

ECS7006 Music Informatics

Week 3 – Onset Detection

School of Electronic Engineering and Computer Science
Queen Mary University of London

prepared by Simon Dixon
using material by Juan Bello, Emmanuel Vincent and Meinard Müller

s.e.dixon@qmul.ac.uk

2023



Rhythm: Introduction and Background

The Role of Rhythm in Music

Definitions of Music

- “Organised sound”
- “The science or art of ordering tones or sounds in succession, in combination, and in temporal relationships to produce a composition having unity and continuity.”
- “The art of arranging sounds in time so as to produce a continuous, unified, and evocative composition, as through melody, harmony, rhythm, and timbre.”
- Rhythm (or more generally the organisation of sound in time) is a fundamental aspect of music.

What is Rhythm?

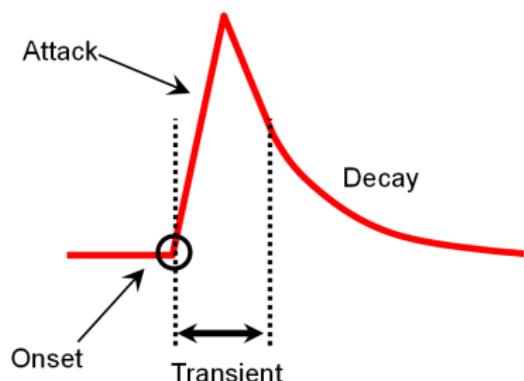
- *Rhythm* refers to the medium-scale temporal aspects of music, perceived as an arrangement of events, beats and accents in time
- Does not include short-term aspects such as the variations within a single musical note or long-term aspects such as the form (structure) of a piece of music
- Arises from the distinctive **grouping** of sounds and silences in time, based on the **durations** of tones, strong and weak **stresses**, and factors like harmony and melodic contour
- Rhythm has the following components:
 - **Tempo**: the rate of a pulse (sequence of beats)
 - **Timing**: when events occur (relative to pulses)
 - **Metre**: the structure of (or relationships between) pulses
 - Grouping: phrase structure (separate from metrical structure)

Event-Based Characterisation of Music

- Although music signals are continuous (ignoring sampling), music is an event-based phenomenon
- In both production and perception, the basic unit of organisation is usually the *note*, or more generally the *event* (unpitched and variable pitch events might not be called notes)
- Detecting and characterising events and their transitions are important tasks in computer music applications
 - e.g. transcription, coding, retrieval, performance studies, audio editing, content-based processing

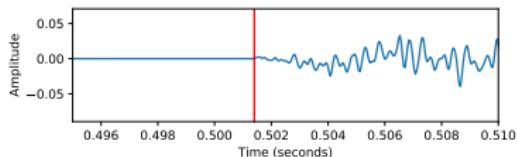
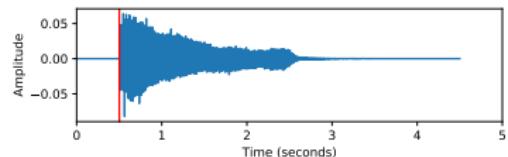
Rhythm: Based on Onset Times

- Rhythm is related to the onset times of events in the signal
- Music signals can be broadly modelled as containing:
 - Steady-state: when sinusoidal features are stable (slowly changing), thus predictable
 - Transients: when features are unstable (rapidly changing) and unpredictable
- The amplitude envelope of musical events is often represented by an attack-decay-sustain-release (ADSR) curve
 - the sustain portion corresponds to the steady-state
 - attack, decay (if fast) and release are transient
- *Onset detection* is the estimation of the beginnings of the attack transients



Definitions

- The *attack* portion of an event is the time interval at the beginning of the event during which its amplitude envelope increases
- A *transient* is a short interval in which the signal evolves quickly in a non-trivial and unpredictable way
- An *onset* is the single instant marking the beginning of an event
 - It usually coincides with the start of the transient
 - Sometimes the physical onset and perceptual onset are distinguished (Vos and Rasch, 1981)



