Audio Content Description

Mathieu Lagrange 

October 17, 2018

Course website: https://github.com/mathieulagrange/musnu

Outline

Outline

Outline

Outline

Outline

Outline

Perceptual vs. physical properties

1. Loudness ⇔ Amplitude
2. Pitch ⇔ Fundamental frequency
3. Timbre ⇔ spectrum and envelope ??

Perceptual vs. physical properties

1. Loudness ⇔ Amplitude

Definition

Loudness is that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud

2. Pitch ⇔ Fundamental frequency
3. Timbre ⇔ spectrum and envelope ??

Perceptual vs. physical properties

1. Loudness ⇔ Amplitude
2. Pitch ⇔ Fundamental frequency

Definition

Pitch is a perceptual property that allows the ordering of sounds on a frequency-related scale

3. Timbre ⇔ spectrum and envelope ??

Perceptual vs. physical properties

1. Loudness ⇔ Amplitude
2. Pitch ⇔ Fundamental frequency
3. Timbre ⇔ spectrum and envelope ??

Definition

Timbre is what makes a particular musical sound different from another, even when they have the same pitch and loudness

Musical properties

Music is about organization of sounds in time

- Dynamics: *pp* to *ff*
- Duration: ♪ 𝅗𝅥 𝅝
- Rhythm: ♫

Definition

rhythm is the timing of events on a human scale

- Beat

Definition

Beat is the rhythm listeners would tap their toes to when listening to a piece of music

Timbre

Five major acoustic parameters

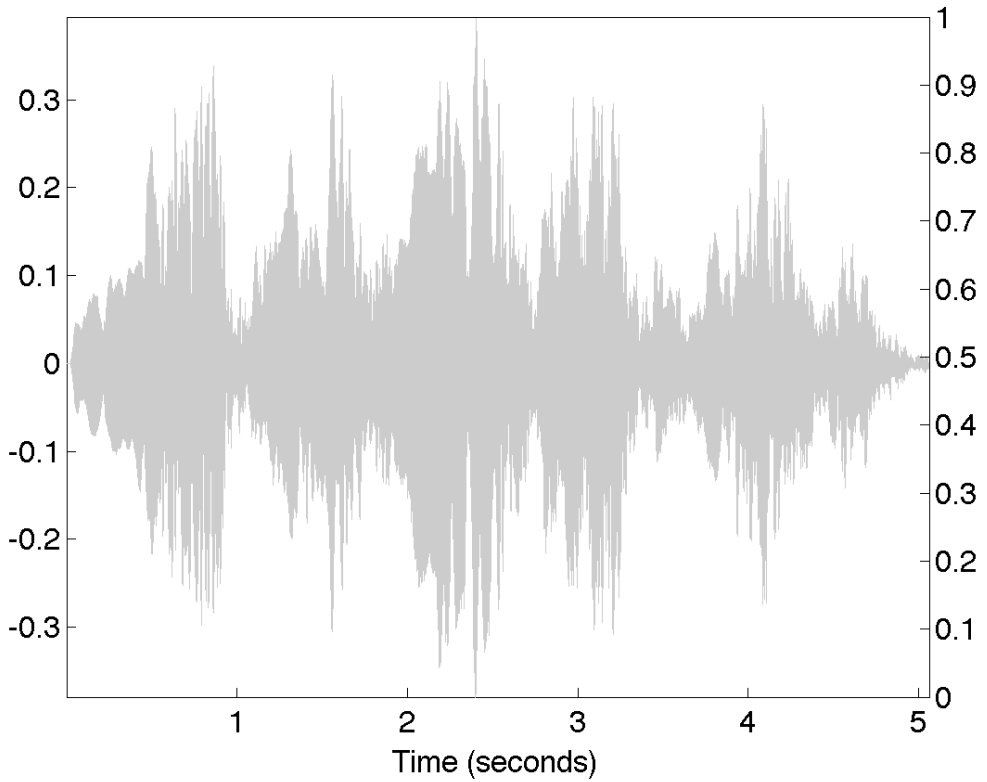
1. The range between tonal and noiselike character
2. The spectral envelope
3. The time envelope in terms of rise, duration, and decay
4. The changes both of spectral envelope (formant-glide) and fundamental frequency (micro-intonation)
5. The prefix, or onset of a sound, quite dissimilar to the ensuing lasting vibration

---

Schouten & al The perception of timbre 1968

Female singing voice

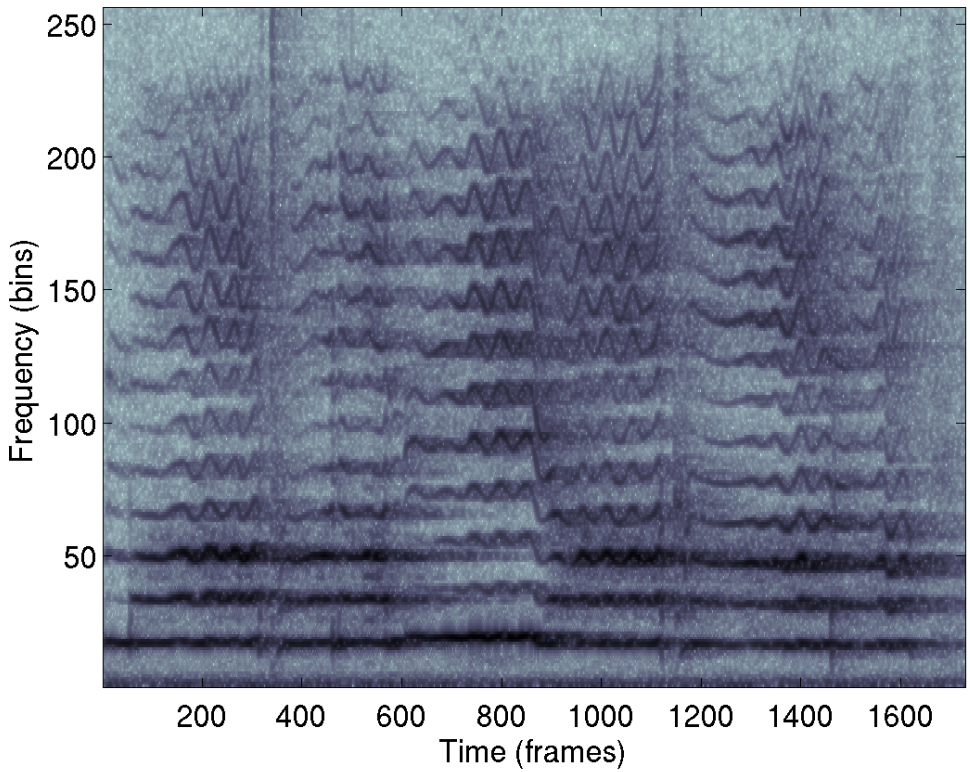
- waveform



- spectrogram
- both

Female singing voice

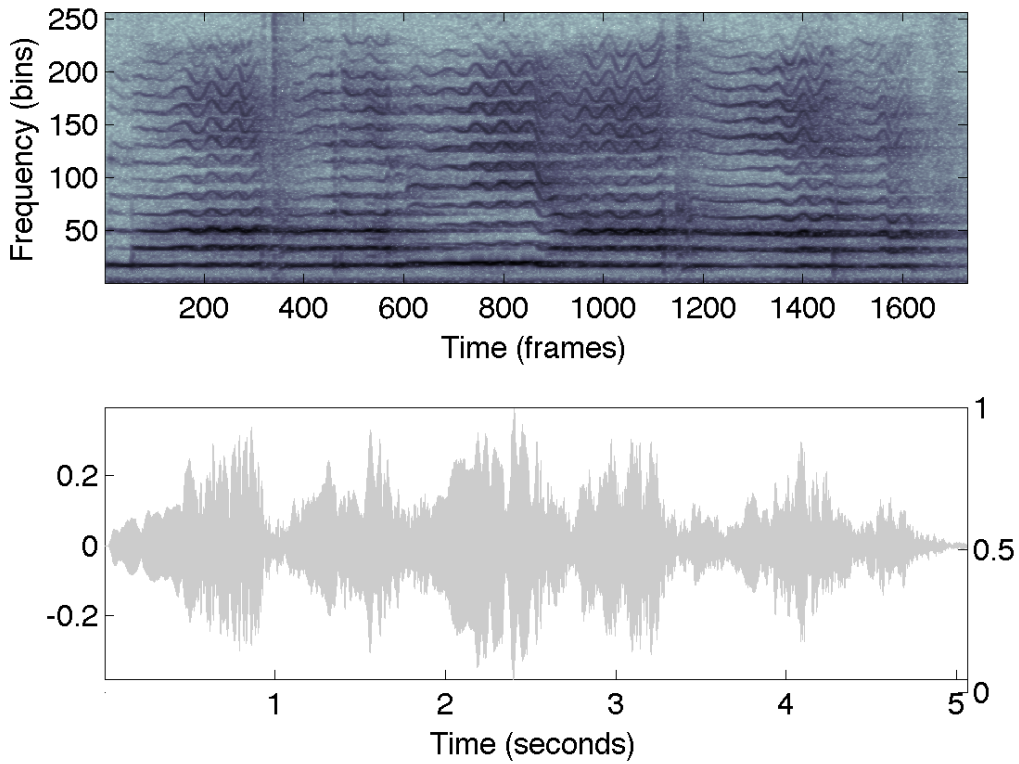
- waveform
- spectrogram



- both

Female singing voice

- waveform
- spectrogram
- both



Loudness

Loudness is the perceptual attribute of sound that correlate with its physical strength

- a useful scale is the deciBels:

$$v_{dB} = 20log_{10}\left(\frac{v(n)}{v_0}\right)$$

- v_0 is a reference constant, set to 1 when $-1 < v_n < 1$

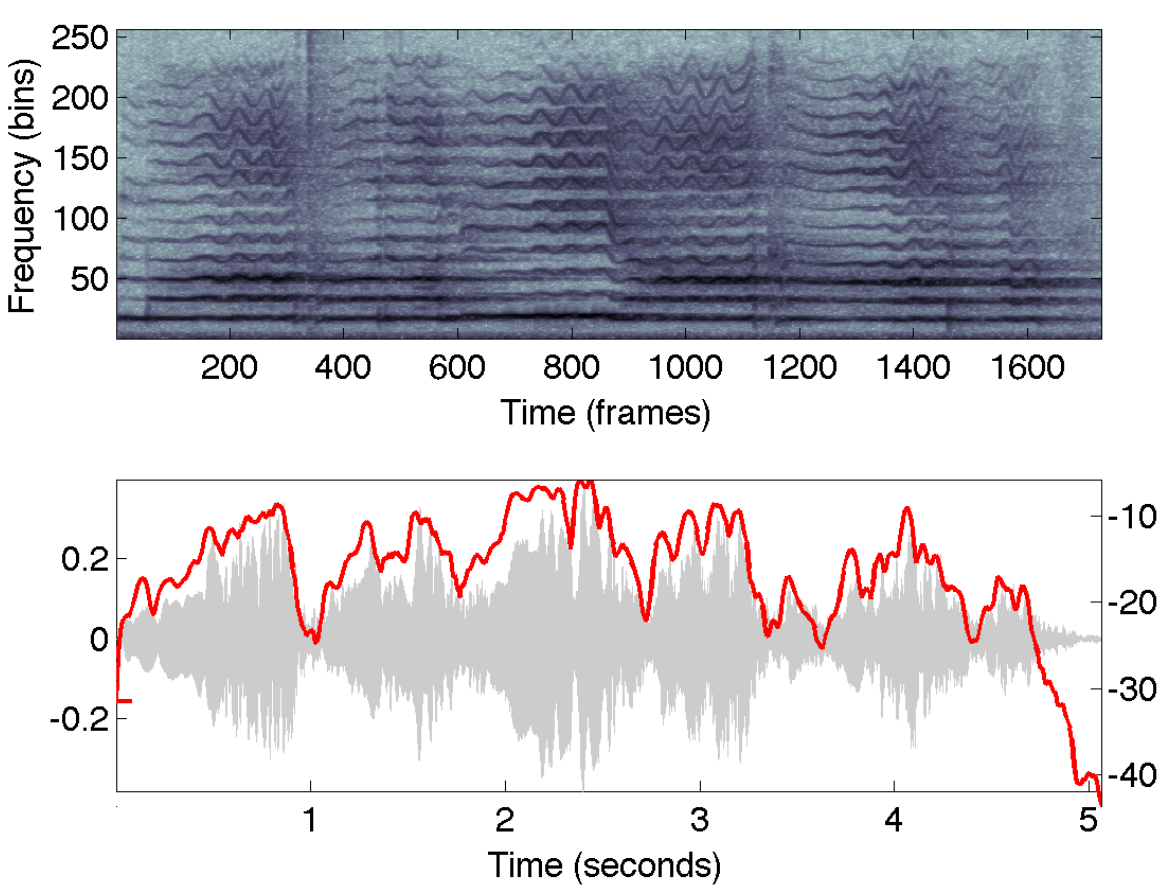
Root Mean Square (RMS)

Definition

The Root Mean Square (RMS) is defined as

$$v(n) = \sqrt{\frac{1}{N}\sum x(i)^2}$$

Root Mean Square (RMS)



