A Tutorial on Onset Detection in Music Signals

J.P. Bello et al.

Goals

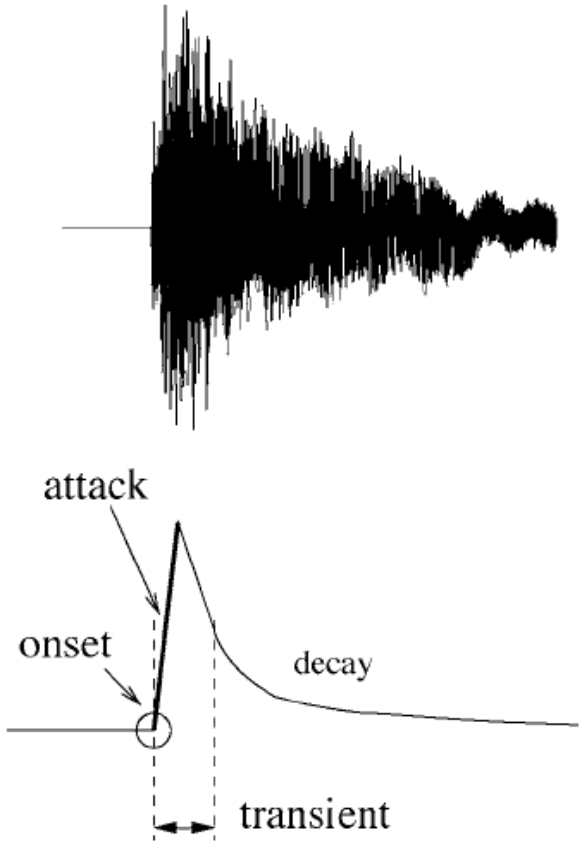
- Detect events in music signals. Specifically the beginning of notes.
- Multiple usage:
  - Proper segmentation of music signals
  - Extraction of important features
  - Segmented compression

This can be generalized to any time series

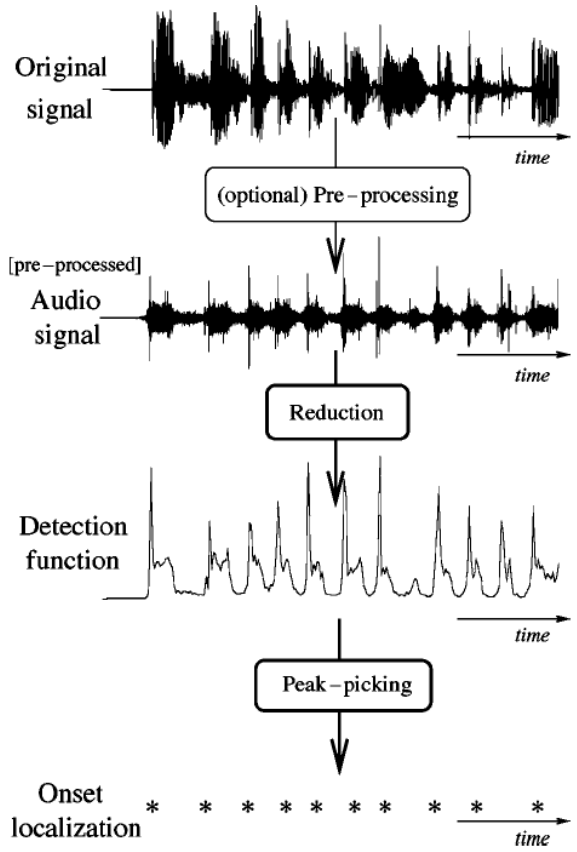
- Detection of transients in different signals:

- **Electrocardiogram (EKG)**
- **Seismograph data**
- **Stock-market results**

Definition of transients



In general, multi-step approach is used



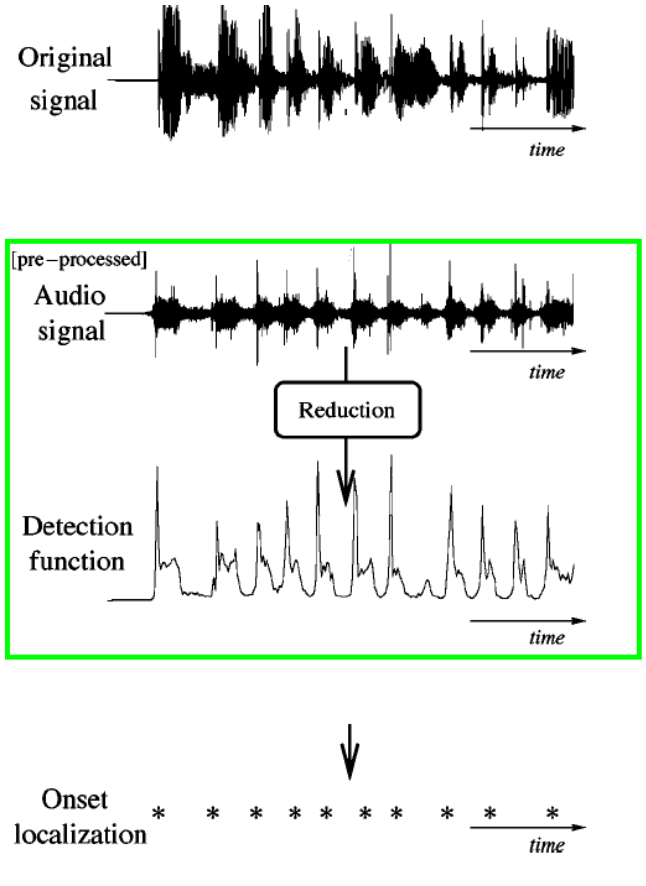
Pre-Processing
Multi-band separation

- Separate signals in multiple bands and combine each band decision to get final decision

Pre-Processing
Signal modelisation

- Model signal as a stationary signal (ex. sum of slowly varying cosines)
- Measure residual signal from diff. between model and original.
  - Burst in energy should indicate transient as our model is inadequate at that moment

Signal Reduction



Signal reduction

- Transform the signal into a detection function
  - Extract relevant features
  - Reduce the complexity of the signal

Signal Reduction

- Two broad categories
  - Reduction based on signal features
    - Temporal features
    - Spectral features
  - Reduction based on probabilistic model
    - Two competing models
    - Surprising moment approach

Temporal features

- Approach based on energy
- Measure the derivative of the energy
- Measure the derivative of the log of the energy (i.e. relative change in energy)

$$\frac{d(\log E)}{dt} = \frac{1}{E}\frac{dE}{dT}$$

Spectral features

- Rapid changes in the envelope usually lead to energy being present across the spectrum
  - Take the short term FFT of the signal
  - Take the spectral energy with a bigger weight on high frequencies

$$\widetilde{E}(n) = \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} W_k \, |X_k(n)|^2$$

Spectral Features

- Alternatively, look for the evolution of the energy per band
  - Rapid rise in energy should be due to transient
  - Example:

$$SD(n) = \sum_{k=\frac{N}{2}}^{\frac{N}{2}-1} \{H\left(|X_k(n)| - |X_k(n-1)|\right)\}^2$$

$$H(x) = (x + |x|)/2,$$

Spectral Features

- Previous methods where based on the amplitude
- Alternatively, we can look at the phase

Spectral Features

- If the signal is a stationary sine wave, phase changes across FFT windows should remain the same:

$$\varphi_k(n) - \varphi_k(n-1) \simeq \varphi_k(n-1) - \varphi_k(n-2).$$

- Take the second derivative and check for variations: $\Delta\varphi_k(n) = \varphi_k(n) - 2\varphi_k(n-1) + \varphi_k(n-2) \simeq 0.$

$$\zeta_p(n) = \frac{1}{N}\sum_{k=1}^{N} |\Delta\varphi_k(n)|$$

Time Frequency Representations

- Fourier analysis contains perfect spectral information, but time of different events is lost (STFFT solves this a bit by windowing)
- TFR contain both some spectral and time information
- Transform the signal with wavelets, in this case Haar wavelets.
- Can give better time resolution

Probabilistic models

- Assume the probability of a given sample is dependant on past samples
- Then measure the “surprise” of obtaining the actual sample.
  - A high surprise value indicates current frame is very different from our model

$$S(n) \equiv S(\mathbf{x}(n)) \stackrel{\text{def}}{=} -\log p(\mathbf{x}(n) \mid \{\mathbf{x}(j) : j < n\})$$