- An *offset* marks the end of a note (often unclear)
- NB: In the MIDI protocol, a note consists of two “MIDI events” (Note-On and Note-Off), plus scalar attributes for pitch and intensity

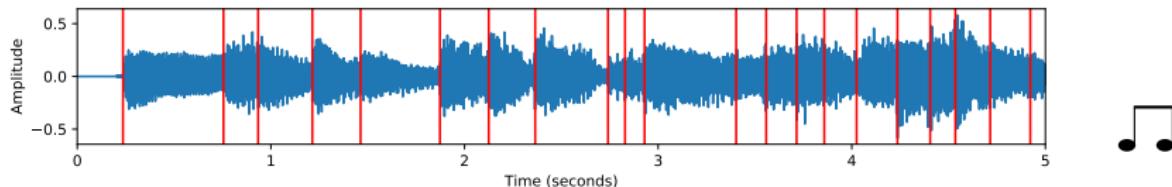
Onset Detection

Onset Detection: Motivation

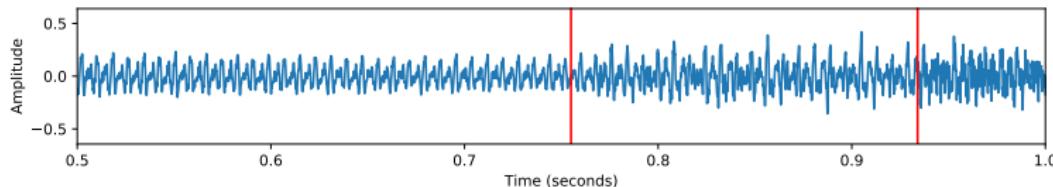
- Onset detection (finding the beginnings of events) is the first step towards understanding the underlying periodicities and accentuations in the signal, i.e. rhythm
- Onset detection is useful for applications including:
 - Audio editing tools (e.g. alignment)
 - Digital audio effects (e.g. time scaling)
 - Audio coding
 - Segmentation for analysis tools (e.g. transcription)
 - Investigation of performance timing

Onset Detection: The Challenge

- There is no unique way to characterise onsets, but they often share common features (e.g. a change in signal properties, a short duration transient, a sudden burst of energy, unpredictable and unstable components followed by a steady-state region)

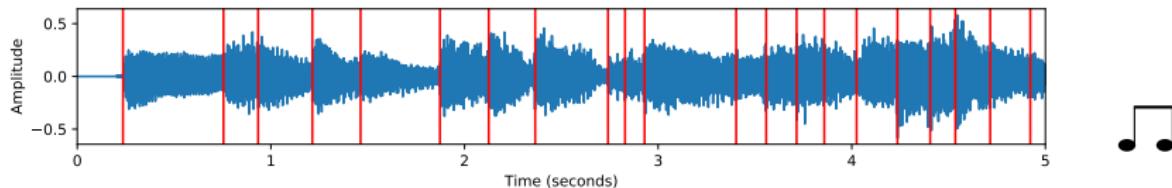


- Onsets can be difficult to identify in time-domain and frequency-domain signals, particularly in polyphonic and multi-instrumental musical signals

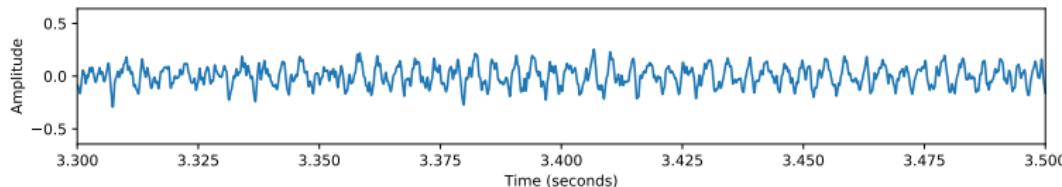


Onset Detection: The Challenge

- There is no unique way to characterise onsets, but they often share common features (e.g. a change in signal properties, a short duration transient, a sudden burst of energy, unpredictable and unstable components followed by a steady-state region)

