of musical notes in audio signals were evaluated [1]. The study focussed on published methods based on spectral features such as the magnitude, phase and complex domain representations, and compared existing methods (spectral flux, phase deviation and complex difference) with proposed improvements to these methods (weighted phase deviation, normalised weighted phase deviation and rectified complex difference). Two test sets were used: a set of short excerpts from a range of instruments (1060 onsets), plus a much larger data set of piano music (106054 onsets). Results showed a similarly high level of performance with a magnitude-based (spectral flux), a phase-based (weighted phase deviation) or a complex domain (complex difference) onset detection function. For MIREX 2006, the following five onset detection functions were submitted: spectral flux, complex domain, rectified complex domain, weighted phase deviation and normalised weighted phase deviation.

Keywords: MIREX, spectral flux, phase deviation, complex domain

1. Introduction

Recent reviews and evaluations of onset detection methods can be found in [2, 3, 4, 1]. The onset detection functions described in this document are more fully described and compared in [1]. Although it is clear that different methods are suitable for different data sets, we focus on simple, general-purpose methods of finding onsets. All methods presented here share the same peak picking algorithm, which limits the closeness of successive onsets. For polyphonic music, this might penalise the algorithms, depending on how the evaluation is performed.

2. Onset Detection Functions

An onset detection function is a function whose peaks are intended to coincide with the times of note onsets. Onset detection functions usually have a low sampling rate (e.g. 100Hz) compared to audio signals; thus they achieve a high level of data reduction whilst preserving the necessary information about onsets. Most onset detection functions are based on the idea of detecting changes in one or more properties of the audio signal.

If an audio signal is observed in the time-frequency plane, an increase in energy (or amplitude) within some frequency band(s) is a simple indicator of an onset. Alternatively, if we consider the phase of the signal in various frequency bands, it is unlikely that the frequency components of the new sound are in phase with previous sounds, so irregularities in the phase of various frequency components can also indicate the presence of an onset. Further, the phase and energy (or magnitude) can be combined in various ways to produce more complex onset detection functions. These ideas form the basis of the onset detection functions described in this paper.

All of the methods presented here make use of a time-frequency representation of the signal based on a short time Fourier transform using a Hamming window $w(m)$, and calculated at a frame rate of 100 Hz. If $X(n,k)$ represents the kth frequency bin of the nth frame, then:

$$X(n,k) = \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} x(hn+m)\, w(m)\, e^{-\frac{2j\pi mk}{N}}$$

where the window size $N = 2048$ (46 ms at a sampling rate of $r = 44100$ Hz) and hop size $h = 441$ (10 ms, or 78.5% overlap).

2.1. Spectral Flux

Spectral flux measures the change in magnitude in each frequency bin, and if this is restricted to the positive changes and summed across all frequency bins, it gives the onset function SF [5]:

$$SF(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} H(|X(n,k)| - |X(n-1,k)|)$$

where $H(x) = \frac{x+|x|}{2}$ is the half-wave rectifier function. Empirical tests favoured the use of the L_1-norm here over the L_2-norm used in [6, 2], and the linear magnitude over the logarithmic (relative or normalised) function proposed by Klapuri [7].

2.2. Phase Deviation

The rate of change of phase in an STFT frequency bin is an estimate of the instantaneous frequency of that component. This can be calculated via the first difference of the phase of $X(n,k)$. Let $\psi(n,k)$ be the phase of $X(n,k)$, that is:

$$X(n,k) = |X(n,k)|\, e^{j\psi(n,k)}$$

where $-\pi < \psi(n,k) \leq \pi$. Then the instantaneous frequency is given by the first difference $\psi'(n,k)$:

$$\psi'(n,k) = \psi(n,k) - \psi(n-1,k)$$

mapped onto the range $(-\pi, \pi]$. The change in instantaneous frequency, which is an indicator of a possible onset, is given by the second difference of the phase:

$$\psi''(n,k) = \psi'(n,k) - \psi'(n-1,k)$$

which is also mapped onto the range $(-\pi, \pi]$. Large discontinuities in the unwrapped phase or its derivatives can wrap around to 0, but the onset detection function based on phase deviation, PD, takes the mean of the absolute changes in instantaneous frequency across all bins [8, 2], which reduces the chance of a missed detection:

$$PD(n) = \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |\psi''(n,k)|$$

2.3. Weighted Phase Deviation

Phase deviation performs poorly because of "noise introduced by components with no significant energy" [2]. That is, the function considers all frequency bins k equally, although the energy of the signal is concentrated around the bins containing the partials of the currently sounding tones. The *weighted phase deviation* (WPD) function takes this into account by weighting the phase deviation values by the magnitude of the corresponding frequency bin:

$$WPD(n) = \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |X(n,k)\, \psi''(n,k)|$$