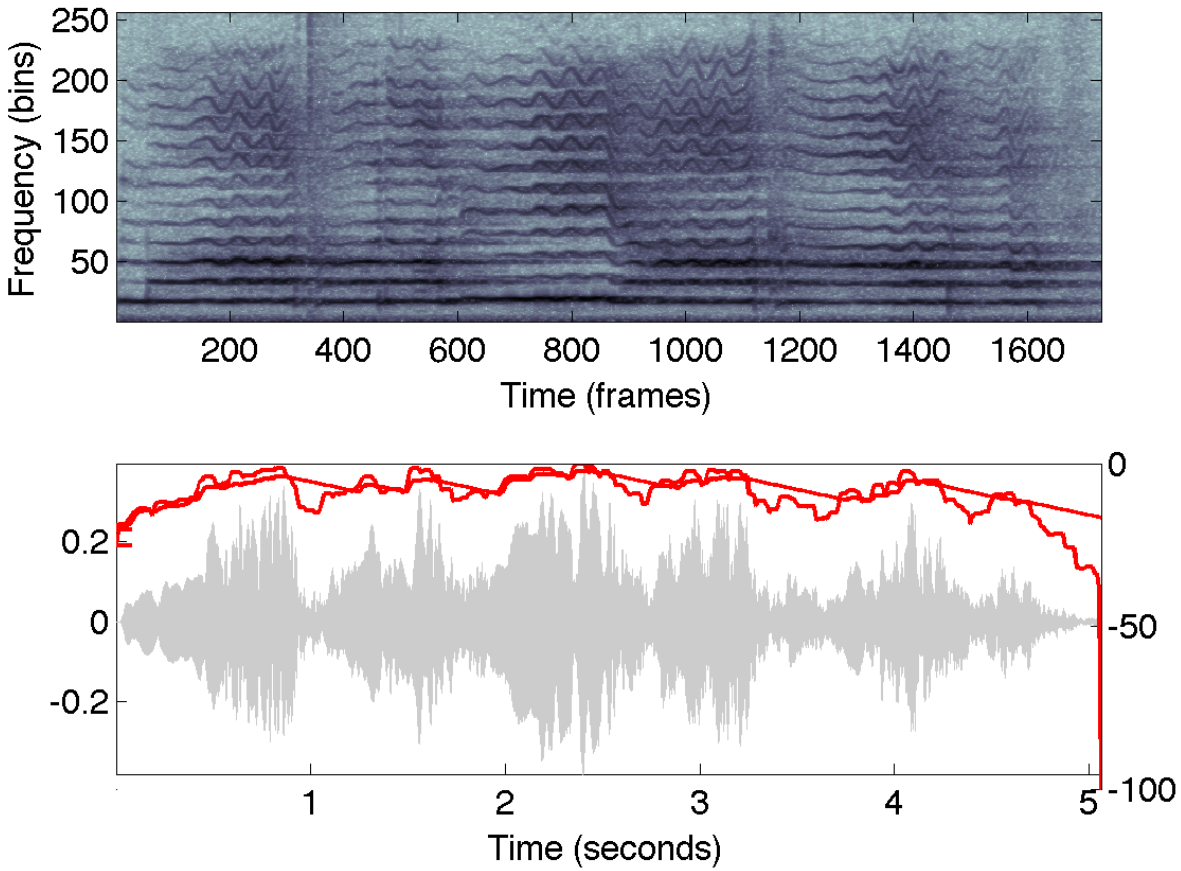
- computation per blocks of hundred of milliseconds
- low values ⇔ low volume
- large values ⇔ high volume

Peak Envelope

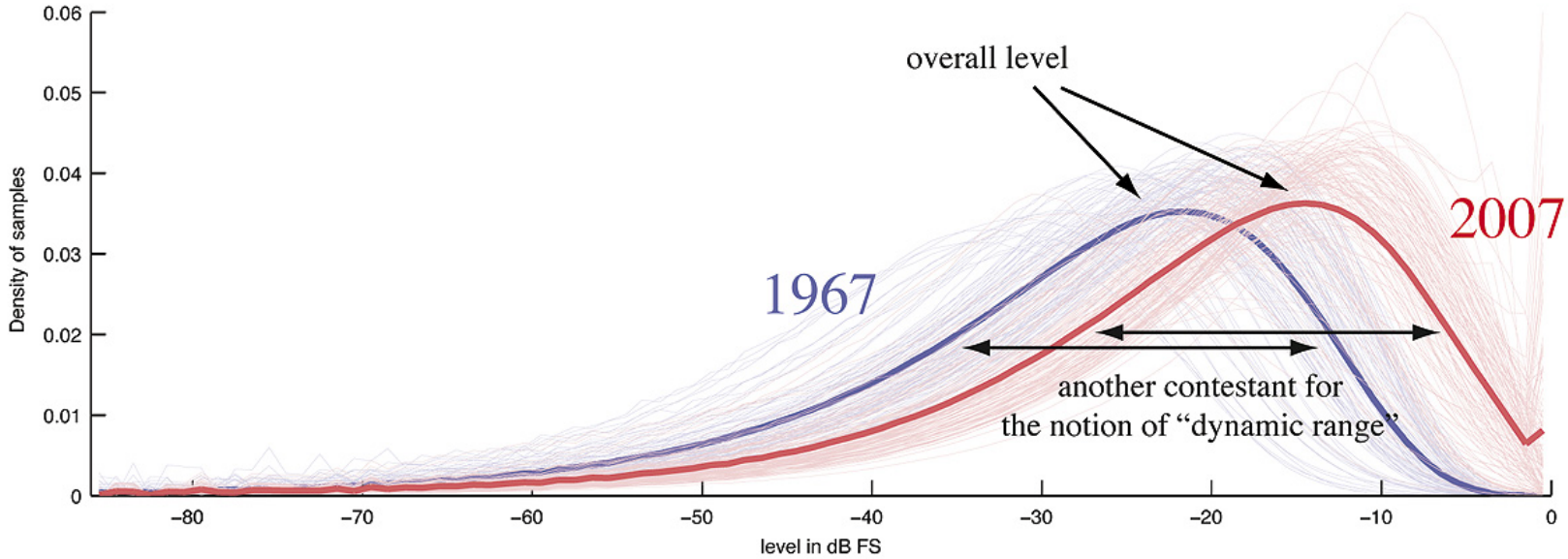
Definition

The Peak Envelope is defined as the maximal value within a given block of samples.

Peak Envelope



The loudness war



Comparisons between signal levels and picture levels as defined in Photoshop result in another interpretation of the loudness war.

Emmanuel Deruty Loudness study of 4.5K tracks

Pre-processing

- Down mixing: average all channels contribution
- DC removal: subtract arithmetic mean
- Normalization: set (maximal / averaged) amplitude to a predefined value

Pre-processing

- Down mixing: average all channels contribution

$$x(i) = \frac{1}{C} \sum_{c=1}^{C} x_c(i)$$

- DC removal: subtract arithmetic mean
- Normalization: set (maximal / averaged) amplitude to a predefined value

Pre-processing

- Down mixing: average all channels contribution
- DC removal: subtract arithmetic mean

$$x(i) = x(i) - \frac{1}{N}\sum_{n=1}^{N} x(n)$$

- Normalization: set (maximal / averaged) amplitude to a predefined value

Pre-processing

- Down mixing: average all channels contribution
- DC removal: subtract arithmetic mean
- Normalization: set (maximal / averaged) amplitude to a predefined value

$$x(i) = \frac{x(i)}{max(|x(i)|)}$$

Pre-processing

- Down mixing: average all channels contribution
- DC removal: subtract arithmetic mean
- Normalization: set (maximal / averaged) amplitude to a predefined value

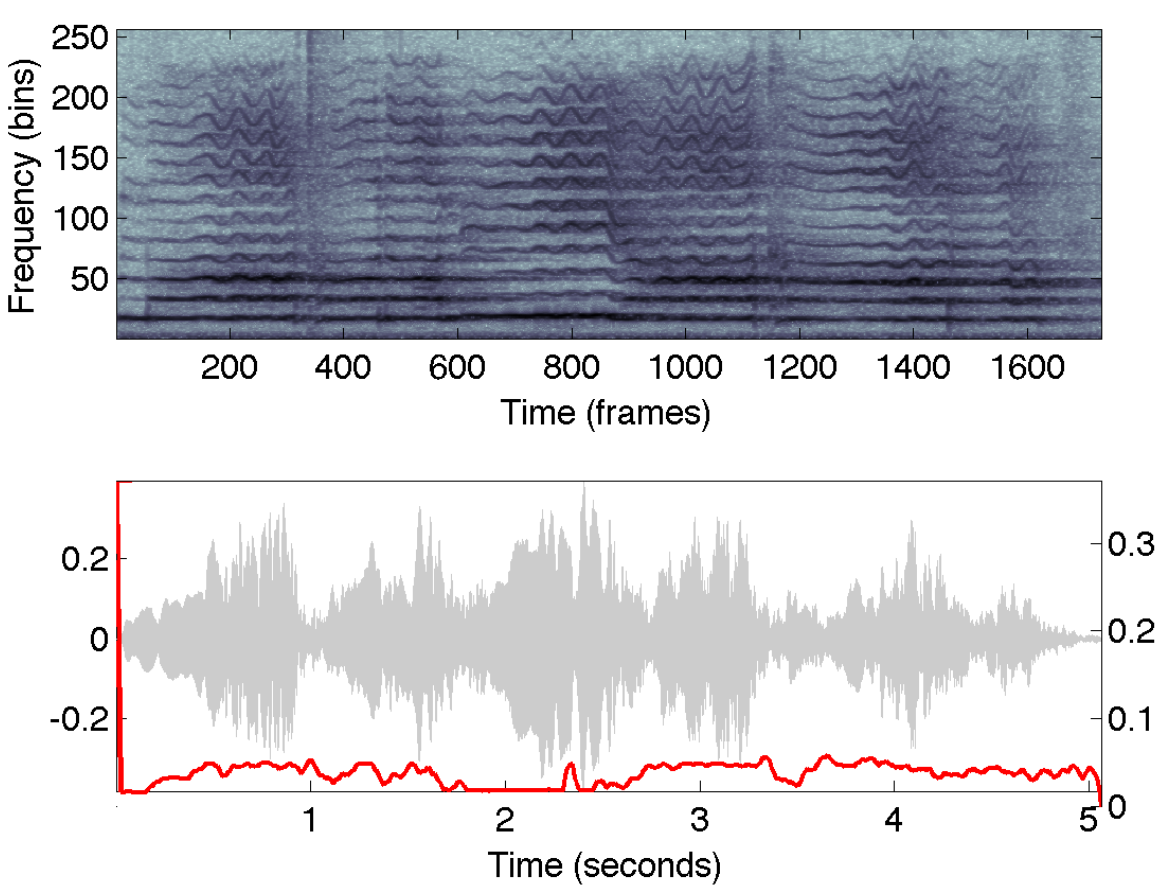
$$x(i) = \frac{x(i)}{\sqrt{\sum_{n=1}^{N} x(n)^2}}$$

The zero crossing rate (ZCR)

Definition

The number of changes of sign of audio samples in consecutive block.

The zero crossing rate (ZCR)



- low values ⇔ low frequency content
- large values ⇔ high frequency content
- large value change across blocks indicates noisy content

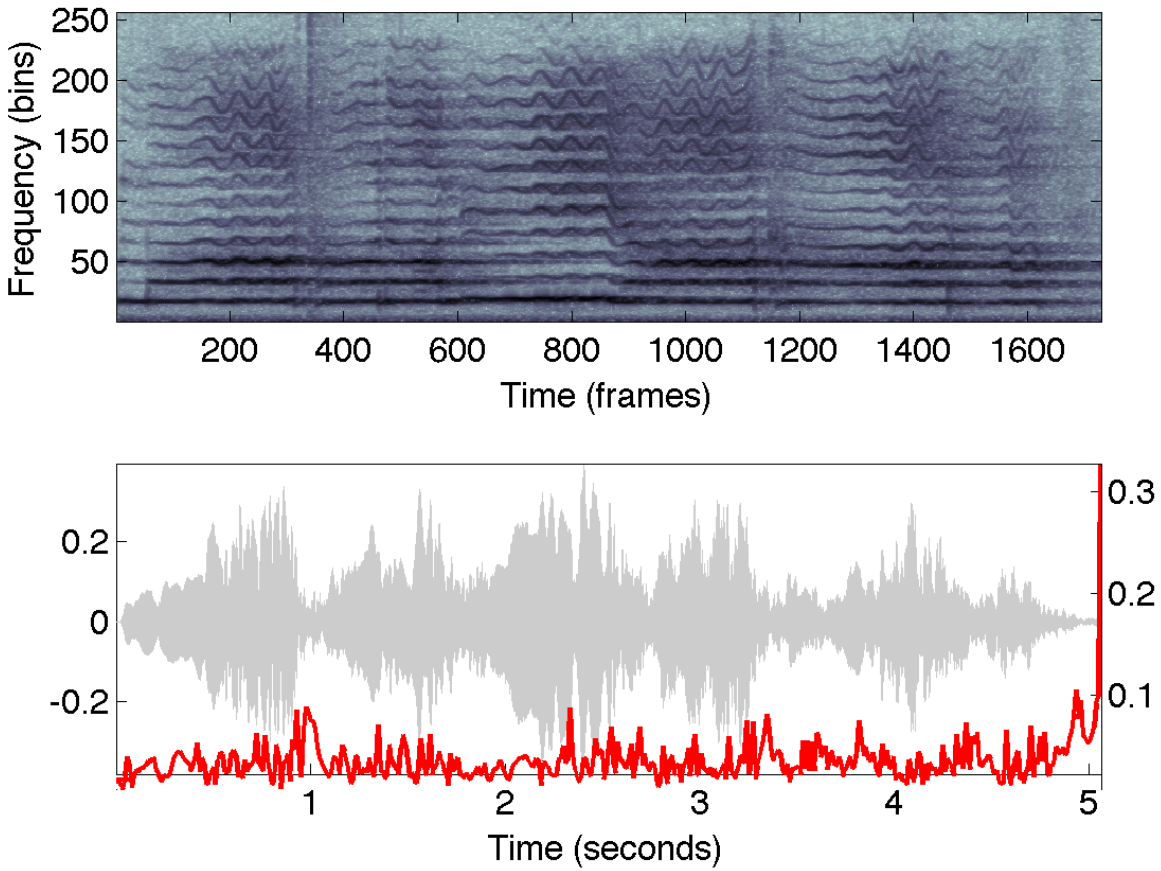
The predictivity ratio

Definition

The predictivity ratio is a measure of how well the audio signal can be predicted by an K-order linear prediction:

$$\hat{x}(i) = \sum_{k=1}^{K} b_l x(i-k)$$

The predictivity ratio



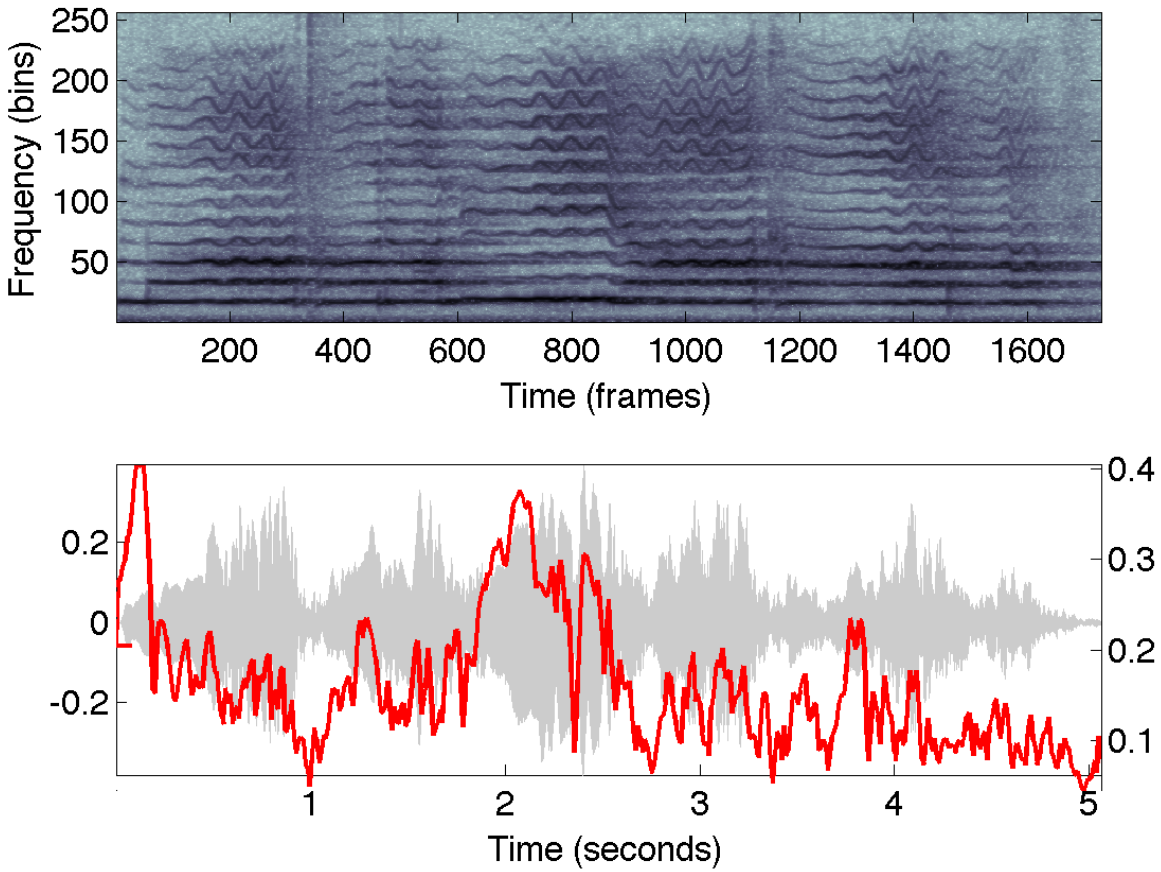
- low values ⇔ tonal signal
- large values ⇔ impredictible signals (noise)

The spectral crest factor

Definition

The spectral crest factor is the ratio of the maximum in the magnitude spectrum with the sum of the magnitude spectral bins.

The spectral crest factor



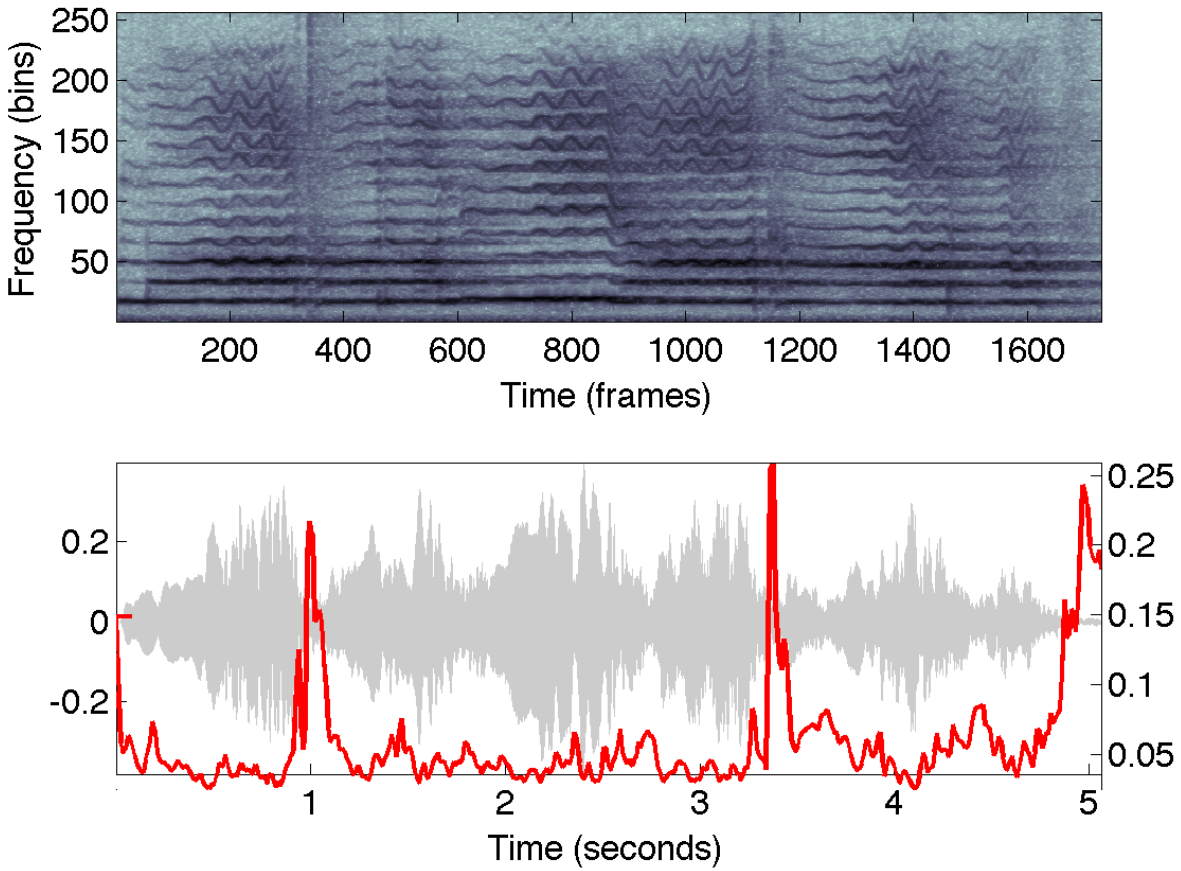
- low values ⇔ flat magnitude spectrum
- large values ⇔ sinusoids

The spectral flatness

Definition

The spectral flatness is the ratio of geometric mean and arithmetic mean of the magnitude spectrum.

The spectral flatness



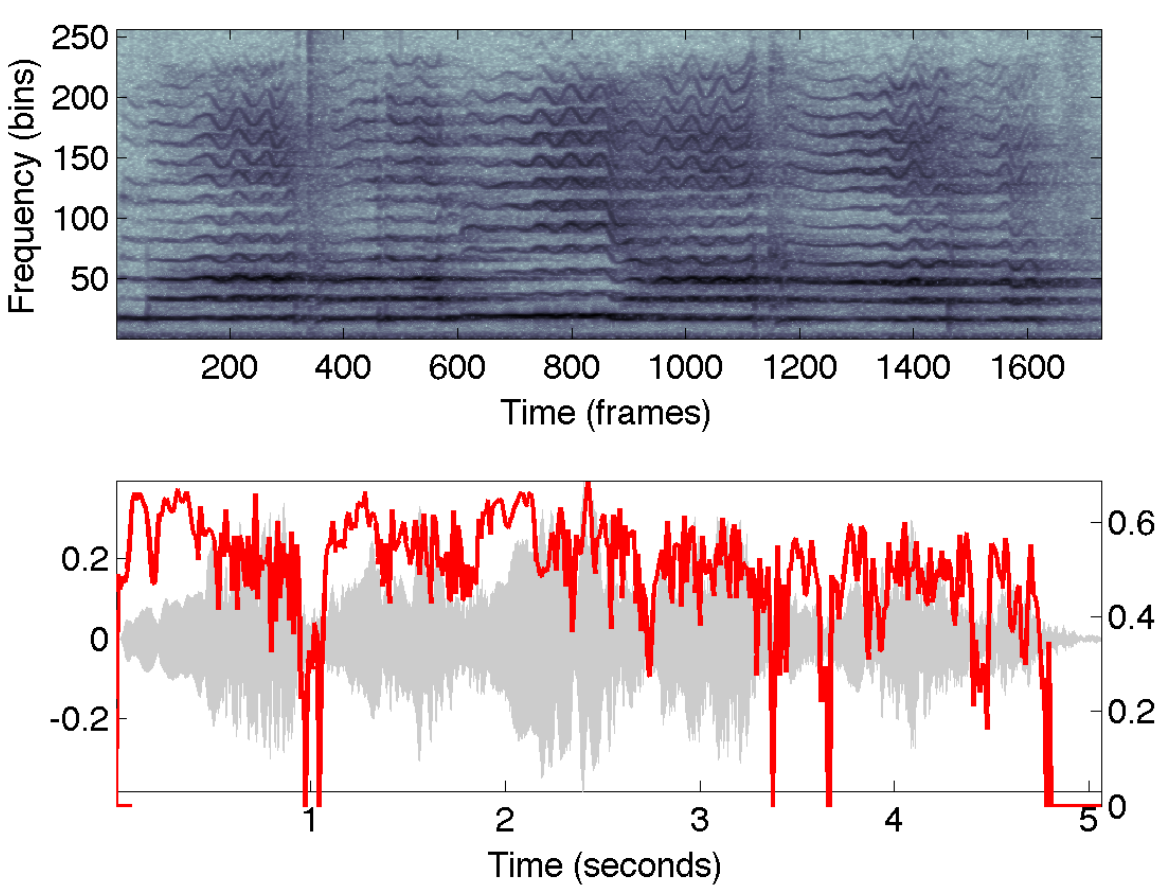
- low values ⇔ possibly tonal components
- large values ⇔ flat (noisy) spectrum

The tonal power ratio

Definition

The tonal power ratio is the ratio between the tonal power and overall power.