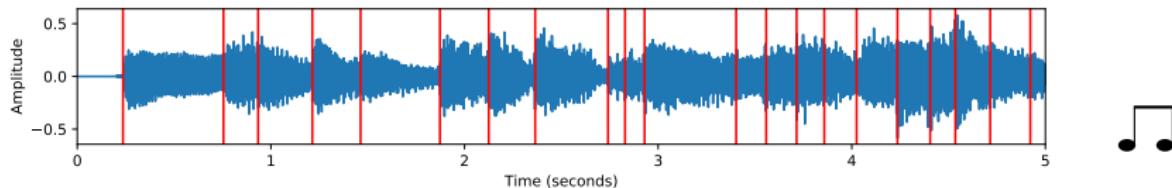


- Onsets can be difficult to identify in time-domain and frequency-domain signals, particularly in polyphonic and multi-instrumental musical signals

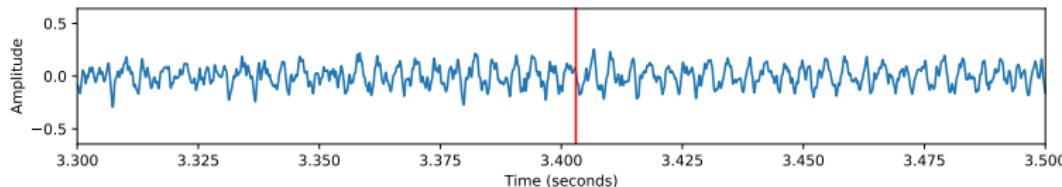


Onset Detection: The Challenge

- There is no unique way to characterise onsets, but they often share common features (e.g. a change in signal properties, a short duration transient, a sudden burst of energy, unpredictable and unstable components followed by a steady-state region)



- Onsets can be difficult to identify in time-domain and frequency-domain signals, particularly in polyphonic and multi-instrumental musical signals



Onset Detection: A Standard Approach

- The performance of onset detection algorithms depends on the technique used and the type of onsets:
 - Hard onsets: related to a percussive event (not hard to detect)
 - Soft onsets: related to a tonal change and/or slow attack (e.g. glissando, legato)
- The standard approach is to use an intermediate representation, which is called an *onset detection function* (or novelty function)
- Onsets are then computed from the detection function by *peak-picking* with suitable thresholds and constraints

Onset Detection Functions (ODFs)

We will consider a range of ODFs, starting with the simplest:

- Time domain onset detection algorithms
- Frequency domain onset detection algorithms
 - using magnitude only
 - using phase only
 - using magnitude and phase together

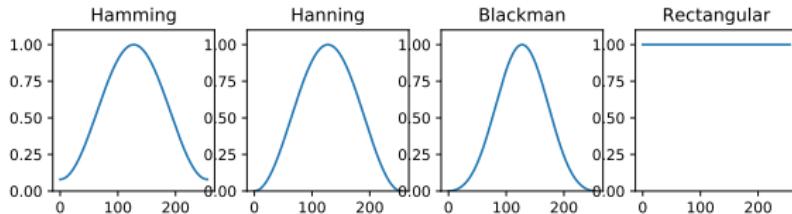
Time Domain Onset Detection

- An onset is often accompanied by an increase in signal amplitude
- So we start with a simple envelope follower (rectifying + smoothing):

