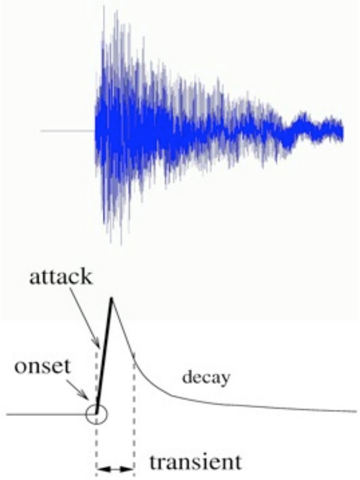
Characterizing temporal events in music signals

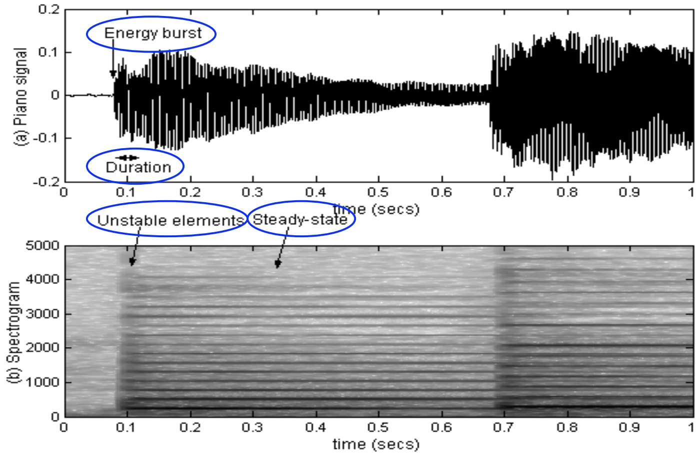
Definitions

- The attack refers to the time interval during which the amplitude envelope increases
- The transients refer to short intervals in which the signal evolves quickly in a non-trivial and unpredictable way
- The onset is the single instant chosen to mark the temporally extended transient. Usually it will coincide with the start of the transient



Detecting onsets

We can exploit some of the most common features of transients to characterize them and estimate their corresponding *onsets*.



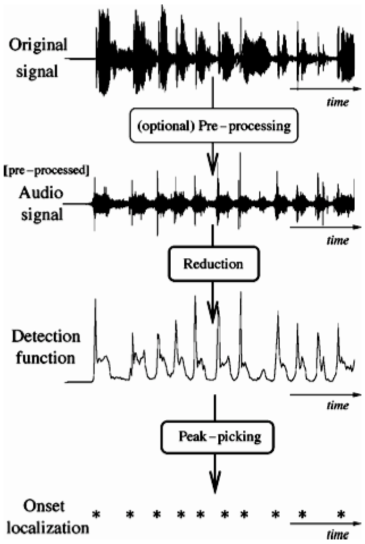
Detecting onsets

- Onset detection is useful for a number of applications including:
  1. Audio editing tools
  2. Digital audio effects (e.g. time scaling)
  3. Audio coding
  4. Synthesis
  5. Segmentation for analysis tools (e.g. transcription)

- Onset detection, i.e. characterizing the temporal location of events in the music signal, is the first step towards understanding the underlying periodicities and accentuations in the signal, i.e. rhythm.

- There are many techniques for onset detection, which perform differently for different types of onsets:
  a. Hard onsets: related to a percussive event
  b. Soft onsets: related to a light tonal change (e.g. glissando, legato)

Onset detection



- It is not possible to look for changes in time-domain waveform as they are both additive and oscillatory.
- This is even more so for common musical signals (polyphonic and multi-instrumental)
- It is thus necessary to use an intermediate representation, i.e. detection or novelty function

Time-domain

- The temporal evolution of music signals usually shows that the occurrence of an onset is often accompanied by an amplitude increase
- Thus using a simple envelope follower (rectifying + smoothing) is an obvious choice:

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

- Where w(m) is an N-length smoothing window and x(n) is the signal.
- Alternatively we can square the signal rather than rectify it to obtain the local energy