The tonal power ratio



Tonal components are

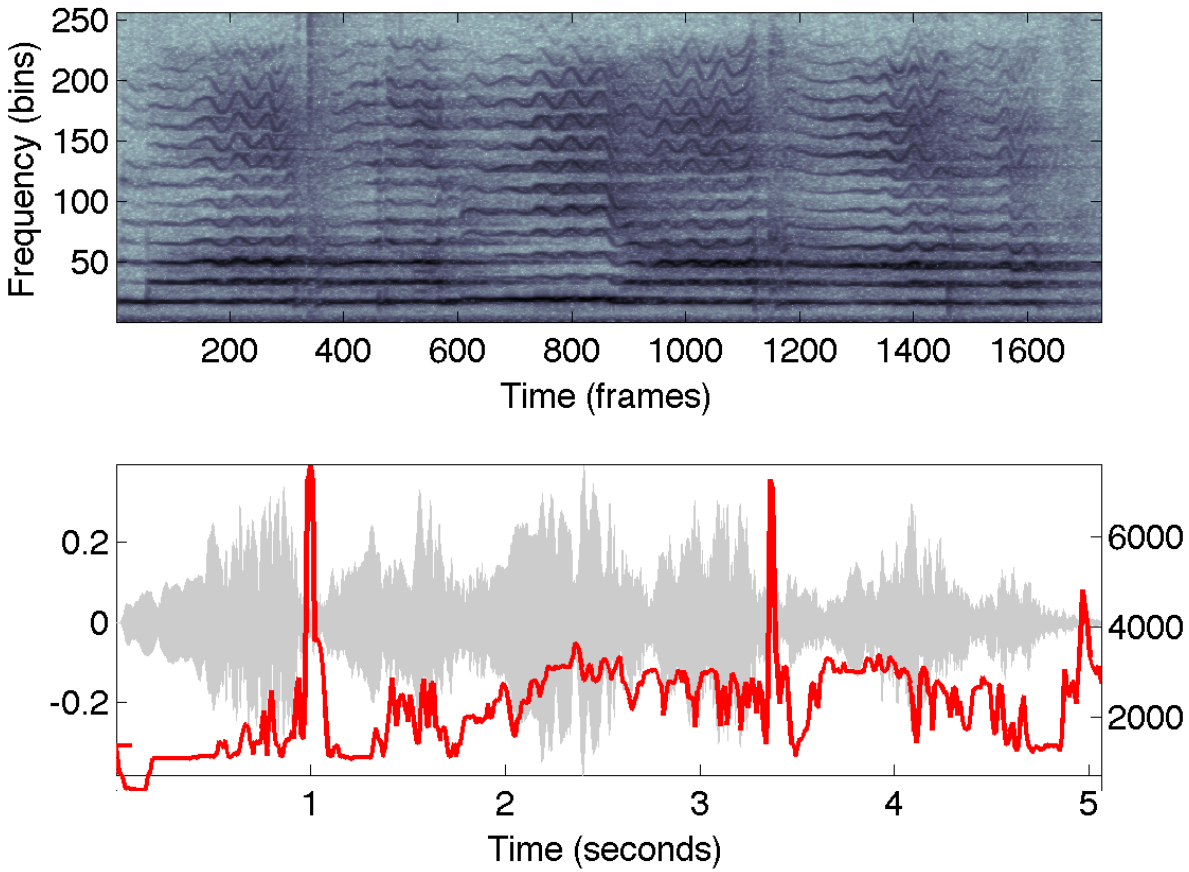
- local maxima: $|X(k-1,n)|^2 < |X(k,n)|^2 > |X(k+1,n)|^2$
- powerful: magnitude above a given threshold

The spectral rolloff

Definition

The frequency bin below which the accumulated magnitude of the spectrum reach a certain percentage of the overall magnitude spectrum.

The spectral rolloff



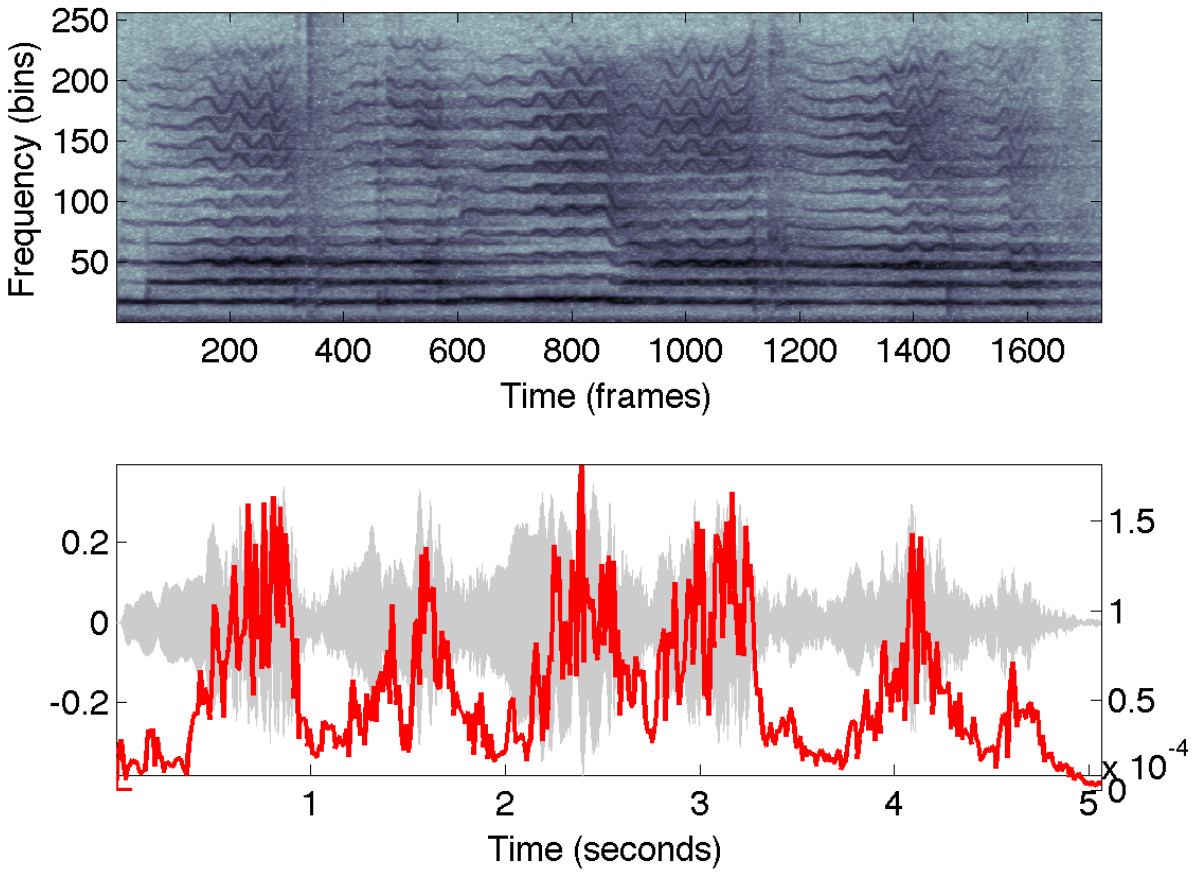
- low values ⇔ low audio bandwidth
- large values ⇔ noise

The spectral flux (roughness)

Definition

The spectral flux is the average difference between consecutive magnitude bins.

The spectral flux (roughness)



- low values ⇔ steady state signals or low input levels
- large values ⇔ abrupt changes

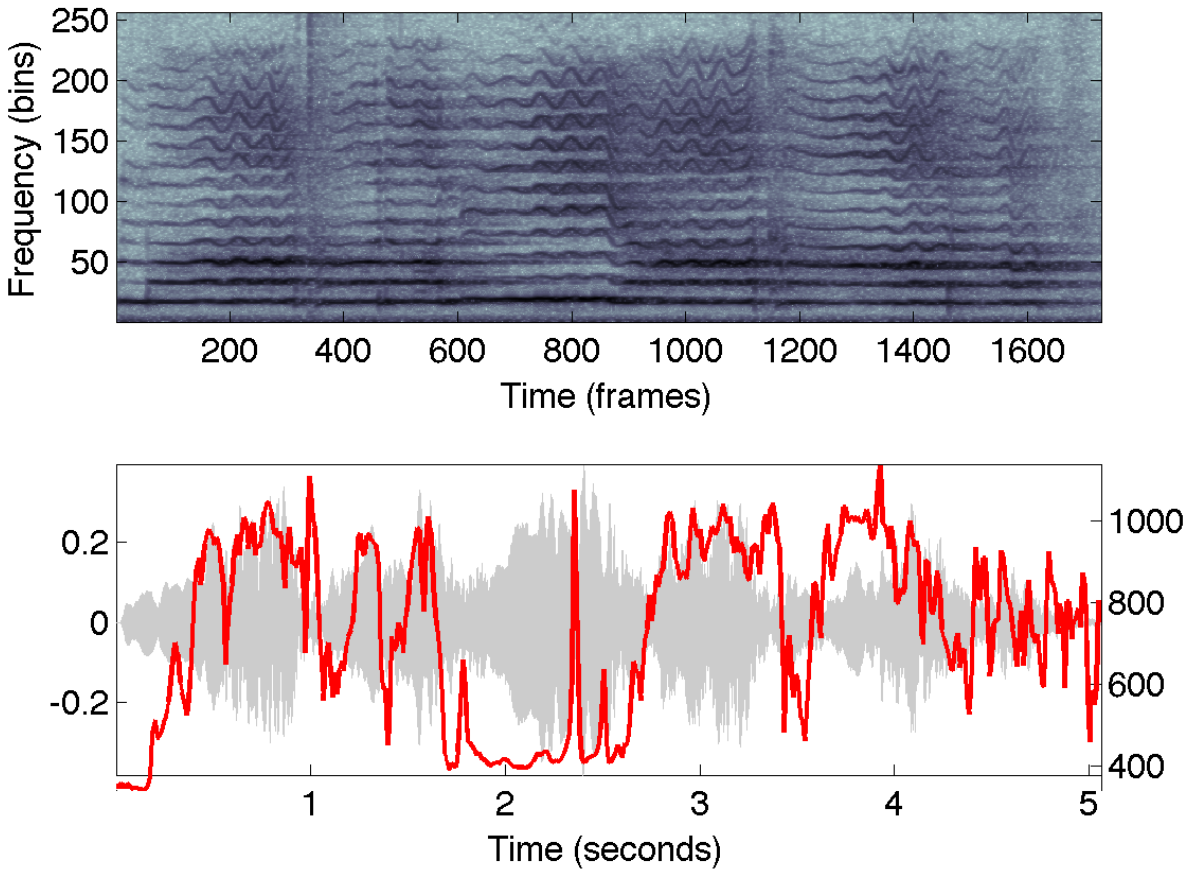
The spectral centroid (brightness / sharpness)

Definition

The spectral centroid is defined as the Center of Gravity (COG) of the spectral energy:

$$sc(n) = \frac{\sum_{k=1}^{K} k|X(k,n)|^2}{\sum_{k=1}^{K} |X(k,n)|^2}$$

The spectral centroid (brightness / sharpness)



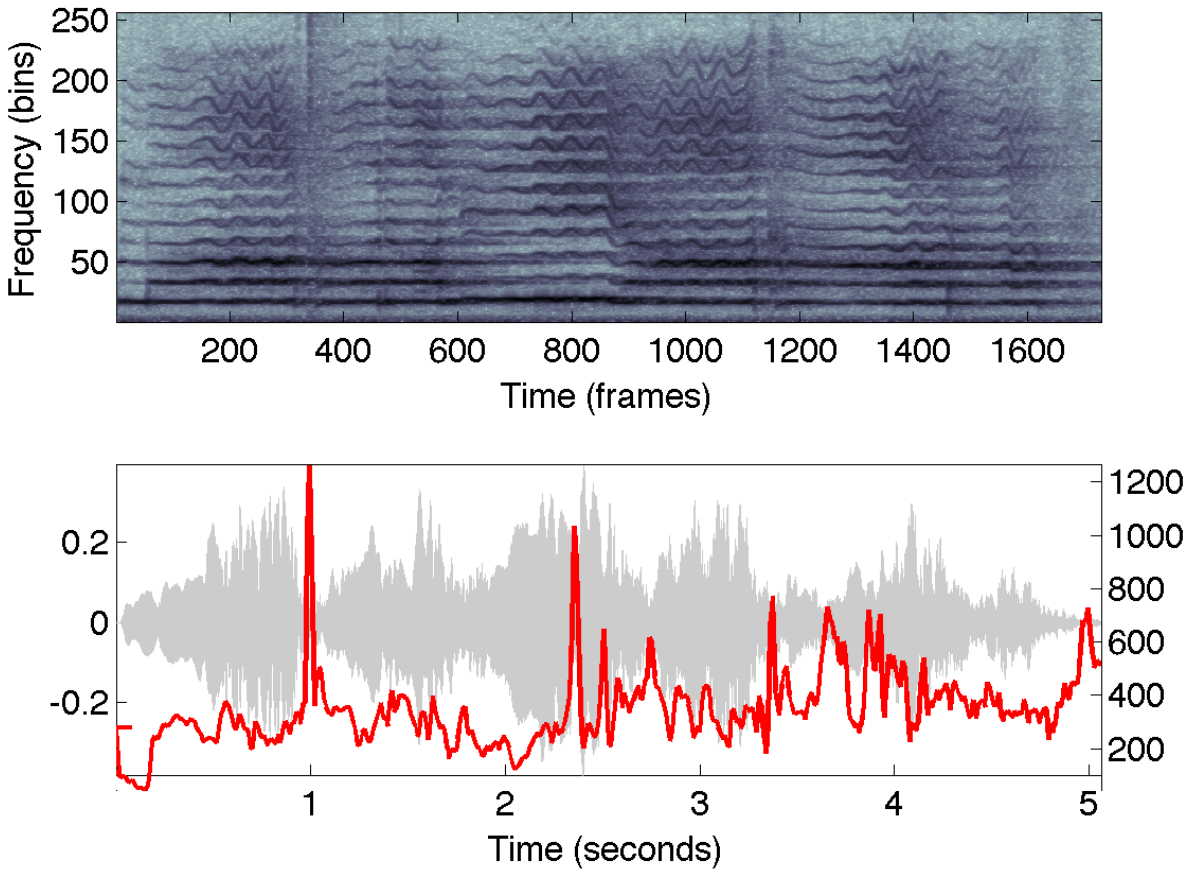
- low values ⇔ dominant low frequencies
- large values ⇔ dominant high frequencies

The spectral spread (instantaneous bandwidth)

Definition

The spectral spread describes the concentration of the spectral energy around the spectral centroid.