$$E_0(n) = \frac{1}{N+1} \sum_{m=-N/2}^{N/2} |x(n+m)| w(m)$$

where $w(m)$ is a smoothing window and $x(n)$ is the signal

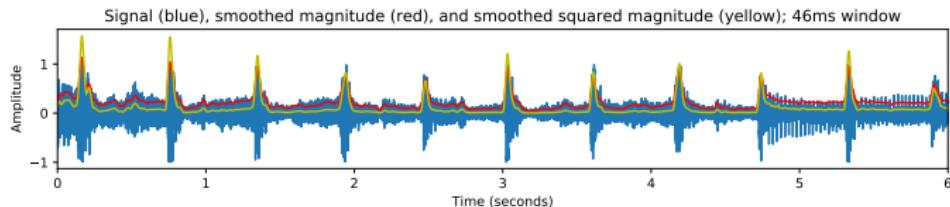
- The smoothing window w can be any positive-valued function with finite support; usually it is smooth and symmetric around a peak in the centre
- Any of the window functions used in STFT analysis can be used



Time Domain Onset Detection

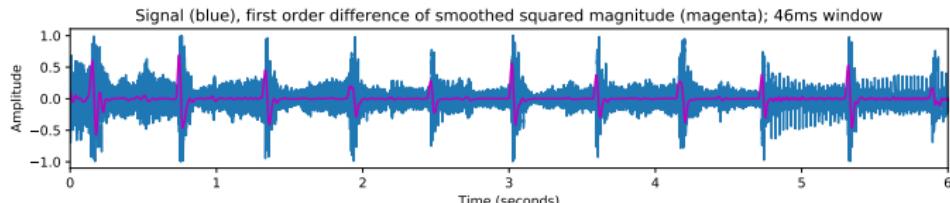
- Alternatively we can square the signal to obtain the power:

$$E(n) = \frac{1}{N+1} \sum_{m=-N/2}^{N/2} (x(n+m))^2 w(m)$$



- The next step is to take the time derivative of the signal, so that sudden rises appear as narrow peaks in the derivative:

$$E'(n) = E(n) - E(n-1)$$



Time Domain Onset Detection

- We can take inspiration from psychoacoustics (Klapuri, 1999)
- Loudness is perceived logarithmically
- The smallest detectable change (*just noticeable difference*, JND) in loudness is approximately proportional to the overall loudness
- Noting that:

$$\frac{\partial(\log E)}{\partial t} = \frac{1}{E} \frac{\partial E}{\partial t}$$

we can use the first time difference of $\log(E(n))$ to simulate human perception of changes in loudness

Frequency Domain Onset Detection

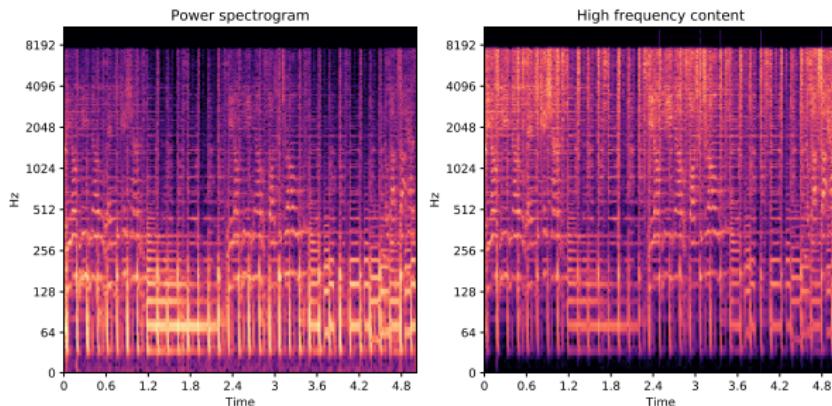
- Frequency domain methods are all based on STFT analysis
- If the hop size is H (in samples) and the sampling rate f_s , then frame n is centred at time $t = nH/f_s$
- If the window size is N , then frequency bin k is centred at frequency $f = kf_s/N$
- For each time-frequency point (n, k) , the complex value $X(n, k)$ represents the magnitude and phase of the frequency content at that point
- Power can also be computed in the frequency domain:

$$E(n) = \frac{1}{N+1} \sum_{k=-N/2}^{N/2} |X(n, k)|^2$$

High Frequency Content

- In the spectral domain, high-frequency energy appears at onsets (transients appear as wide-band noise), but dissipates rapidly
- The **high frequency content** (HFC) of a signal can be enhanced by applying a linear weighting to the power:

$$\text{HFC}(n) = \frac{4}{N(N+2)} \sum_{k=-N/2}^{N/2} |k| \cdot |X(n, k)|^2$$



Spectral Flux

- As for time domain methods, changes in the spectrum are better indicators of onsets than instantaneous measures such as HFC
- The **spectral flux** (SF) onset detection function is given by:

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

where $H(x)$ is the half-wave rectifier:

$$H(x) = \frac{x + |x|}{2} = \max(x, 0)$$

so that only the increases in energy are taken into account

- An alternative version squares the summands

Improvements on Spectral Flux: Superflux

- **Superflux** was designed to suppress false onsets due to vibrato
- Use a high frame rate (200 FPS)
- Use a logarithmic frequency scale (0.5 semitone spacing)
- Use logarithmic compression on the spectrum ($C > 1$):

$$X_1[n, k] = \log(1 + C \cdot |X[n, k]|)$$

- Use a maximum filter on the frequency bins:

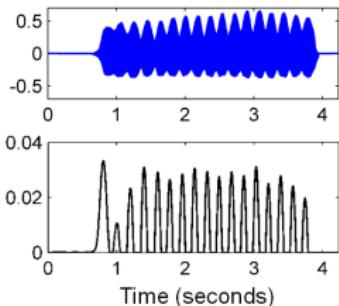
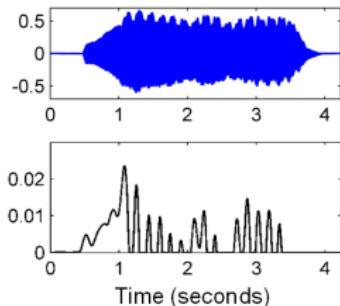
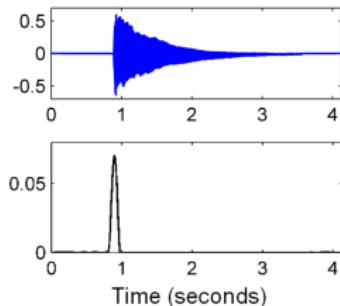
$$X_2[n, k] = \max(X_1[n, k - 1 : k + 1])$$

- Undo overlapping of windows (e.g. $\mu = 2$ for 50% overlap):

$$\text{Superflux}(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} H(|X_2(n, k)| - |X_2(n - \mu, k)|)$$

Limitations of Magnitude-Based Onset Detection

- The above methods are based on energy, power or magnitude
- This approach is effective for detecting percussive onsets (e.g. drums, piano, guitar)
- It is not as effective for detecting softer onsets (e.g. bowed strings or woodwinds played legato)



- It is also not effective when energy profiles of weaker notes are masked by stronger notes, as often occurs in polyphonic music
- Can we use phase information instead or as well as magnitude?

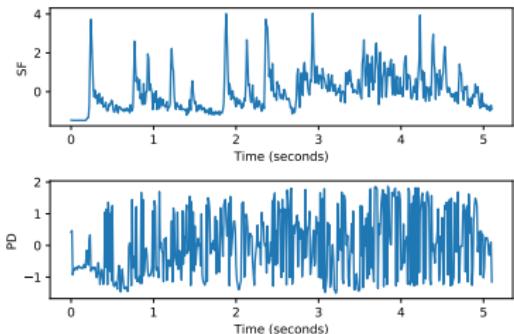
Phase-only Onset Detection

- If $X(n, k) = |X(n, k)| e^{j\phi(n, k)}$, then the **phase deviation** (PD) onset detection function is given by the mean absolute deviation from the expected (steady state) phase:

$$\begin{aligned} \text{PD}(n) &= \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |\text{princarg}(\phi''(n, k))| \\ &= \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |\text{princarg}(\phi(n, k) - 2\phi(n-1, k) + \phi(n-2, k))| \end{aligned}$$

where $\text{princarg}(\phi)$ wraps ϕ to the range $(-\pi, \pi]$

- Problem: PD is sensitive to noise. Frequency bins containing low energy are weighted equally with bins containing high energy, but bins containing low-level noise have random phase