The *normalised weighted phase deviation* ($NWPD$) function is similar, except that the sum of the weights is factored out, to give a weighted average phase deviation:

$$NWPD(n) = \frac{\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |X(n,k)\, \psi''(n,k)|}{\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |X(n,k)|}$$

2.4. Complex Domain

Another way of jointly considering amplitude and phase is to search for departures from "steady-state" behaviour in the complex domain, by calculating the expected amplitude and phase of the current bin $X(n,k)$, based on the previous two bins $X(n-1,k)$ and $X(n-2,k)$. The target value $X_T(n,k)$ is estimated by assuming constant amplitude and rate of phase change:

$$X_T(n,k) = |X(n-1,k)|\, e^{\psi(n-1,k)+\psi'(n-1,k)}$$

and therefore a complex domain onset detection function CD can be defined as the sum of absolute deviations from the target values:

$$CD(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |X(n,k) - X_T(n,k)|$$

This formulation is simpler but equivalent to the complex domain detection function in [2, 9].

2.5. Rectified Complex Domain

One problem with the CD method is that it does not distinguish between increases and decreases in amplitude of the signal, so that onsets are not distinguished from offsets. The rectified complex domain (RCD) onset detection function uses half-wave rectification to preserve the complex differences only in spectral bins where energy is increasing:

$$RCD(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} RCD(n,k)$$

where

$$RCD(n,k) = \begin{cases} |X(n,k) - X_T(n,k)|, & \text{if} |X(n,k)| \geq \\ & |X(n-1,k)| \\ 0, & \text{otherwise} \end{cases}$$

3. Onset Selection

The onsets are selected from the detection function by a peak-picking algorithm which finds local maxima in the detection function, subject to various constraints. The thresholds and constraints used in peak-picking have a large impact on the results, specifically on the ratio of false positives to false negatives. For example, a higher threshold generally reduces the number of false positives and increases the number of false negatives. The best values for thresholds are dependent on the application and the relative undesirability of false positives and false negatives.

Peak picking is performed as follows: each onset detection function $f(n)$ is normalised to have a mean of 0 and standard deviation of 1. Then a peak at time $t = \frac{nh}{r}$ is selected as an onset if it fulfils the following three conditions:

$$f(n) \geq f(k) \text{ for all } k \text{ such that } n - w \leq k \leq n + w$$
$$f(n) \geq \frac{\sum_{k=n-mw}^{n+w} f(k)}{mw + w + 1} + \delta$$
$$f(n) \geq g_\alpha(n-1)$$

	PN	PP	NP	CM	Sonatas	Error
SF	0.952	0.984	0.967	0.882	0.964±0.017	8.8
WPD	0.947	0.912	0.966	0.836	0.912±0.028	9.6
NWPD	0.938	0.971	0.958	0.879	0.944±0.021	10.3
CD	0.946	0.978	0.936	0.876	0.966±0.015	12.8
RCD	0.963	0.981	0.963	0.877	0.955±0.018	9.3

Table 1. *Results of onset detection tests for 5 onset detection functions (SF, WPD, NWPD, CD and RCD). The first four columns show the maximum of the F-measure for the four subsets of data set 1: pitched non-percussive (PN), pitched percussive (PP), non-pitched percussive (NP) and complex mixture (CM). The last 2 columns show results for data set 2 (Sonatas): the F-measure with standard deviation of F-measures across sonatas and average absolute error in ms.*

where $w = 3$ is the size of the window used to find a local maximum, $m = 3$ is a multiplier so that the mean is calculated over a larger range before the peak, δ is the threshold above the local mean which an onset must reach, and $g_\alpha(n)$ is a threshold function with parameter α given by:

$$g_\alpha(n) = \max(f(n),\ \alpha g_\alpha(n-1) + (1-\alpha)f(n))$$

Experiments were performed with various values of the two parameters δ and α, and it was found that best results were obtained using both parameters, but the improvement in results due to the use of the function $g_\alpha(n)$ was marginal, assuming a suitable value for δ is chosen.