The spectral spread (instantaneous bandwidth)



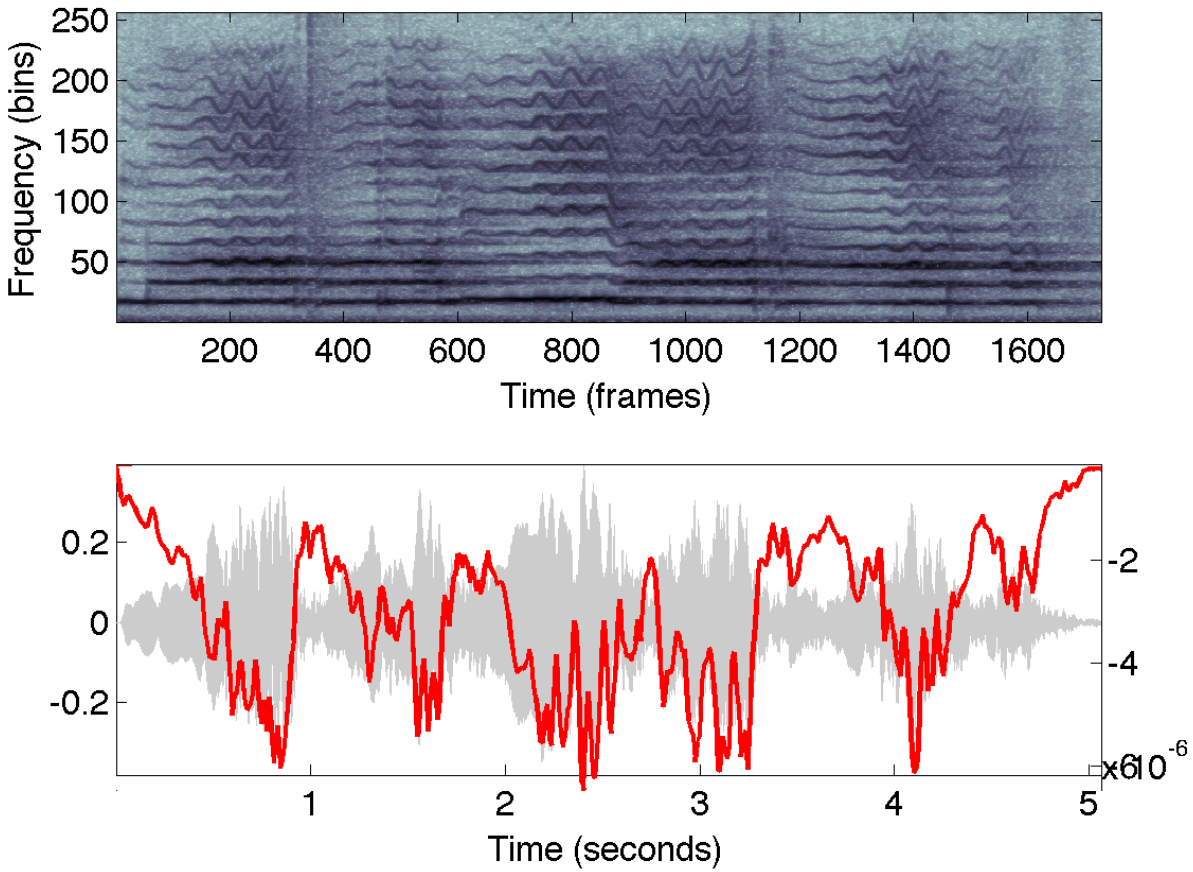
- low values ⇔ during notes (monophonic case)
- large values ⇔ during pauses and transients

The spectral slope

Definition

The spectral slope is computed from a linear regression over frequency of the magnitude spectrum.

The spectral slope



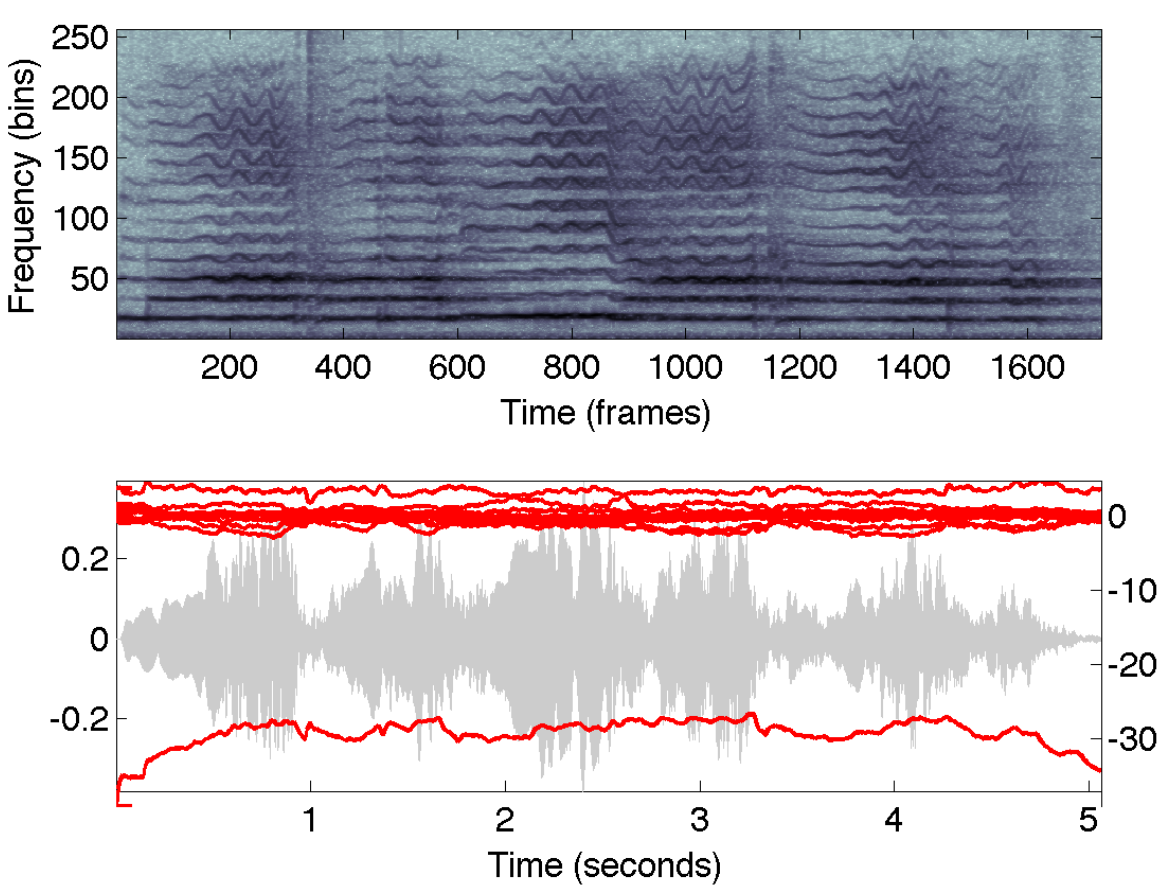
- low values ⇔ "flat" spectrum
- large values ⇔ few higher harmonics

Mel Frequency Cepstral Coefficients (MFCC)s

Definition

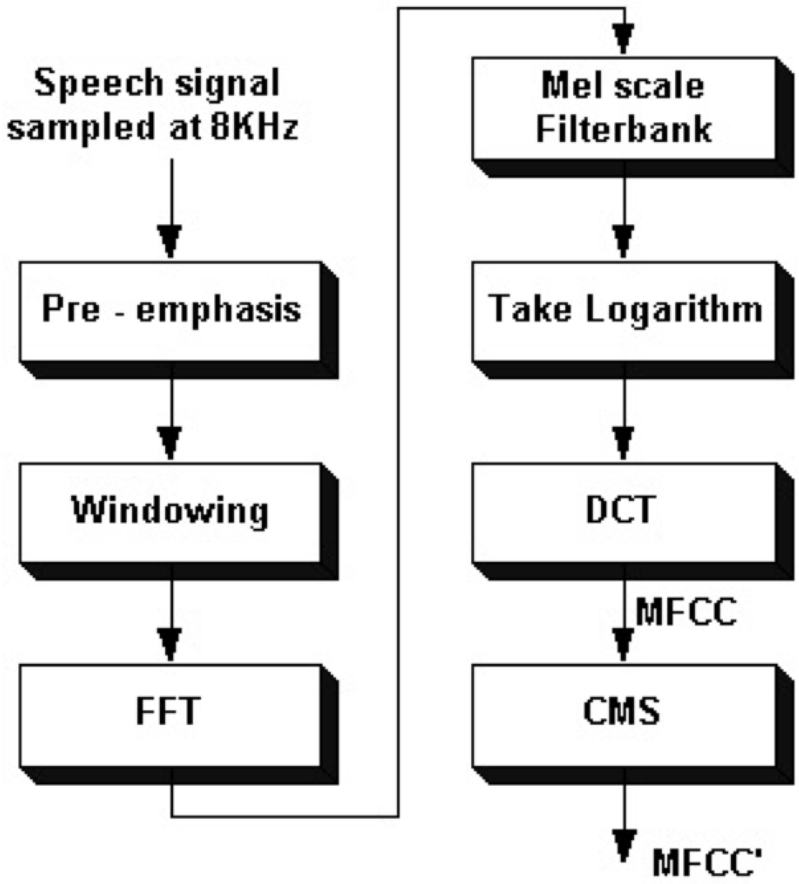
The Mel Frequency Cepstral Coefficients (MFCC)s can be seen as a compact description of the shape of the spectral envelope of an audio signal.

Mel Frequency Cepstral Coefficients (MFCC)s



- based on a mel-warped spectrum
- a logarithm of the magnitude spectrum
- apply a DCT on the resulting spectral representation

Mfccs


Speech signal sampled at 8KHz
Pre - emphasis
Windowing
FFT
Mel scale Filterbank
Take Logarithm
DCT
MFCC
CMS
MFCC'

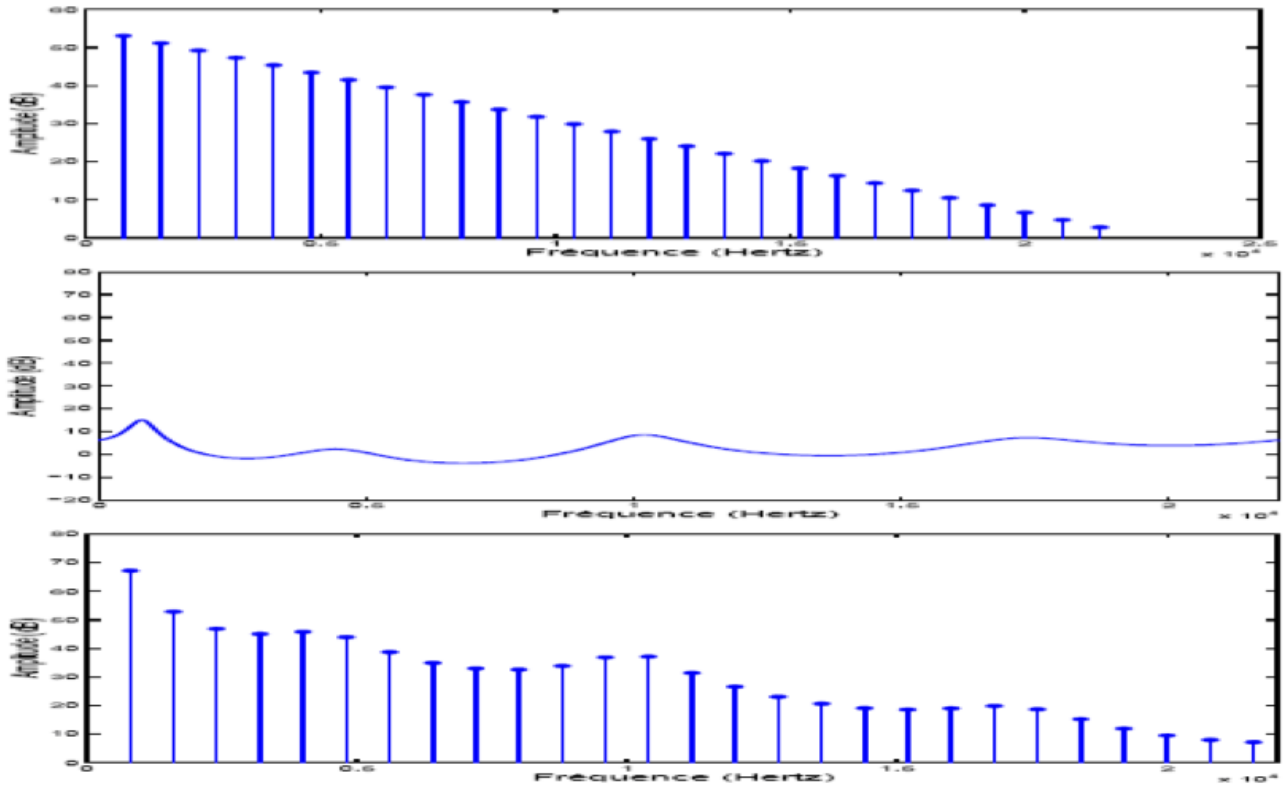
Mel Frequency Cepstral Coefficients (MFCC)s

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform (DCT)
5. The MFCCs are the amplitudes of the resulting spectrum.