Probabilistic models

- Can be applied to multiple samples (frames).
  - Split the frame in two, and use a joint distribution estimate to measure conditional probability (and then measure “surprise”)

$$p(x_2 \mid x_1) = \frac{p(x_2, x_1)}{p(x_1)}$$

$$S(x_2) = \log p(x_1) - \log p(x_1, x_2)$$

Independent component analysis models

- Assume that the frame x is the linear combination of s independently distributed random variables: x=As where A is a matrix
- We can then measure the probability of x:
- (and then measure “surprise”)

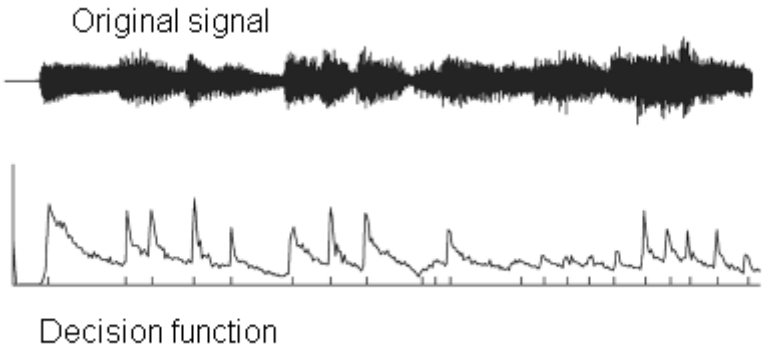
$$-\log p(\mathbf{x}) = -\sum_{i=1}^{N} \log p_i(s_i) + \log \det \mathbf{A}$$

Probabilistic models

- Unfortunately, they need training on the data to estimate the parameters (computationally expensive)
- Based on certain model assumptions, we can derive methods based on spectral computations
  - Probabilistic models therefore can be seen as a superset of our other models

Peak Picking

- Once we have reduced the signal, we need to trig based on decision function
- Search for peaks in detection function



Thresholding

- Absolute thresholding d(n)>cte
  - Not very flexible, not robust on dynamic signals
- Relative thresholding: take into account values of local d(n)

$$\ddot{\delta}(n) = \delta + \lambda \sum_{i=-M}^{M} w_i d^2(n+i) \qquad \dot{\delta}(n) = \delta + \lambda \,\mathrm{median}\,\{|d(n-M)|, \ldots, |d(n+M)|\}.$$

  - Takes into account relative amplitude of d(n)

Comparison

- 5 different reduction methods on 1065 different signals
  - High Frequency Content $\widetilde{E}(n) = \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} W_k \left|X_k(n)\right|^2$
  - Spectral Difference $SD(n) = \sum_{k=\frac{N}{2}}^{\frac{N}{2}-1} \left\{H\left(\left|X_k(n)\right| - \left|X_k(n-1)\right|\right)\right\}^2$
  - Phase Deviation $\zeta_p(n) = \frac{1}{N} \sum_{k=1}^{N} \left|\Delta\varphi_k(n)\right|$
  - Wavelet regularity modulus (Haar)
  - ICA Negative Log-likelihood $S(\mathbf{x}_2) = \log p(\mathbf{x}_1) - \log p(\mathbf{x})$

Comparison

- Peak picking was done with relative threshold based on the median of d(n).
  - Parameters of thresholding function where chosen manually for each reduction methods
  - Only static threshold constant was changed for comparison

Comparison

- 4 groups of signals, all at 44.1 kHz
- Onset labeling done manually on all signals
  - Somewhat imprecise
    - Successful detection is <50ms