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

Time-domain

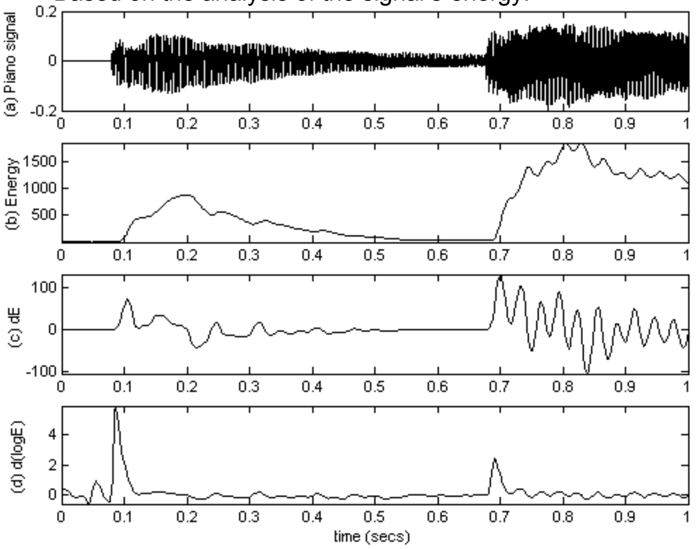
- A further refinement is to use the derivative of energy w.r.t. time, thus converting sudden rises in energy into narrow peaks in the derivative
- Furthermore, the study of psychoacoustics indicate that loudness is perceived logarithmically.
- For humans, the smallest detectable change in loudness is approximately proportional to the overall loudness of the signal (smaller changes are equally relevant in quieter signals), thus:

$$\frac{\partial E/\partial n}{E} = \frac{\partial(\log E)}{\partial n}$$

- Calculating the first difference of logE(n) w.r.t. time simulates the ear’s perception of loudness (Klapuri, 1999)

Time-domain

Based on the analysis of the signal’s energy:



Frequency-domain

- Many approaches exploit the behavior of the signal in the frequency-domain to characterize onsets.
- If $X_k(n)$ is the STFT of the signal x(n) times the N-length smoothing window w(m), then the local energy in the frequency domain is defined as:

$$E(n) = \frac{2}{N} \sum_{k=0}^{k=N/2} \left| X_k(n) \right|^2$$

- In the spectral domain, energy increases related to transients tend to appear as wide-band noise. This is more noticeable at high frequencies. We can emphasize that by using linear weighting

$$HFC(n) = \frac{2}{N} \sum_{k=0}^{k=N/2} \left| X_k(n) \right|^2 k$$

Frequency-domain

- As with the time-domain estimations, it is more robust to characterize changes in the spectrum than rely on instantaneous measures.
- The goal is to formulate the detection function as a distance metric between neighboring STFT frames.
- E.g. HFC differences, spectral differences (flux).
- An example is the L2 norm on the rectified difference:

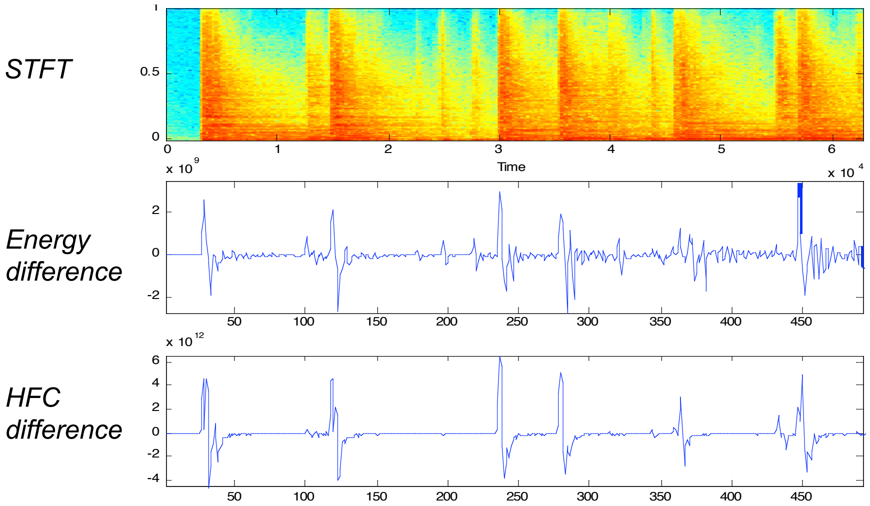
$$SD(n) = \sum_{k=0}^{k=N/2} \left\{ H\left(\left| X_k(n) \right| - \left| X_k(n-1) \right| \right) \right\}^2$$