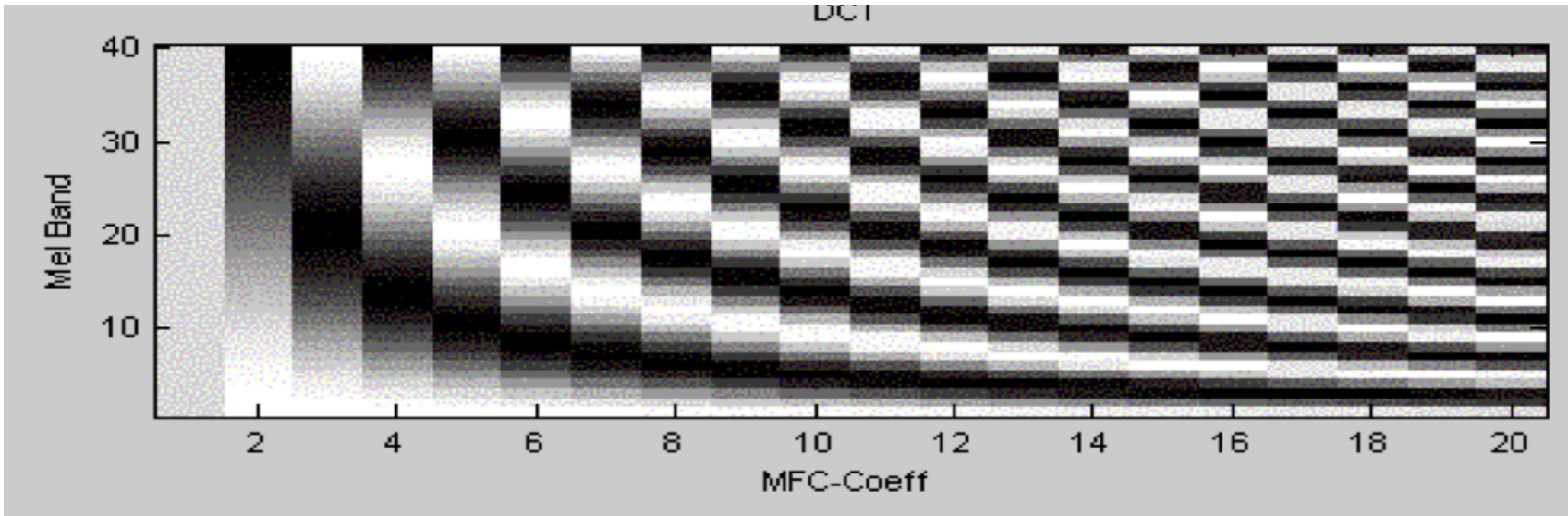
Mfccs : log spectrum

linear frequency log frequency

Mfccs : source / filter model



Mfccs : dct basis



Projecting the magnitude spectrum over this basis allows us to capture the envelope with the first resulting dct coefficients.

Pitch

Many tonal sounds produced by many musical instruments

- can be considered as a sum of sinusoidal components with frequencies that are harmonically related: $f_0, 2f_0, 3f_0, \cdots$
- then the fundamental frequency f_0 dominates the pitch perception
- mostly a relative perception as the fundamental frequency may be missing

Scales

The relation between the fundamental frequency and the perceived pitch is non-linear

- this non-linearity is tied to the frequency resolution of the human cochlea
- Mel scale: the scale is a measure of tone height. One analytic model:

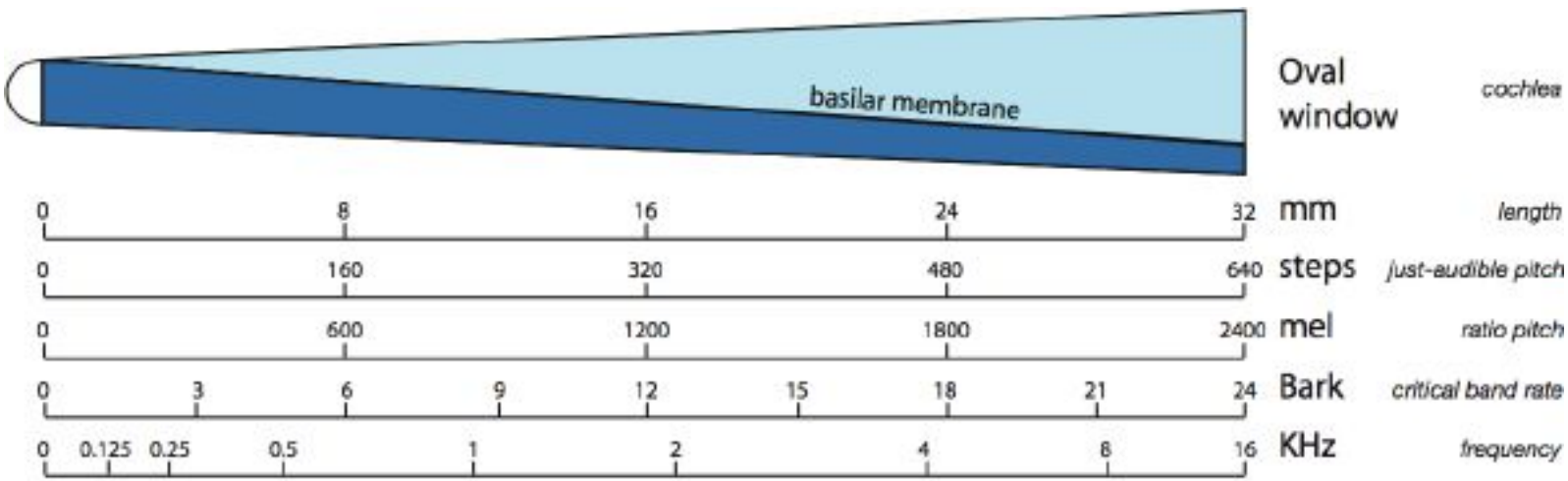
$$m_F(f) = 1000 log_2\left(1 + \frac{f}{1000}\right)$$

- Bark scale: the scale come from the bandwidth of measured critical bands:

$$b_F(f) = 7 \operatorname{arcsinh}\left(\frac{f}{650}\right)$$

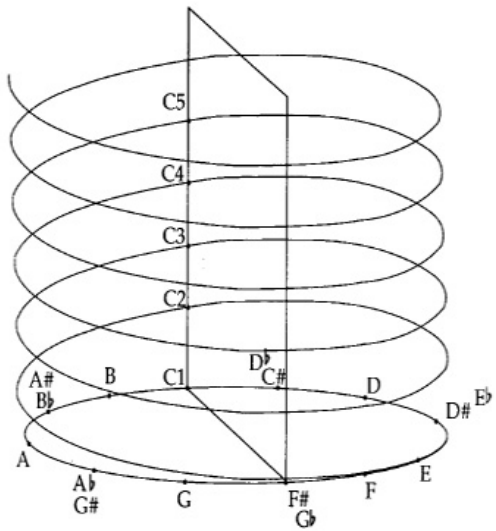
- the 2 scales are related by 1 bark = 100 mel

Scales



Physical relationships between the different scales

Chroma perception



Humans also tends to group pitches with ratios of 2.

The frequency of musical pitch

The MIDI standard

- uses the equal temperament which results in enharmonic equivalence
- each semi-tone has a distance of 1 to its nearest neighbor
- from MIDI pitch to frequency:

$$p(f) = 69 + 12\log_2\left(\frac{f}{f_{A4}}\right)$$

- from frequency to MIDI pitch:

$$f(p) = f_{A4}2^{\frac{p-69}{12}}$$

Tuning frequency

The tuning frequency f_{A4}

- is the frequency of the concert pitch or standard pitch
- is standardized internationally to 440 Hz
- that said the exact frequency used by musicians can vary due to the use historic instruments or timbre preferences

Tuning frequency

The tuning frequency f_{A4}

- is the frequency of the concert pitch or standard pitch
- is standardized internationally to 440 Hz
- that said the exact frequency used by musicians can vary due to the use historic instruments or timbre preferences

Tuning frequency

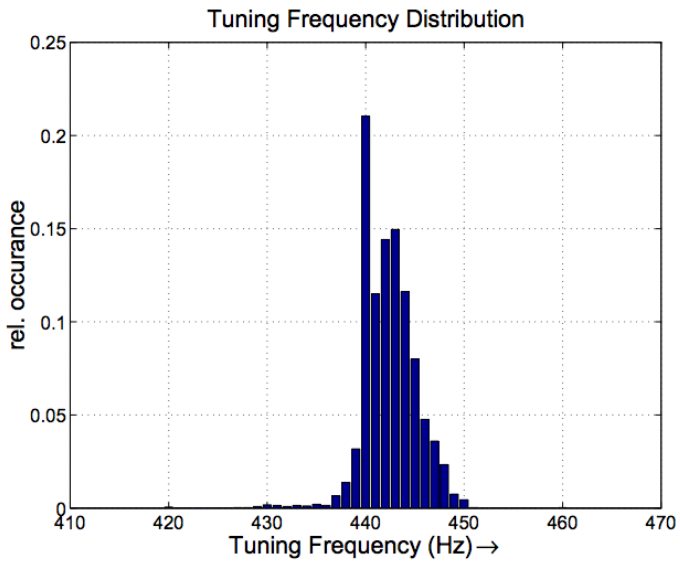
The tuning frequency f_{A4}

- is the frequency of the concert pitch or standard pitch
- is standardized internationally to 440 Hz
- that said the exact frequency used by musicians can vary due to the use historic instruments or timbre preferences

Tuning frequency

The tuning frequency f_{A4}

- is the frequency of the concert pitch or standard pitch
- is standardized internationally to 440 Hz
- that said the exact frequency used by musicians can vary due to the use historic instruments or timbre preferences



Alexander Lerch On the Requirement of Automatic Tuning Frequency Estimation ISMIR 2006

Identification of tonal components

For a wide range of applications (coding, analysis, synthesis), identifying the tonal components is of crucial interest.
The input signal can be assumed to be:

- a time variant mixture of tonal and non-tonal components
- an undefined number of tonal voices

The tonal components can be assumed to be

- salient: with a certain intensity and not masked by nearby components
- deterministic: its phase cannot change erratically
- stationary for a minimum length of time

Methods

- Local maximum: binary decision
- Peakiness: magnitude difference between the local maxima and neighbors
- Thresholding: adaptive relative to a smoothed magnitude spectrum
- Frequency coherence: analysis of the phase

Fundamental frequency detection

Time domain methods:

- Zero Crossing Rate (ZCR): the interval between 2 zero crossings related directly to the fundamental period length
- Autocorrelation function (ACF): the goal of the autocorrelation is to correlate the signal and a shifted version of itself. The mathematical definitions is:

$$y(n) = \sum_{k=1}^{M} x(n)x(n+k)$$

If the signal is periodic, the autocorrelation function y(n) also will be, and if the signal is harmonic the autocorrelation function will have peaks in multiples of the fundamental frequency. This technique is most efficient at mid to low frequencies.

Fundamental frequency detection

Frequency domain methods:

- Harmonic Product Spectrum:

$$X(k,n) = \prod_{j=1}^{O} |X(jk,n)|^2$$

- Harmonic Sum Spectrum:

$$X(k,n) = \sum_{j=1}^{O} |X(jk,n)|^2$$

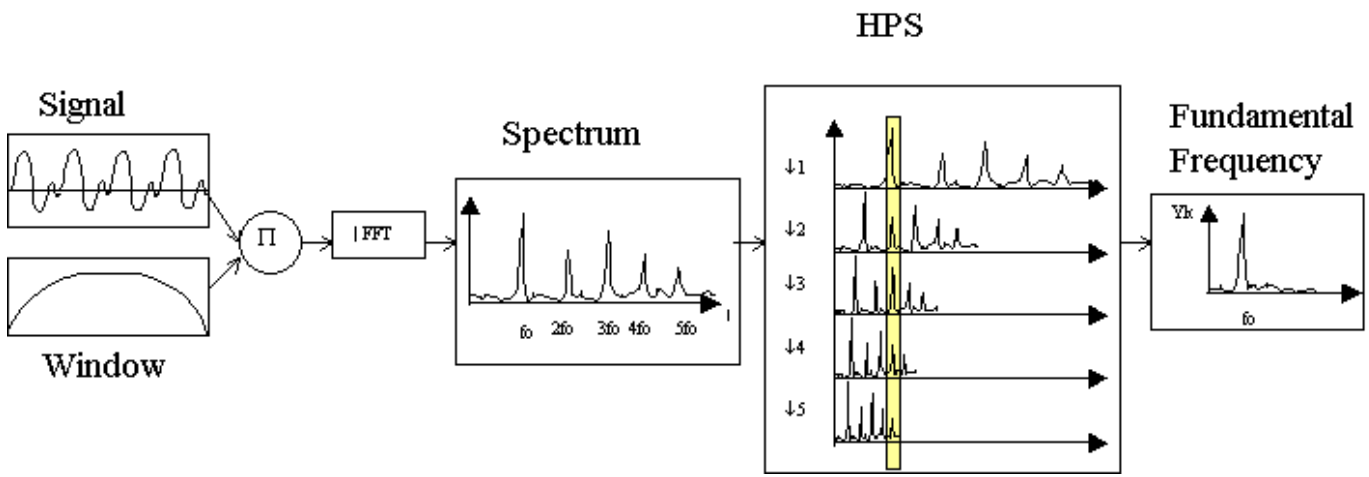
- ACF of the magnitude spectrum
- Cepstral domain