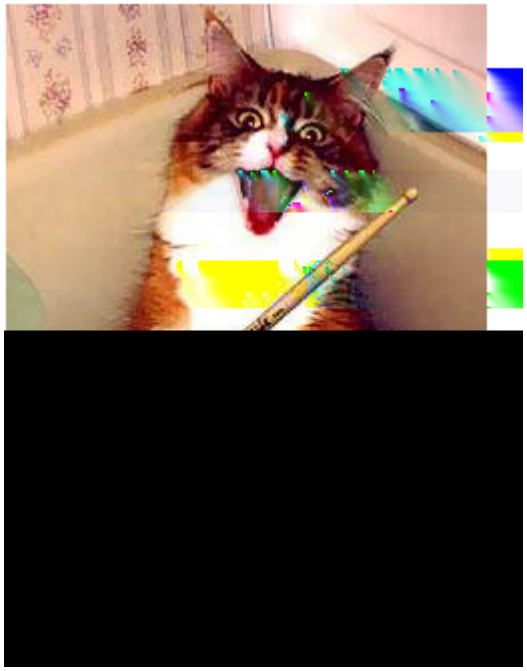
Types of signals

- Pitched non percussive

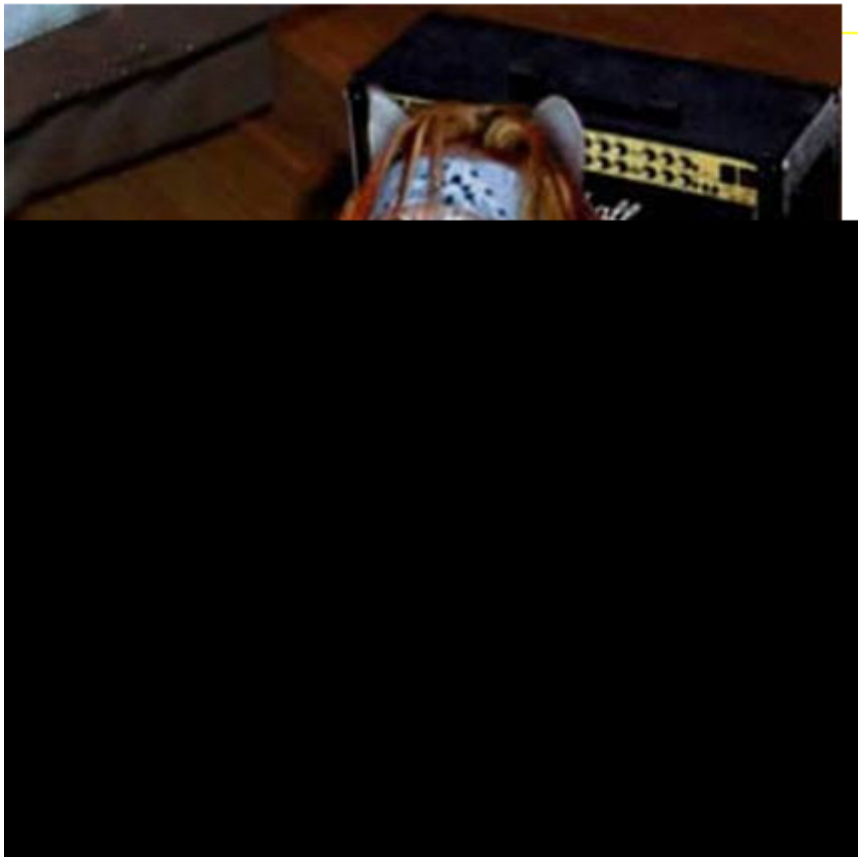
- Pitched percussive

Types of signal

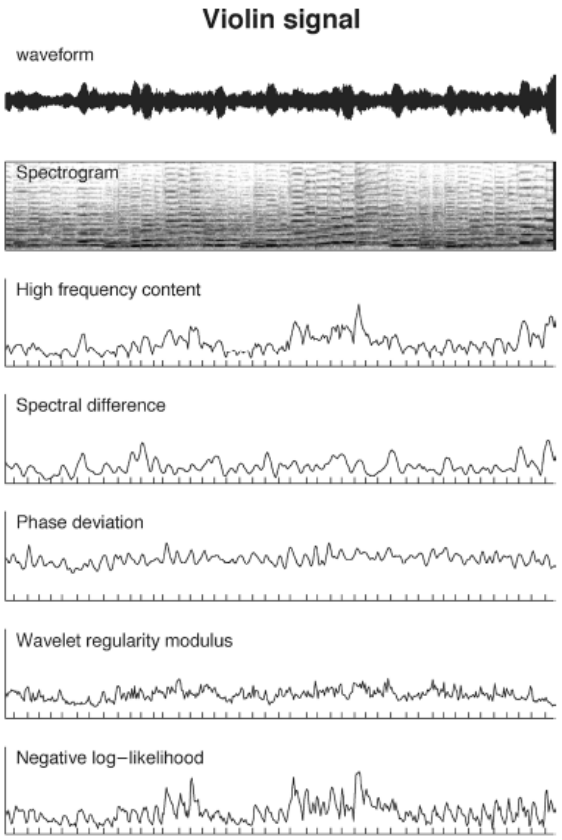
- Non pitched percussive

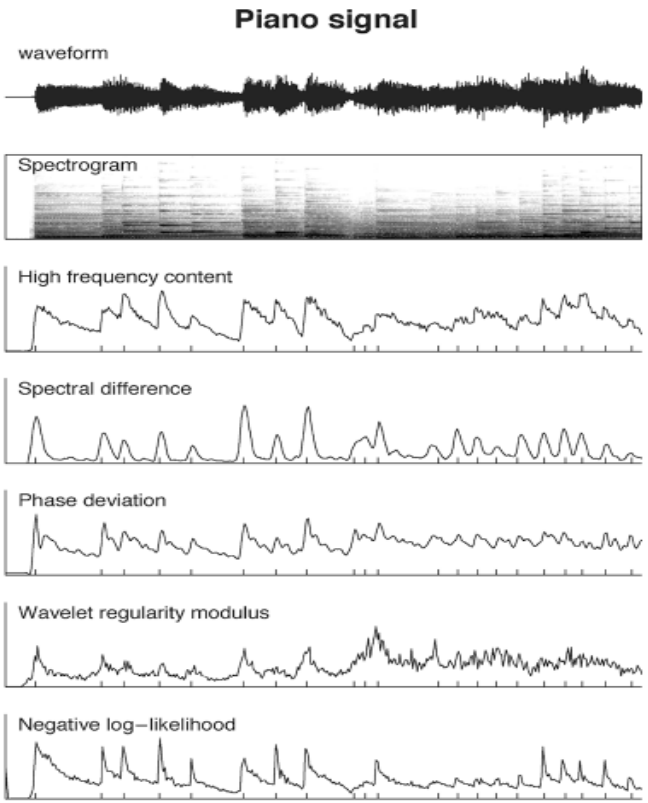


- Pop music

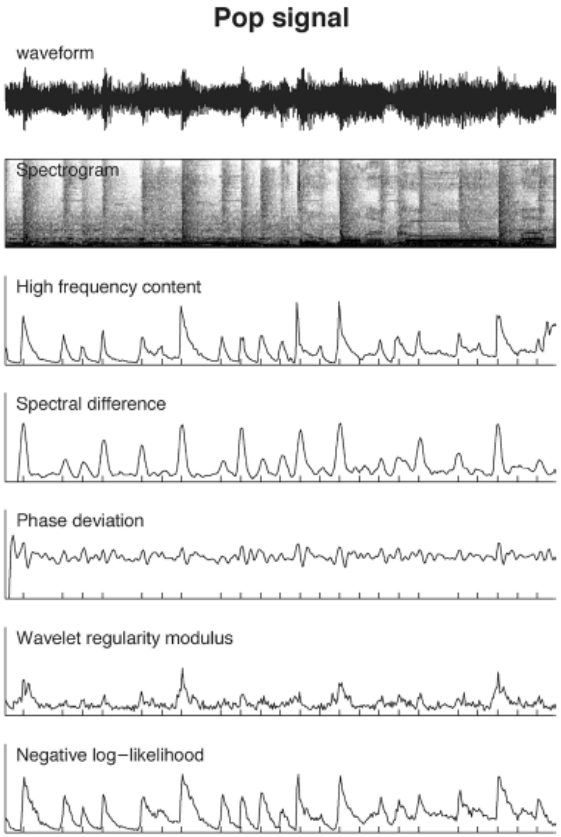


Comparison of detection functions



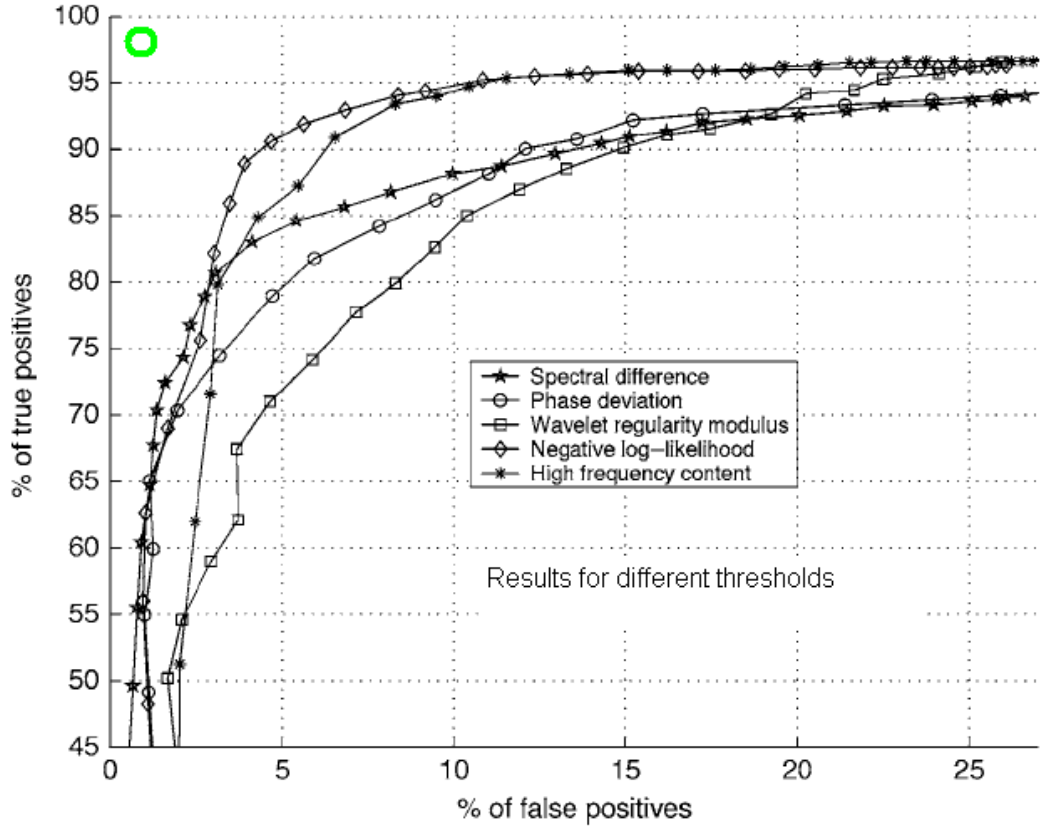


Comparison of detection functions



- Percussion signals easier to spot

Overall Results



Overall Results

- Optimal point for each method (distance)
  - Log-likelihood: 90.6%, 4.7%
  - HFC: 90%, 7%
  - Spectral diff. 83%,4.1%
  - Phase dev. 81.8%,5.6%
  - Wavelets 79.9%,8.3

Overall Results

- Log-likelihood gives best overall results
- HFC also give good Positive/Negative ratio
- Wavelets are not that good