- where:

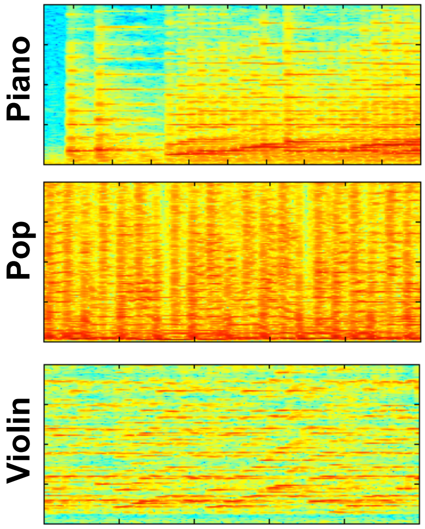
$$H(x) = (x + |x|)/2$$

is zero for negative arguments (so only energy increases are taking into account)

Frequency-domain



Energy-based detection



All studied methods are based on the energy content of the signal.

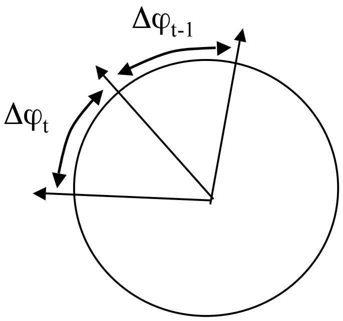
Energy-based detection is effective for percussive signals.

However it is not as effective when energy profiles of weaker notes are masked by those of stronger notes as is the case in polyphonic mixtures.

It also has troubles identifying softer onsets (e.g. bowed strings, woodwinds)

Phase-based detection

- An Alternative is to use phase information, as phase carries all timing information from the signal.
- Captures tonal changes (good for soft onsets)



$$\Delta\varphi_t \cong \Delta\varphi_{t-1}$$

- The deviation of the phase prediction for a given bin k is:

$$d_\varphi = princ\arg(\varphi_t - 2\varphi_{t-1} + \varphi_{t-2})$$

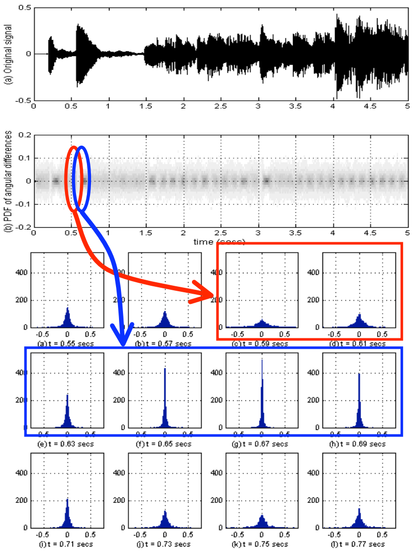
Phase-based detection

If we analyze the distributions of these phase deviations for all *k* along the time axis, we obtain a sequence of distributions that are:

Spread with a low central lobe during transients

Sharp with a high central lobe during steady-state

By quantifying these observations we can produce an onset detection function



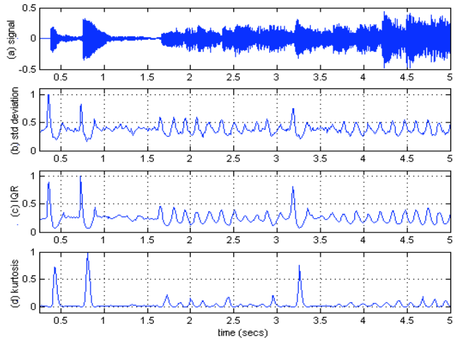
(a) t = 0.55 secs (b) t = 0.57 secs (c) t = 0.59 secs (d) t = 0.61 secs