4. Results

Before submission, two data collections were used for testing the onset detection functions. The data from Bello et al. [2], consists of 4 sets of short excerpts from a range of instruments, classed into the following groups: NP — non-pitched percussion, such as drums (119 onsets); PP — pitched percussion, such as piano and guitar (577 onsets); PN — pitched non-percussion, in this case solo violin (93 onsets); and CM — complex mixtures from popular and jazz music (271 onsets). The second data collection consists of about 4 hours of Mozart Piano Sonatas (106054 onsets) — two orders of magnitude more than that used in other evaluations — and includes complex passages such as trills, fast scale passages with pedal and arpeggiated chords. The level of complexity is such that a human annotator would not be able to mark all the onsets precisely.

Table 1 shows the results across these two data sets. In each case, the results are shown for the point on the ROC curve which gives the maximum value of the F-measure. That is, the ground-truth data was used to select optimal values of δ and α. Further issues involving evaluation are discussed in [1].

In these results, the spectral flux, weighted phase deviation and complex domain methods all achieved a similar level of performance on this data, so that the choice of a suitable algorithm could be based on other factors such as simplicity of programming, speed of execution and accuracy of correct onsets (right column), which all speak for the spectral flux onset detection function (SF).

Entry	Precision	Recall	F-measure
roebel-3	0.836	0.779	0.788
roebel-2	0.831	0.769	0.780
roebel-1	0.861	0.746	0.777
du	0.797	0.799	0.762
brossier-hfc	0.752	0.774	0.734
dixon-sf	0.736	0.790	0.726
brossier-dual	0.769	0.735	0.724
brossier-complex	0.780	0.725	0.721
dixon-rcd	0.735	0.765	0.716
dixon-cd	0.709	0.776	0.710
brossier-specdiff	0.764	0.701	0.707
dixon-wpd	0.663	0.786	0.685
dixon-nwpd	0.524	0.908	0.620

Table 2. *Average precision, recall and F-measure for the best parameter setting for each of the MIREX 2006 entries, sorted by F-measure.*

The results from the MIREX 2006 competition are shown in Table 2. The performance of the onset detection functions is much lower than in Table 1. There are a number of reasons why this is the case: first, the parameter settings used in Table 1 were refined with knowledge of the onset times, allowing some amount of overfitting to the data. From the results in Table 2, it is clear that the range of parameter settings for the submitted onset detection functions was too narrow, so that the optimal point on the ROC curve was not reached in each case. This is particularly clear for the case of the NWPD function, where a bug in the submitted code led to wrong parameter values being used. It is also worth noting that the data used in Table 1 are relatively easy for onset detection; the first part consists of simple music, and the second part consists of complex music played on a simple-to-detect instrument, the piano. Further analysis of the results will yield insights into the specific strengths and weaknesses of the individual algorithms.

5. Acknowledgements

This work was supported by the Vienna Science and Technology Fund, project CI010 *Interfaces to Music*, and the EU project S2S^2. OFAI acknowledges the support of the ministries BMBWK and BMVIT. Thanks to Juan Bello for providing test data, and to the MIREX team for conducting the MIREX evaluation.

References

[1] S. Dixon, "Onset detection revisited," in *Proceedings of the 9th International Conference on Digital Audio Effects*, 2006, pp. 133–137.

[2] J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in musical signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[3] N. Collins, "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions," in *118th Convention of the Audio Engineering Society*, Barcelona, Spain, 2005.

[4] J.S. Downie, "2005 MIREX contest results - audio onset detection," www.music-ir.org/evaluation/mirex-results/audio-onset, 2005.

[5] P. Masri, *Computer Modeling of Sound for Transformation and Synthesis of Musical Signal*, Ph.D. thesis, University of Bristol, Bristol, UK, 1996.

[6] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," in *Proceedings of the 5th International Conference on Digital Audio Effects*, 2002, pp. 33–38.

[7] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Phoenix, Arizona, 1999.

[8] C. Duxbury, J.P. Bello, M. Davies, and M. Sandler, "A combined phase and amplitude based approach to onset detection for audio segmentation," in *Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS-03)*, 2003, pp. 275–280.

[9] 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, pp. 553–556, 2004.