Type of Onset Results

PITCHED NON-PERCUSSIVE - 93 ONSETS

METHOD	TP %	FP %
High frequency content	81.7	14.7
Spectral difference	87.1	8.6
Phase deviation	95.7	4.3
Wavelet reg. modulus	92.5	10.1
Neg. log-likelihood	96.8	3.2

PITCHED PERCUSSIVE - 489 ONSETS

METHOD	TP %	FP %
High frequency content	94.1	5.4
Spectral difference	94.9	1.6
Phase deviation	95.5	0.3
Wavelet reg. modulus	92.7	5.1
Neg. log-likelihood	92.4	3.1

NON-PITCHED PERCUSSIVE - 212 ONSETS

METHOD	TP %	FP %
High frequency content	96.7	0.0
Spectral difference	81.6	5.5
Phase deviation	80.7	5.5
Wavelet reg. modulus	88.7	2.2
Neg. log-likelihood	92.9	1.7

COMPLEX MIX - 271 ONSETS

METHOD	TP %	FP %
High frequency content	84.5	10.8
Spectral difference	80.4	10.4
Phase deviation	80.1	24.7
Wavelet reg. modulus	81.9	27.7
Neg. log-likelihood	86.0	10.8

Type of Onset Results

- Phase based methods perform poorly on non-pitched sounds but outperform HFC on pitched non percussive
  - No harmonics present vs no aggressive attack
- HFC performs better on percussive sounds
  - More abrupt onsets with percussive instruments lead to more high frequency contents at onsets
- Complex signals have a lower success rate
  - Phase based methods suffer from richness of music

Conclusions

- There is no best method. Computation cost and type of signal must be taken into account
- For percussive signals, temporal methods suffice
- HFC a good complexity/precision compromise
  - But if purely non-percussive, phase based approached might be better
- If computation costs are not a problem, probabilistic approach is recommended
- Advantage of wavelets is very precise time localization vs spectral, phase based approach