(e) t = 0.63 secs (f) t = 0.65 secs (g) t = 0.67 secs (h) t = 0.69 secs

(i) t = 0.71 secs (j) t = 0.73 secs (k) t = 0.75 secs (l) t = 0.77 secs

Phase-based detection

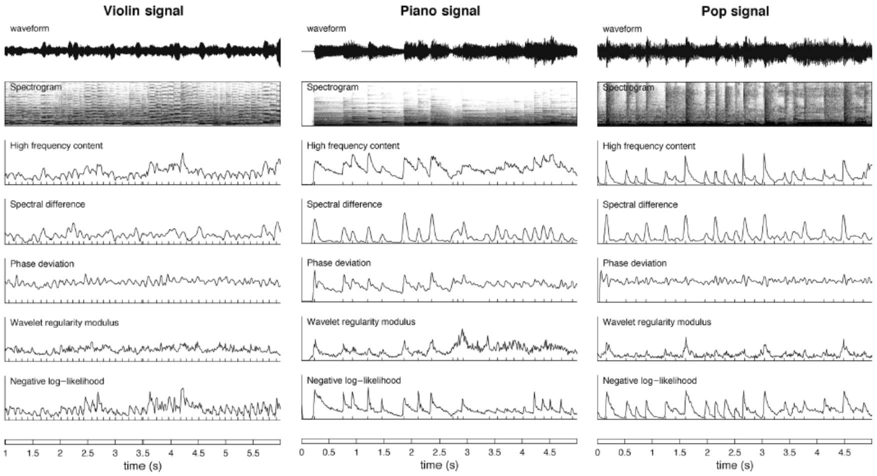
- Several approaches have been proposed to quantify this behavior (standard deviation, inter-quartile range, kurtosis)



- Perhaps the most efficient and easy to implement is the mean absolute phase deviation:

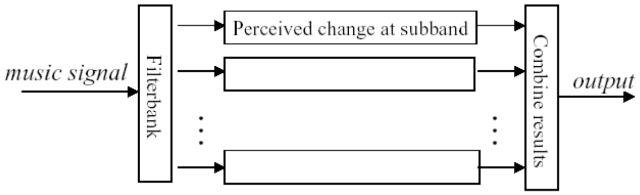
$$\eta_p(n) = \frac{2}{N}\sum_{k=0}^{N/2}\left|d_\varphi(n,k)\right|$$

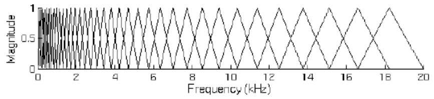
Choice of detection function



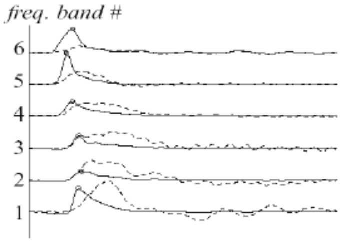
Further improvements

- It has been shown that sub-band decompositions bring benefits as events from independent bands do not mask each other
- A good example is the work by Klapuri et al. (1999)

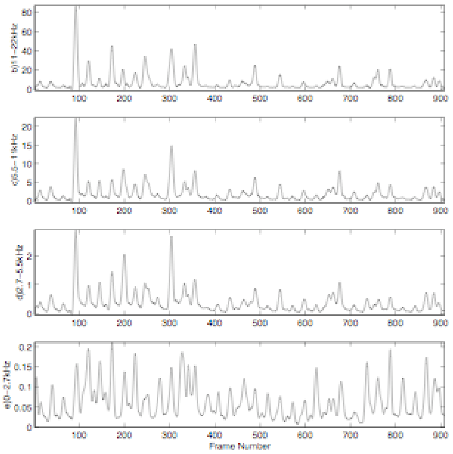




36 triangular windows, equally distributed in the mel-frequency scale



Why sub-bands?



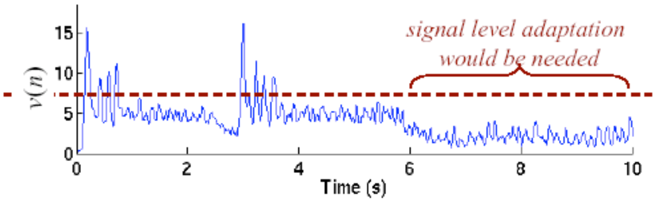
High sub-bands (better localization, prone to noise and miss-detection of tonal onsets)

Low sub-bands (robust to noise, high accuracy, poor resolution, poor localization)

Post-processing and peak-picking

- Post-processing facilitates peak-picking
- Examples include smoothing, normalization, DC-removal, differentiation, etc

- Peaks above a threshold are considered as onsets.
- This threshold can be fixed, however it is hard to choose a value that will operate in all signals (or even just in a whole song)