Fundamental frequency detection

Frequency domain methods:

- Harmonic Product Spectrum:

$$X(k,n) = \prod_{j=1}^{O} |X(jk,n)|^2$$



- Harmonic Sum Spectrum:

$$X(k,n) = \sum_{j=1}^{O} |X(jk,n)|^2$$

Pitch Chroma

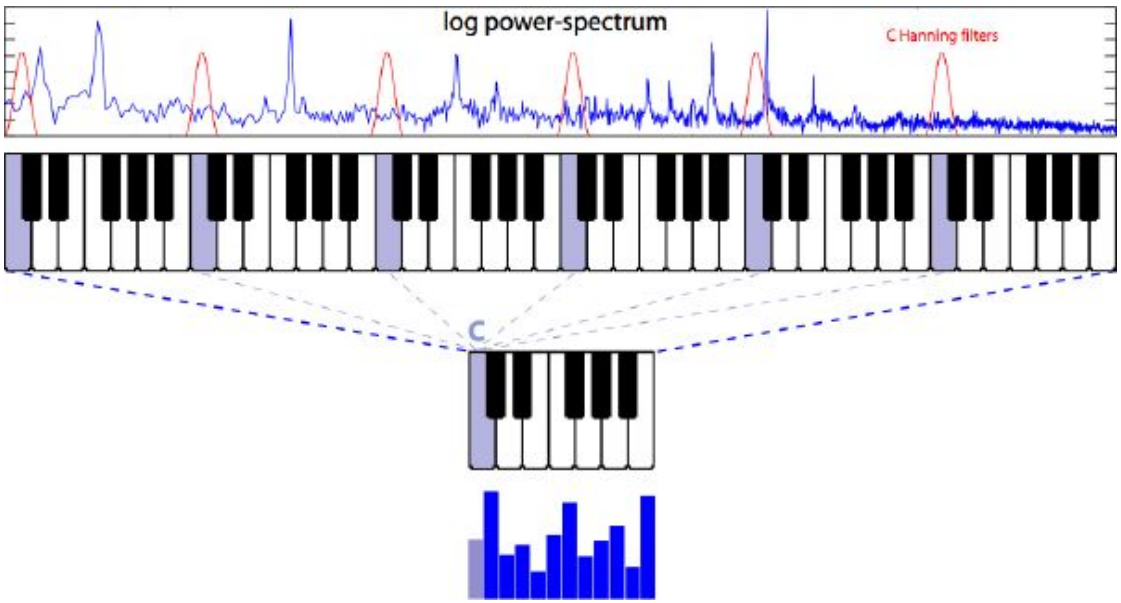
The pitch chroma

is the histogram-like 12-dimensional vector with each dimension representing one pitch class.

The main building blocks are:

- frequency representation grouped into semi tone bands
- a measure of salience in each band
- computation of the sum of bands over all octaves

Pitch Chroma



This features is largely invariant to:

- octave change
- timbre fluctuation
- noise

Human perception of temporal events

- Onset: the start of a sound event defined as its time and strength
- Attack time: 5 to 200 ms depending on the musical instrument
- Tempo: rate at which different pulses with equal duration units occur at a moderate and natural rate. Perceived tempo is the tactus or the foot tapping rate (typical value is 100 Beats per Minute (BPM))

$$T = \frac{60}{t(j+1) - t(j)}$$

- Meter: alternation of strong and weak musical elements which are grouped with a length of normally 3 to 7 beats or of duration of around 5 seconds.
- Rhythm: perceptual groups of musical elements of length between one beat and the length of the meter

Human perception of temporal events

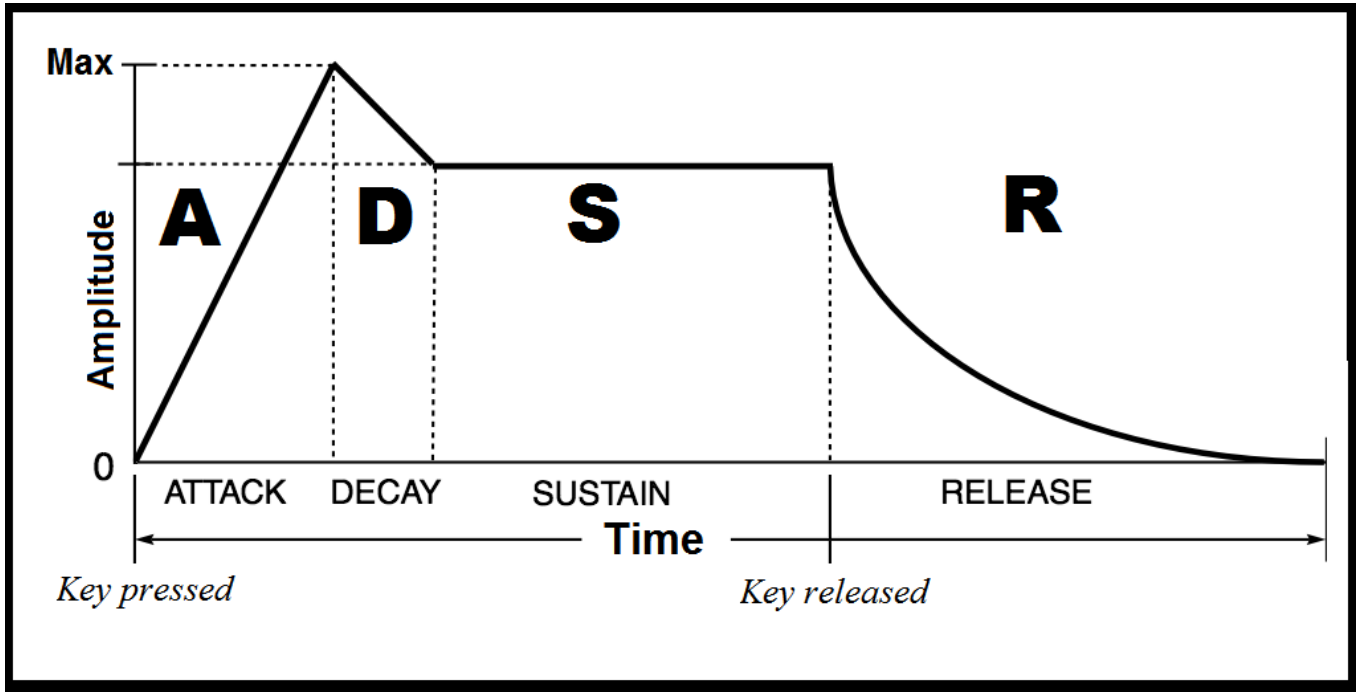
- Onset: the start of a sound event defined as its time and strength
- Attack time: 5 to 200 ms depending on the musical instrument
- Tempo: rate at which different pulses with equal duration units occur at a moderate and natural rate. Perceived tempo is the tactus or the foot tapping rate (typical value is 100 Beats per Minute (BPM))

$$T = \frac{60}{t(j+1) - t(j)}$$

- Meter: alternation of strong and weak musical elements which are grouped with a length of normally 3 to 7 beats or of duration of around 5 seconds.
- Rhythm: perceptual groups of musical elements of length between one beat and the length of the meter

Human perception of temporal events

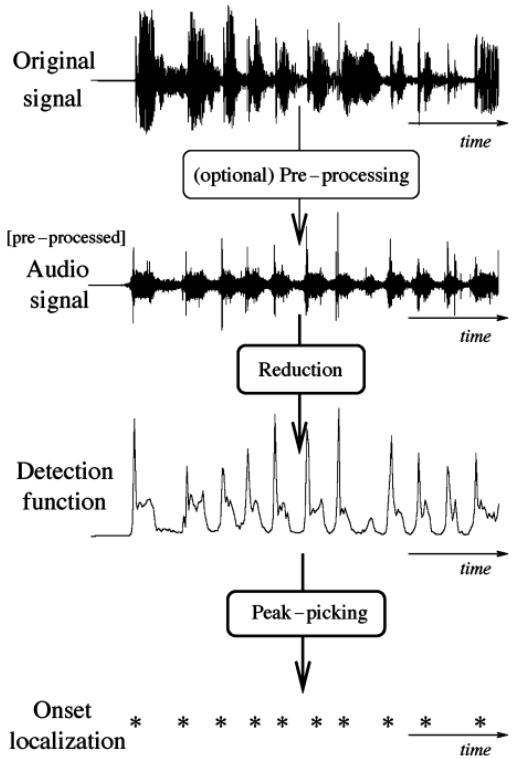
- Onset: the start of a sound event defined as its time and strength
- Attack time: 5 to 200 ms depending on the musical instrument



- Tempo: rate at which different pulses with equal duration units occur at a moderate and natural rate. Perceived tempo is the tactus or the foot tapping rate (typical value

Onset detection

Onset detection systems are usually built on 3 successive processing steps



Bello & al A Tutorial on Onset Detection in Music Signals IEEE transactions on Speech and Audio Processing, vol. 13, no. 5, September 2005

Novelty function

Something new is happening:

- spectral flux:

$$n_f = \sum_{k=k_{min}}^{k_{max}} \sqrt{|X(n,l)|} - \sqrt{|X(n-1,k)|}$$

- logarithmic distance:

$$n_f = \sum_{k=k_{min}}^{k_{max}} \log_2 \frac{|X(n,l)|}{|X(n-1,k)|}$$

- complex difference:

$$n_f = \sum_{k=k_{min}}^{k_{max}} |X(n,l) - X(n-1,k)|$$

Peak picking

- Fixed threshold:

$$T = \lambda$$

- Moving Average (MA) smoothing:

$$T(n) = \lambda + \sum b(j) n_f(n-j)$$

- Median filtering

Beat histogram

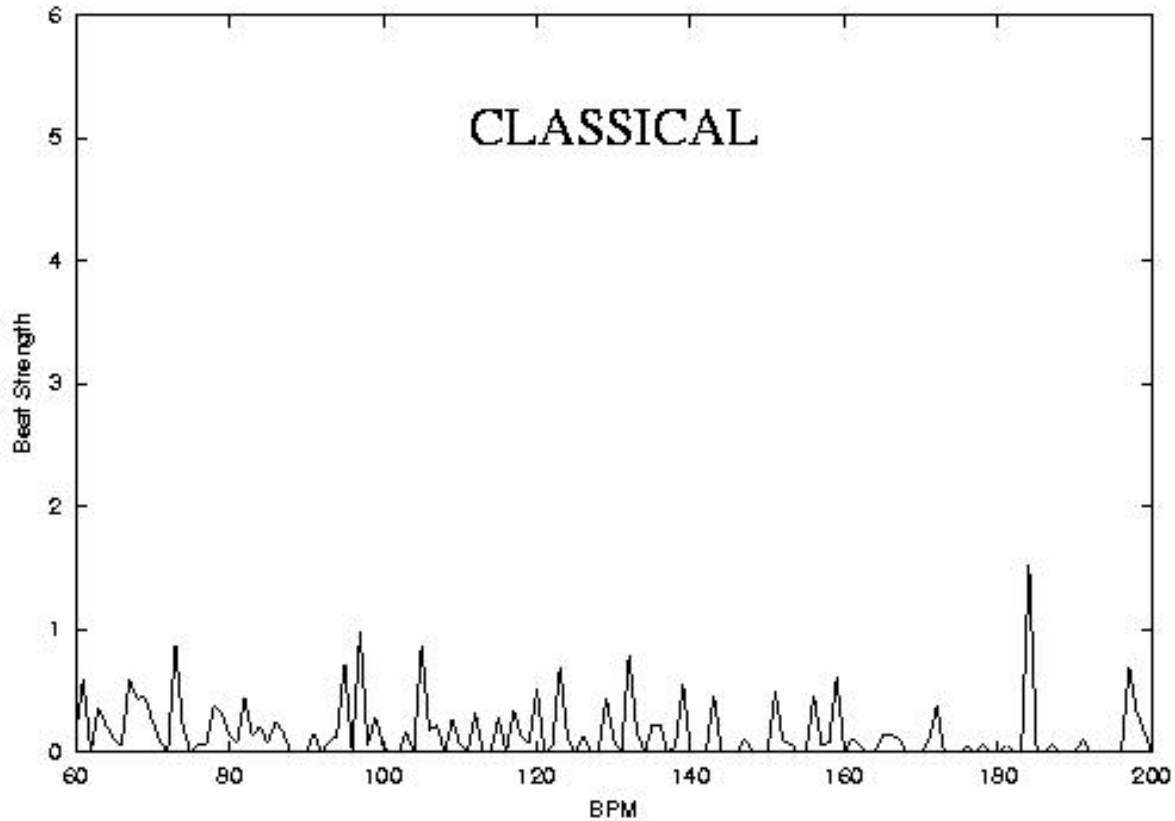
The Beat histogram

can be interpreted as the frequency domain representation of the novelty function. It is a way to visualize some rhythmic properties of the signal.