Combining Phase and Magnitude with Thresholding

- The solution is to combine magnitude and phase values in the onset detection function, which can be done in several ways
- First way: only consider phase values for bins with magnitude above a fixed threshold α
- Gives the **thresholded phase deviation** (TPD) onset detection function:

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

where

$$T(n, k) = \begin{cases} |\phi''(n, k)|, & |X(n, k)| > \alpha \\ 0, & |X(n, k)| \leq \alpha \end{cases}$$

Combining Phase and Magnitude with Weighting

- Alternatively, the phase values can be weighted by the magnitude of the signal in each frequency bin, to give the **weighted phase deviation** (WPD) onset detection function:

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

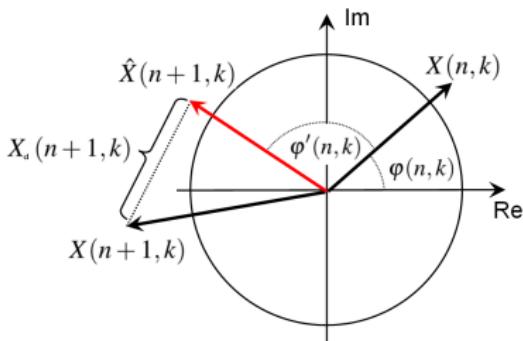
- The WPD function is higher for signals with higher amplitude; a relative function can be constructed by normalising the WPD to give the **normalised weighted phase deviation** (NWPD) onset detection function:

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

Using Phase and Magnitude in the Complex Domain

- An alternative approach to combining phase and magnitude is to consider the STFT bin values as vectors in the complex domain
- In the steady-state, the magnitude of bin k at time $n + 1$ is equal to its magnitude at time n
- Also, the phase at time $n + 1$ is the sum of the phase at n and the rate of phase change ϕ' at n
- Thus the target value is:

$$\hat{X}(n + 1, k) = |X(n, k)| e^{j(\phi(n, k) + \phi'(n, k))}$$



Complex Domain Onset Detection

- The **complex domain** (CD) onset detection function is defined as the sum of absolute deviations of observed values from the target values:

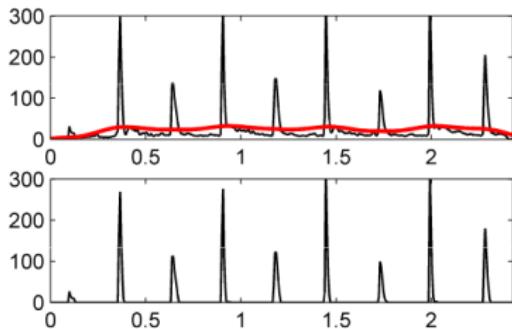
$$CD(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |X(n, k) - \hat{X}(n, k)|$$

- To distinguish between onsets and offsets, the sum can be restricted to bins with increasing magnitude to give the **rectified complex domain** (RCD) onset detection function:

$$RCD(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} \begin{cases} |X(n, k) - \hat{X}(n, k)|, & \text{if } |X(n, k)| \geq |X(n-1, k)| \\ 0, & \text{otherwise} \end{cases}$$

Post-Processing and Peak-Picking

- Peaks in the ODF should correspond to onset times, so the basic idea is to select local peaks above a threshold as onsets
- Fixed threshold values perform poorly, due to variety within and between recordings
- Dynamic (adaptive) thresholds can be generated from local properties of the signal e.g. by low pass or median filters
- e.g. subtract the local average and half-wave rectify