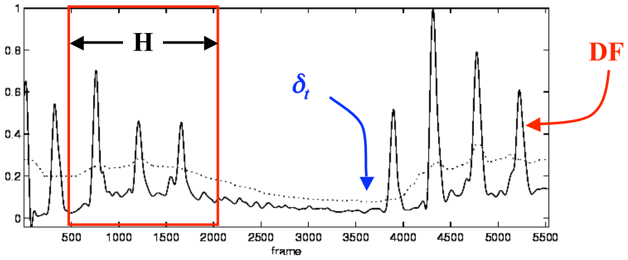
Post-processing and peak-picking

- Adaptive thresholding is a more realistic option for real signals.
- Methods include LPF, non-linear functions and percentiles (e.g. the median)

$$\delta_t(m) = \alpha + \beta \cdot median\left(DF\left(k_m\right)\right), k_m \in \left[m - \frac{H}{2}, m + \frac{H}{2}\right]$$

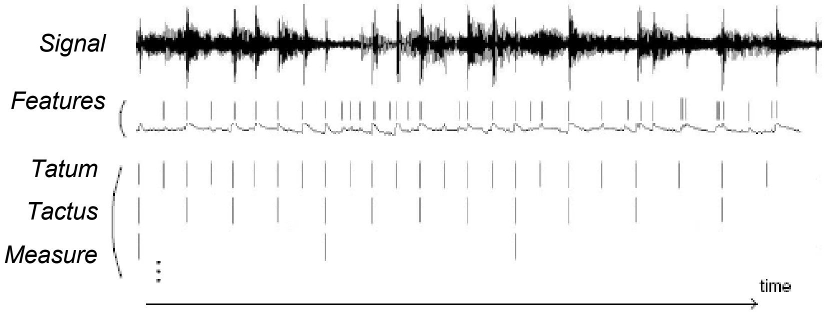
Offset value

Weighting value



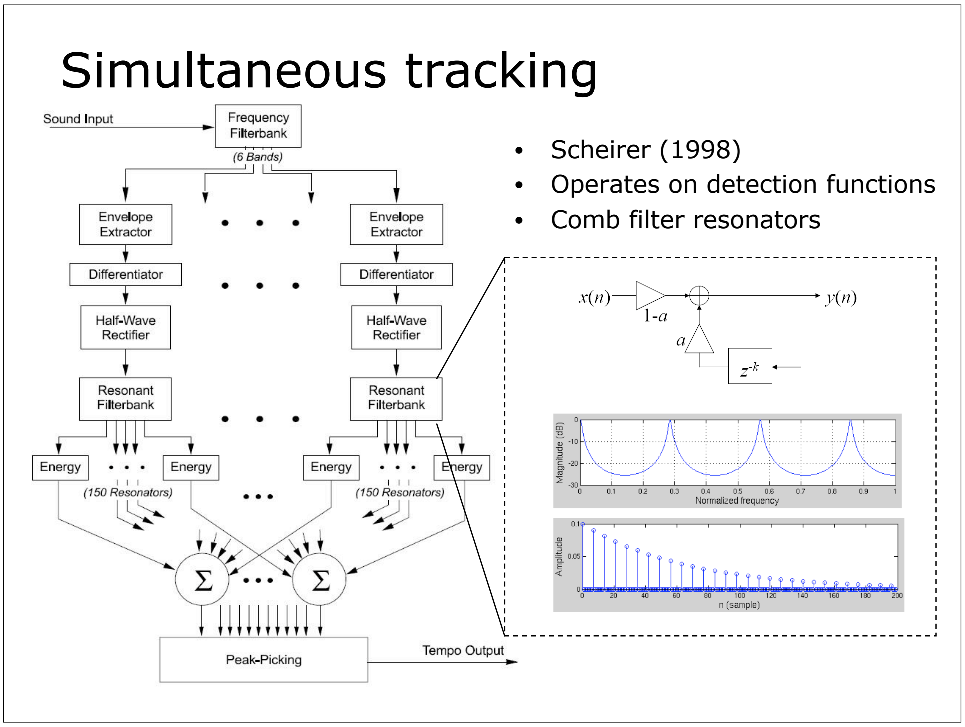
Understanding rhythm

- Goal: to detect moments of musical stress and process them as to uncover the underlying temporal regularities of the signal.
- It is hierarchical in structure, related to the perception of pulses at different time scales (From Gouyon, 2005):