Typical processing blocks are:

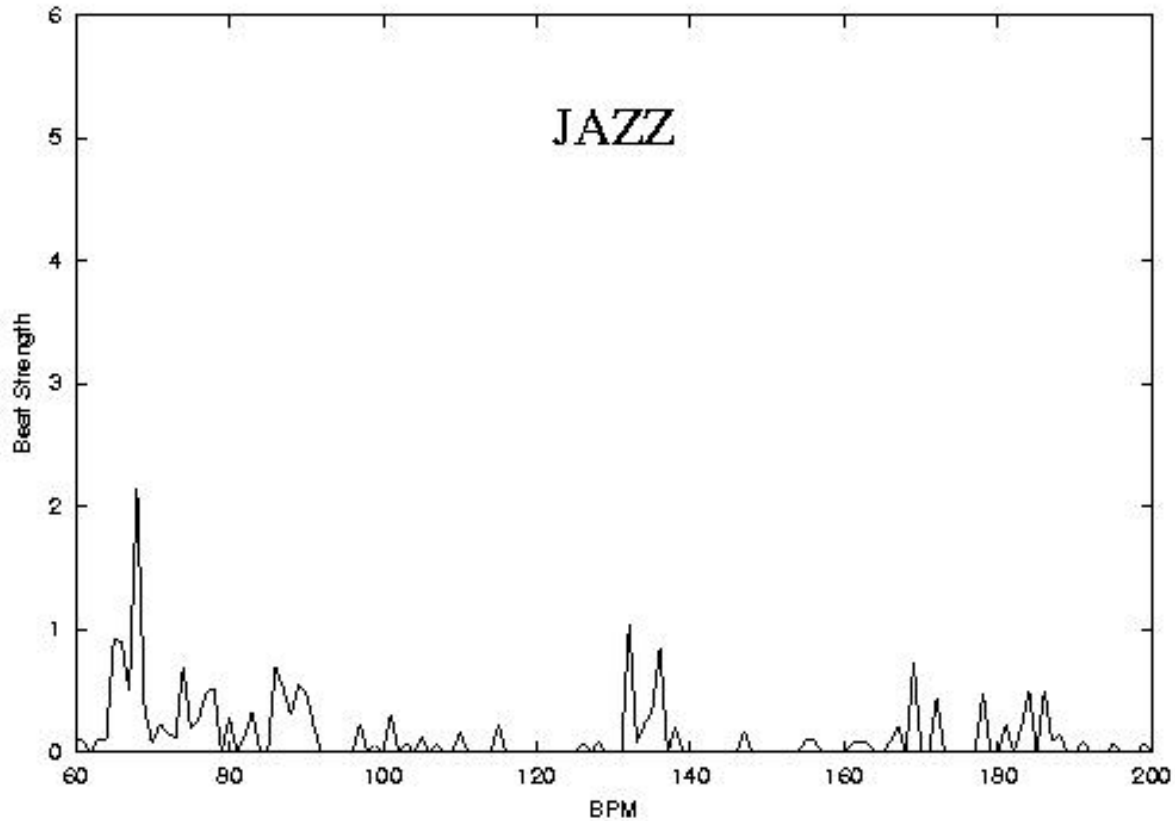
1. split the signal into several frequency bands (4)
2. Full Wave Rectification (FWR): compute absolute value
3. envelope smoothing: low pass filtering
4. down sampling: reducing the sampling rate
5. DC removal: subtracting the arithmetic mean
6. frequency analysis: ACF
7. amplitude thresholding: select the 3 most prominent components and add them to the histogram

Beat histogram



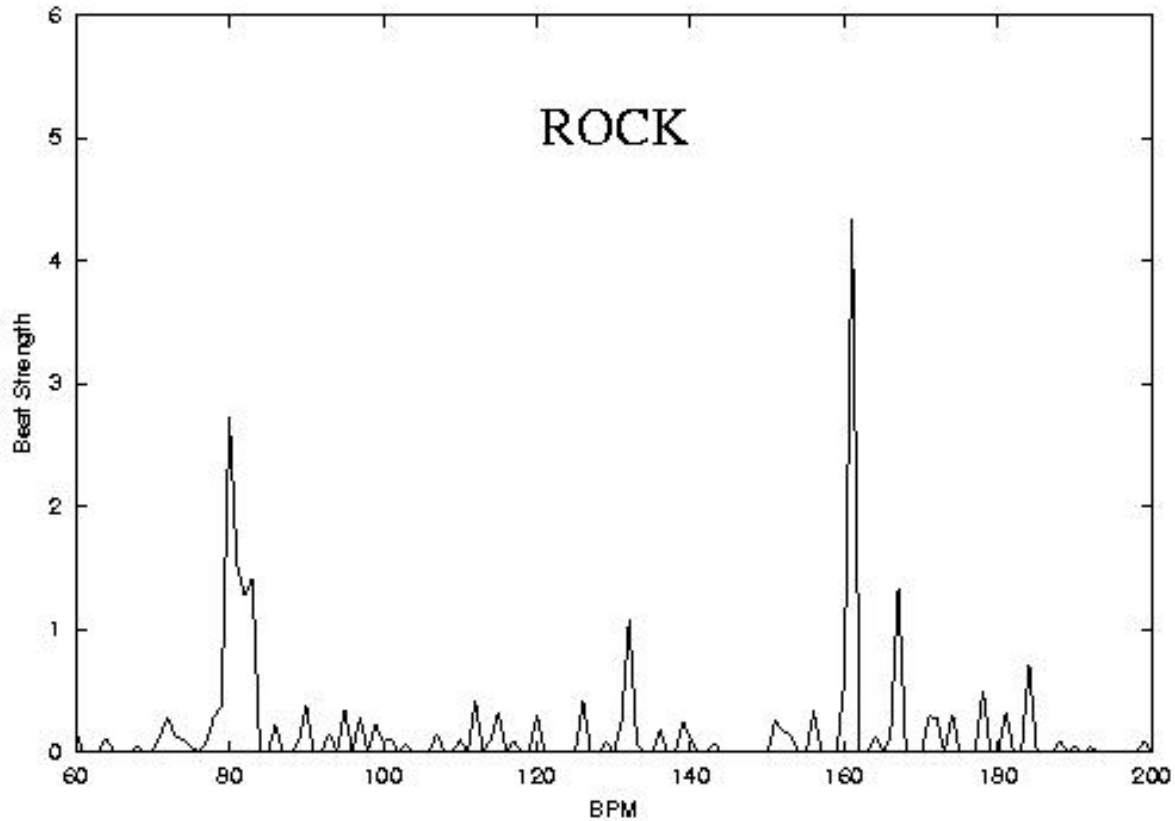
Tzanetakis & al Musical Genre Classification of Audio Signals IEEE Transactions on Speech and Audio Processing, 10(5), July 2002

Beat histogram



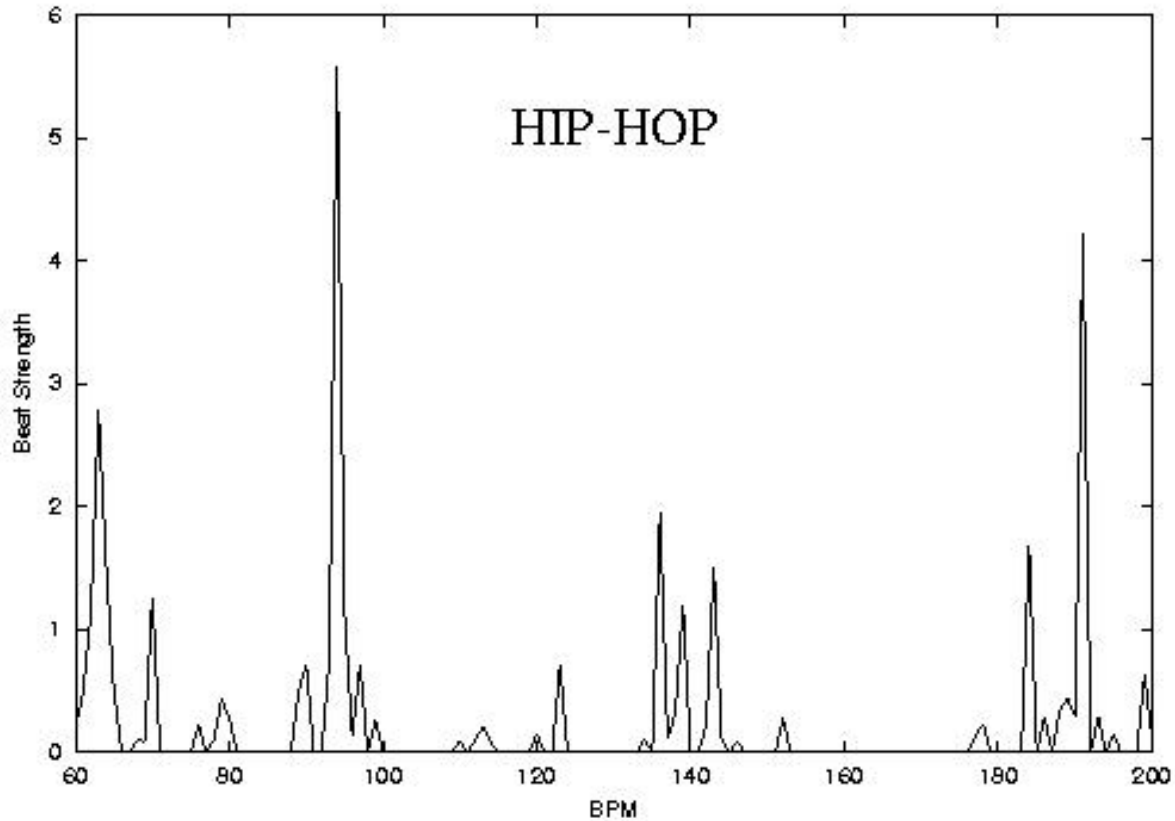
Tzanetakis & al Musical Genre Classification of Audio Signals IEEE Transactions on Speech and Audio Processing, 10(5), July 2002

Beat histogram



Tzanetakis & al Musical Genre Classification of Audio Signals IEEE Transactions on Speech and Audio Processing, 10(5), July 2002

Beat histogram



Tzanetakis & al Musical Genre Classification of Audio Signals IEEE Transactions on Speech and Audio Processing, 10(5), July 2002

Common machine listening tasks

Machine listening tasks

- Visualization
- Query by example
- Classification

For each of those tasks, the features are usually not considered directly.

Derived features

- Time derivative:

$$v_{j,\Delta} = v_j(n) - v_j(n-1)$$

- Low pass filtered version: to focus on long term variations
- Monomials:

$$v_{jl} = v_j(n).v_l(n)$$

Normalization

- if the features can reasonably be considered as symmetric and similar shape (Gaussian distributed), the most standard way is the standardization

$$v_{j,N} = \frac{v_j(n) - \mu_{v_j}}{\sigma_{v_j}}$$

- if not, one can use some features transform such as the Box-Cox transform:

$$v_\lambda = \frac{v^\lambda - 1}{\lambda}$$

- or consider the median

Pooling

For classification tasks, it is common to smooth out local variations by considering a texture window of several frames. The pooling operator is:

- image processing: max
- audio processing: sum
- typical window length ranges from 1 to 15 seconds
- in both, pyramidal schemes are or may be used.

Dimensionality reduction

From a description point of view, the higher number of features, the better.
Depending on the learning scheme, it may or may not be beneficial. Ideally, one would like to have a feature set that is:

- discriminative
- decorrelated (no redundancy between features)
- invariant to obvious signal modifications
- reasonable computational complexity

Features Selection

- Brute force subset selection: test all combinations
- Single variable classification: sort individual features per classification performance (do not model combined usefulness)
- Sequential forward selection: iteratively add the best feature addition to the pool
- Sequential backward elimination: iteratively remove the worst feature of the pool

Feature-Space Transformation

The aim is to reduce the number of features while maintaining most of the variability of the data. The most standard way to do this is to perform a Principal Component Analysis (PCA) of the feature set.

- Let us seek for a transformation matrix T such that:

$$U = T^T.V$$

- T is a linear projection:

$$T = [c_1, c_2, \cdots, c_N]$$

such that

- the vectors c_i are the direction of highest variance
- the vectors c_i are orthogonal to each others:

$$c_i^T c_j = 0, \forall i \neq j$$

Computation of the PCA

How:

1. Compute deviations from the mean:

$$B = V - h\mu^T$$

2. Compute the correlation matrix:

$$C = \frac{1}{N-1} B^*.B$$

3. Perform an eigenvalue decomposition:

$$T^{-1} C T = D$$

   where D is the diagonal matrix of eigenvalues of C

4. sort eigenvectors starting from highest eigenvalue

Computation of the PCA

Usage:

- visualization: project data on the 2 or 3 eigenvectors with the highest eigenvalues
- features selection: prune out eigenvectors with related eigenvalue lower than 1
- features selection: weights of principal eigenvectors are good relevance indicators

Similarity

Task: identify the nearest items of a database of a given seed item

- Model features distribution per item
- compare models
- Validation by considering ranking metrics: k-precision, Mean Average Precision (MAP)

Classification

Task: classify unknown item into pre-defined classes

- Feature Extraction
- Texture Blocking
- Classification decision per block
- Majority vote
- Validation by computing accuracy using N-folds cross validation: train on N-1 folds, test on 1 fold

Classification schemes

Generative classification scheme: model the joint probability distribution $p(v, l)$ of v the features and l the labels

- Gaussian Mixture Models
- Hidden Markow Models
- ...

Discriminative classification scheme: model the separation

- K-Nearest Neighbors
- Support Vector Machines (SVM)s
- ...