- A constraint can be applied to suppress peaks which are too close to another peak (double detections)

Evaluation of Onset Detection Functions

- Empirical evaluations require audio data annotated with the times of all the onsets
- Manual annotations are limited in scope and prone to errors and inaccuracy
- Automatic annotations can be generated from synthesised music and computer-monitored instruments (e.g. Bösendorfer's CEUS piano)
- Evaluations count the number of correctly detected onsets C (within some error threshold), missed onsets (*false negatives*, F_N) and false alarms (*false positives*, F_P)
- The accuracy of correctly detected onsets (timing error) is also relevant, and can be expressed in terms of bias (average error) and variance (mean square error)

Evaluation of Onset Detection Functions

- Several statistics can be reported:
 - Precision, the fraction of reported onsets which are correct:

$$P = \frac{C}{C+F_P}$$

- Recall, the fraction of correct onsets which are reported:

$$R = \frac{C}{C+F_N}$$

- F-measure: $F = \frac{2PR}{P+R} = \frac{2C}{2C+F_P+F_N}$

- Onset detection functions are very sensitive to thresholds; an ROC curve plotting P against R is often useful for understanding the relative performance of algorithms (area under curve, AUC)
- Depending on the application, certain types of errors might be more or less undesirable; usually thresholds can be adjusted to trade off precision against recall

Empirical Comparison

	PN data	PP data	NP data	CM data
SF	0.952	0.984	0.967	0.882
PD	0.770	0.619	0.831	0.704
WPD	0.947	0.912	0.966	0.836
NWPD	0.938	0.971	0.958	0.879
CD	0.946	0.978	0.936	0.876
RCD	0.963	0.981	0.963	0.877

Results of onset detection tests, showing the F-measure (F) for 4 data sets: pitched non-percussive (PN), pitched percussive (PP), non-pitched percussive (NP) and complex mixture (CM)

Ref: (Dixon, DAFX 2006)

Empirical Comparison

	P	R	F	E (ms)
SF	0.958	0.969	0.964 ± 0.017	8.8
PD	0.555	0.868	0.677 ± 0.044	19.5
WPD	0.903	0.921	0.912 ± 0.028	9.6
NWPD	0.945	0.944	0.944 ± 0.021	10.3
CD	0.970	0.962	0.966 ± 0.015	12.8
RCD	0.952	0.958	0.955 ± 0.018	9.3

Results of onset detection tests, showing precision (P), recall (R), F-measure (F) and absolute error in timing (E) for a data set consisting of over 106000 piano onsets (13 Mozart sonatas)

Ref: (Dixon, DAFX 2006)

Onset Detection: State of the Art

- If sufficient training data is available, neural networks (NNs) can be trained to recognise onsets
- Since 2010, several onset detection methods have been proposed with various network architectures:
 - bidirectional long short-term memory NNs (Eyben et al., 2010)
 - recurrent NNs (Böck et al., 2012)
 - convolutional NNs (Schlüter and Böck, 2013, 2014)
- These methods outperform approaches based on signal processing in public evaluations (MIREX, the Music Information Retrieval Evaluation eXchange):
https://www.music-ir.org/mirex/wiki/2019:Main_Page
- Current leading approaches use convolutional networks (CNNs) and are trained on tens of thousands of onsets

Onset Detection: Resources

- *A Tutorial on Onset Detection in Musical Signals*, IEEE Transactions on Speech and Audio Processing, 13 (5): 1035–1047 (J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies and M. Sandler, 2005)
- *Onset Detection Revisited*, 9th Int. Conf. on Digital Audio Effects (S. Dixon, 2006)
- *Fundamentals of Music Processing*, Section 6.1 (M. Müller, 2015)
- *Deep Learning for Event Detection, Sequence Labelling and Similarity Estimation in Music Signals*, PhD thesis, Johannes Kepler University, Linz (J. Schlüter, 2017)