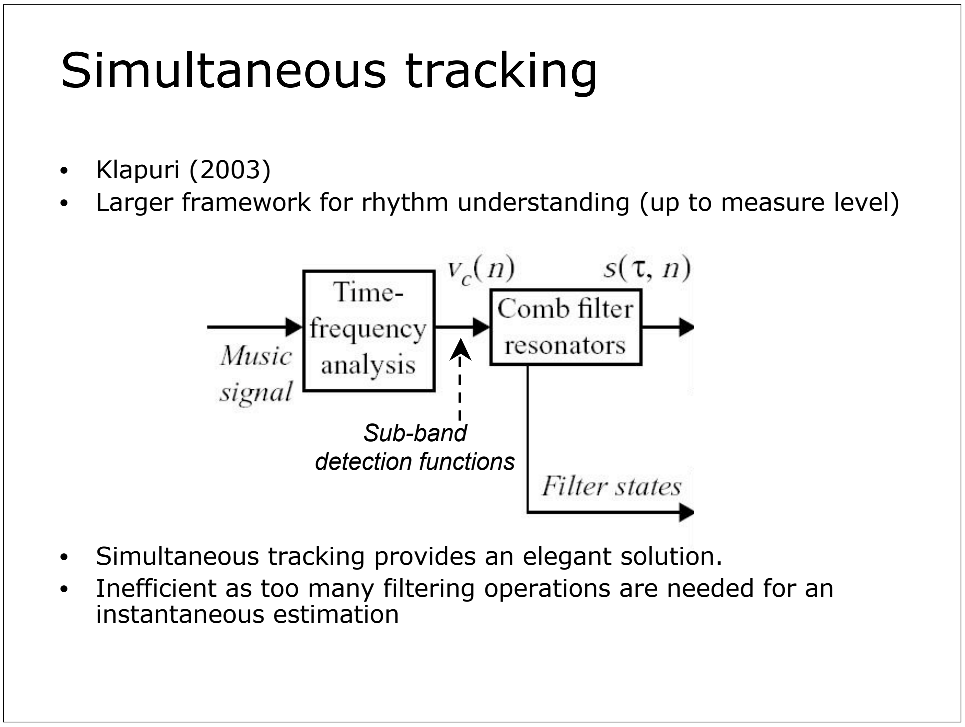
Tempo

- Tempo refers to the pace of a piece of music and is usually given in beats per minutes (BPM)
- We can think of it as a global quality but more realistically it is an evolving characteristic of musical performances.
- Thus, in computational terms we differentiate between tempo estimation and tempo (beat) tracking.
- In tracking, beats are not only described by their rate (frequency) but by their phase (time location).
- Many approaches have been proposed: Goto 97, Scheirer 98, Dixon 01, Tzanetakis 01, Gouyon 02, Klapuri 03, Davies 05, etc. (see MIREX 2004, 2005)
- They roughly divide between those that simultaneously estimate periodicity and phase and those that do it sequentially



Simultaneous tracking

- Scheirer (1998)
- Operates on detection functions
- Comb filter resonators



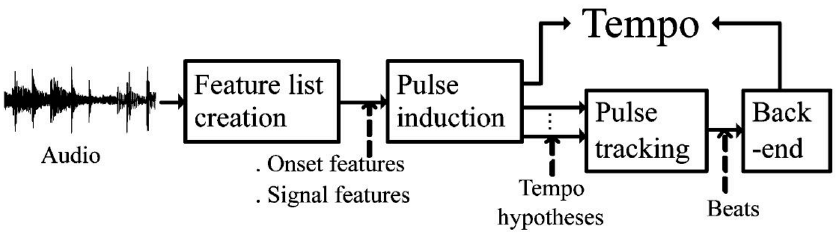
Simultaneous tracking

- Klapuri (2003)
- Larger framework for rhythm understanding (up to measure level)

- Simultaneous tracking provides an elegant solution.
- Inefficient as too many filtering operations are needed for an instantaneous estimation

Sequential tracking

- Periodicity and phase estimation are performed sequentially, thus separating tempo estimation from tracking
- Examples include Dixon (2001), Gouyon (2002) and Davies (2005)

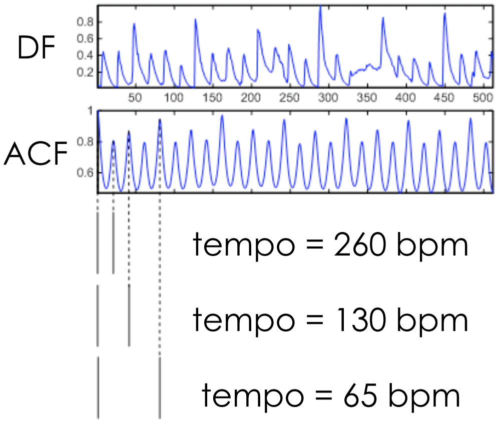


- Feature sets differ: onsets, inter-onset intervals, low-level features within segments, detection functions, etc
- Separating the tasks allow you to select different feature sets that maximize results

Sequential tracking

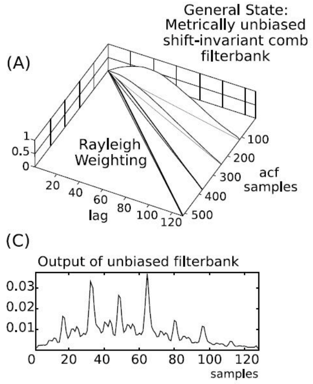
- For example, the autocorrelation sequence of the detection function is better at characterizing periodicities
- However simple peak-picking in the ACF is not enough for tempo estimation

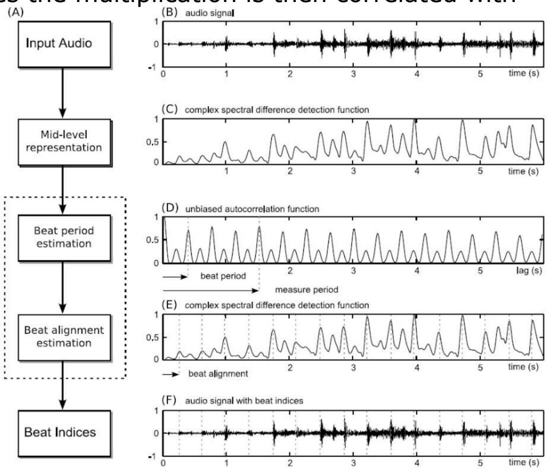
$$r(\tau) = \frac{1}{N} \sum_{n=0}^{N-\tau-1} x(n)x(n+\tau)$$



Sequential tracking

- Davies (2005) performs the dot multiplication of the ACF of the DF with a weighted comb filterbank.
- The filter that maximizes the multiplication is then correlated with the DF to obtain phase





(A)
Input Audio
Mid-level representation
Beat period estimation
Beat alignment estimation
Beat Indices
(B) audio signal
(C) complex spectral difference detection function
(D) unbiased autocorrelation function
beat period
measure period
(E) complex spectral difference detection function
beat alignment
(F) audio signal with beat indices
time (s)
lag (s)

References

- Bello , J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M. and Sandler, M.B. A tutorial on onset detection in music signals. IEEE Transactions on Speech and Audio Processing. 13(5), Part 2, pages 1035-1047, September, 2005.
- Klapuri. " Sound Onset Detection by Applying PsychoacousticKnowledge," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Phoenix, Arizona, 1999.
- Eric D. Scheirer. “Tempo and beat analysis of acoustic musical signals”, Journal of the Acoustical Society of America, January, 1998
- S. Dixon. Automatic Extraction of Tempo and Beat from Expressive Performances. Journal of New Music Research, 30 (1), 2001, pp 39-58.
- M. E. P. Davies and M. D. Plumbley. Beat Tracking With A Two State Model. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2005), Vol. 3, pp241-244 Philadelphia, USA, March 19-23, 2005
- Gouyon, F. A computational approach to rhythm description --- Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing. Ph.D. Thesis. UPF, Spain, 2005. http://www.iua.upf.edu/mtg/publications/9d0455-PhD-Gouyon.pdf
- Klapuri, A. " Musical meter estimation and music transcription ". Paper presented at the Cambridge Music Processing Colloquium, Cambridge University, UK, 2003.