9T1: Spectral-based audio features

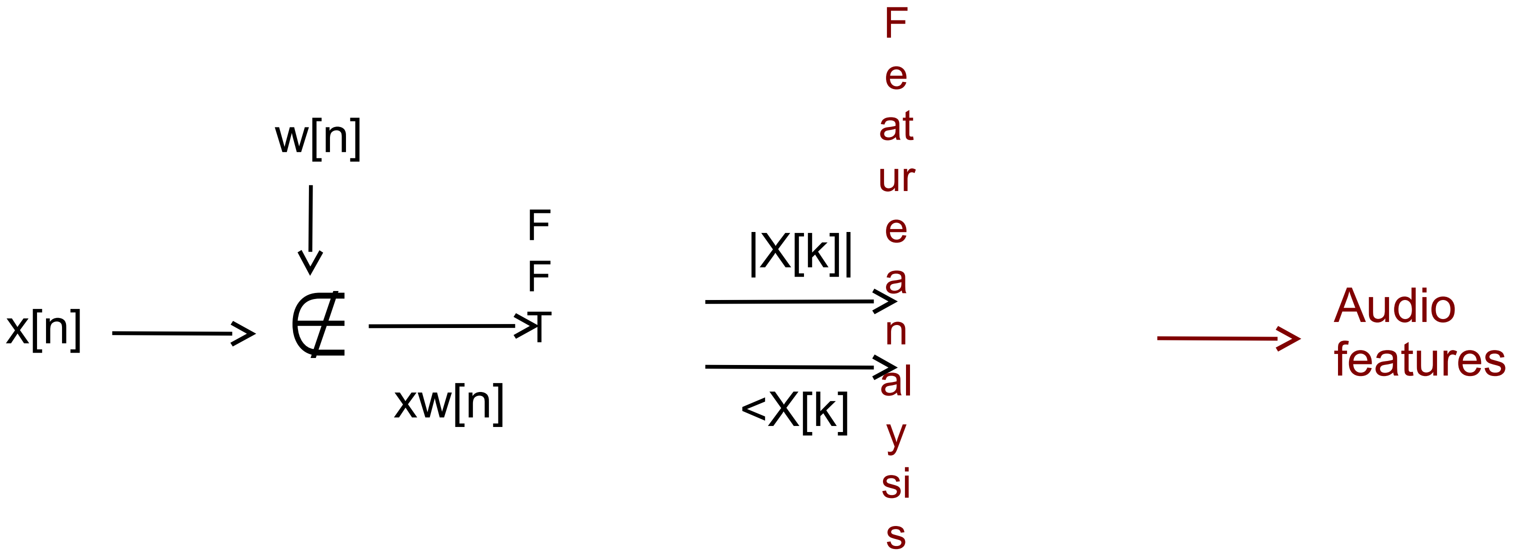
Xavier Serra

rsitat Pompeu Fabra, Barcelona

Index

- Introduction: audio features
- Single-frame spectral features
- Multiple-frames spectral features

Audio features



w[n]

x[n]

∉

xw[n]

FFT

|X[k]|

<X[k]

Feature analysis

Audio features

Essentia descriptors

- **Spectral descriptors:** BarkBands, MelBands, ERBBands, MFCC, GFCC, LPC, HFC, SpectralContrast, Inharmonicity and Dissonance, ...
- **Time-domain descriptors:** EffectiveDuration, ZCR, Loudness, ...
- **Tonal descriptors:** PitchSalienceFunction, PitchYinFFT, HPCP, TuningFrequency, Key, ChordsDetection, ...
- **Rhythm descriptors:** BeatTrackerDegara, BeatTrackerMultiFeature, BpMHistogramDescriptors, NoveltyCurve, OnsetDetection, Onsets, ...
- **SFX descriptors:** LogAttackTime, MaxToTotal, MinToTotal, TCToTotal,...
- **High-level descriptors:** Danceability, DynamicComplexity, FadeDetection, SBic, ...

Single-frame spectral features

- Energy, RMS, Loudness
- Spectral centroid
- Mel-frequency cepstral coefficients (MFCC)
- Pitch salience
- Chroma (Harmonic pitch class profile, HPCP)

Energy, RMS, Loudness

Energy:

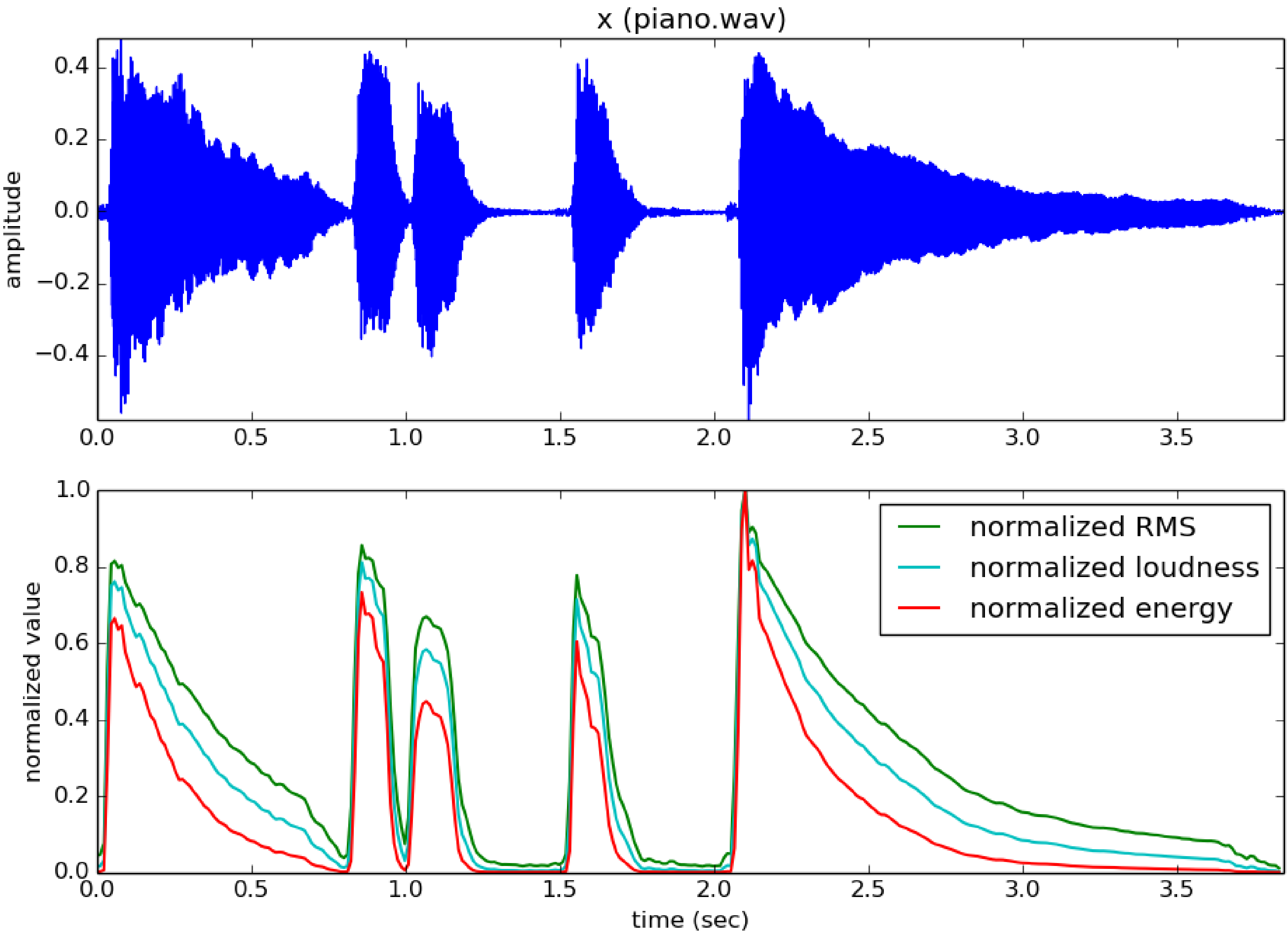
$$energy_l = \sum_{k=0}^{N-1} |X_l[k]|^2$$

Root mean square:

$$RMS_l = \sqrt{\frac{1}{N^2} \sum_{k=0}^{N-1} |X_l[k]|^2}$$

Steven's power law:

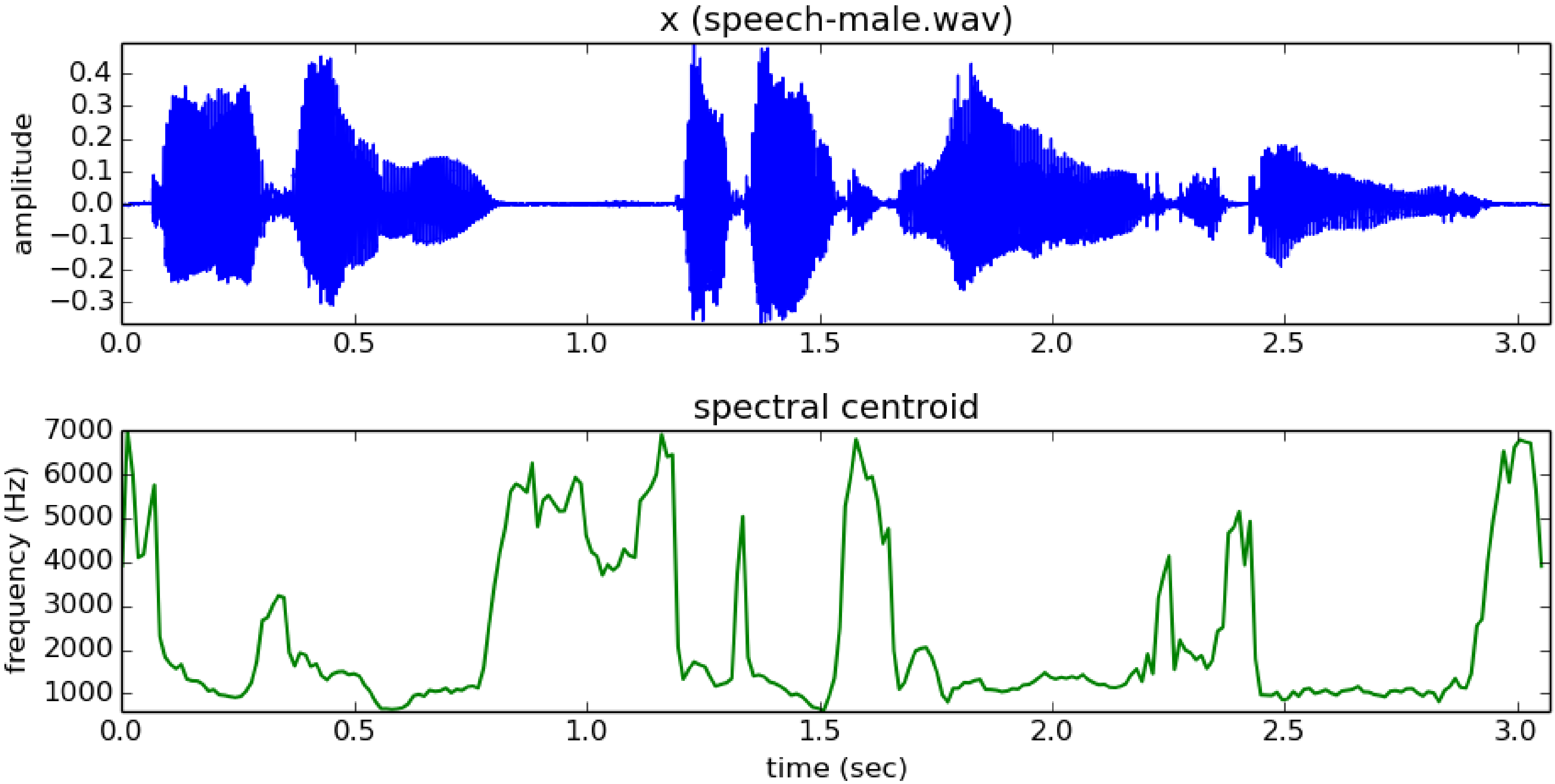
$$loudness_l = \left(\sum_{k=0}^{N-1} |X_l[k]|^2 \right)^{0.67}$$


x (piano.wav)
amplitude
0.4
0.2
0.0
−0.2
−0.4
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
normalized value
1.0
0.8
0.6
0.4
0.2
0.0
normalized RMS
normalized loudness
normalized energy
time (sec)

Spectral centroid

Is influenced by the F0 of the sound.
(sounds with higher F0s have higher spectral centroid)

$$centroid_l = \frac{\sum_{k=0}^{N/2} k|X_l[k]|}{\sum_{k=0}^{N/2} |X_l[k]|}$$



Mel frequency cepstral coefficients

Series of PERCEPTUALLY RELEVANT coefficients which are much fewer than the frequency bins present in a spectrum

$$mfcc_l = DCT\left(\log_{10}\left(\sum_{k=0}^{N/2} \mho X_l[k]\mho H_i[k]\right)\right)$$

i is the number of output cepstral coefficients (goes up to H), H is the number of filters

INDEPENDENT FROM PITCH and LOUDNESS, it is a "spectrum shape feature"

where

$\mho X[k]\mho$ is the positive magnitude spectrum

$H_i[k]$ is the mel scale filter bank for each filter i

$$DCT[m]\text{(Discrete Cosine Transform)} = \sum_{n=0}^{N-1} f[n]\cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)m\right)$$

"sampled version" of the original spectrum

linear magnitude spectrum

$$\mho X_l[k]\mho \longrightarrow H_i[k] \longrightarrow \log_{10} \longrightarrow DCT \longrightarrow \text{mfcc}$$

apply some filters (h)

The DCT (discrete cosine transform) of a DFT (the DCT of a frequency spectrum) compares shapes of sine waves -with different period- against the shape of the magnitude spectrum, and ~~calculates~~ the similarity between them

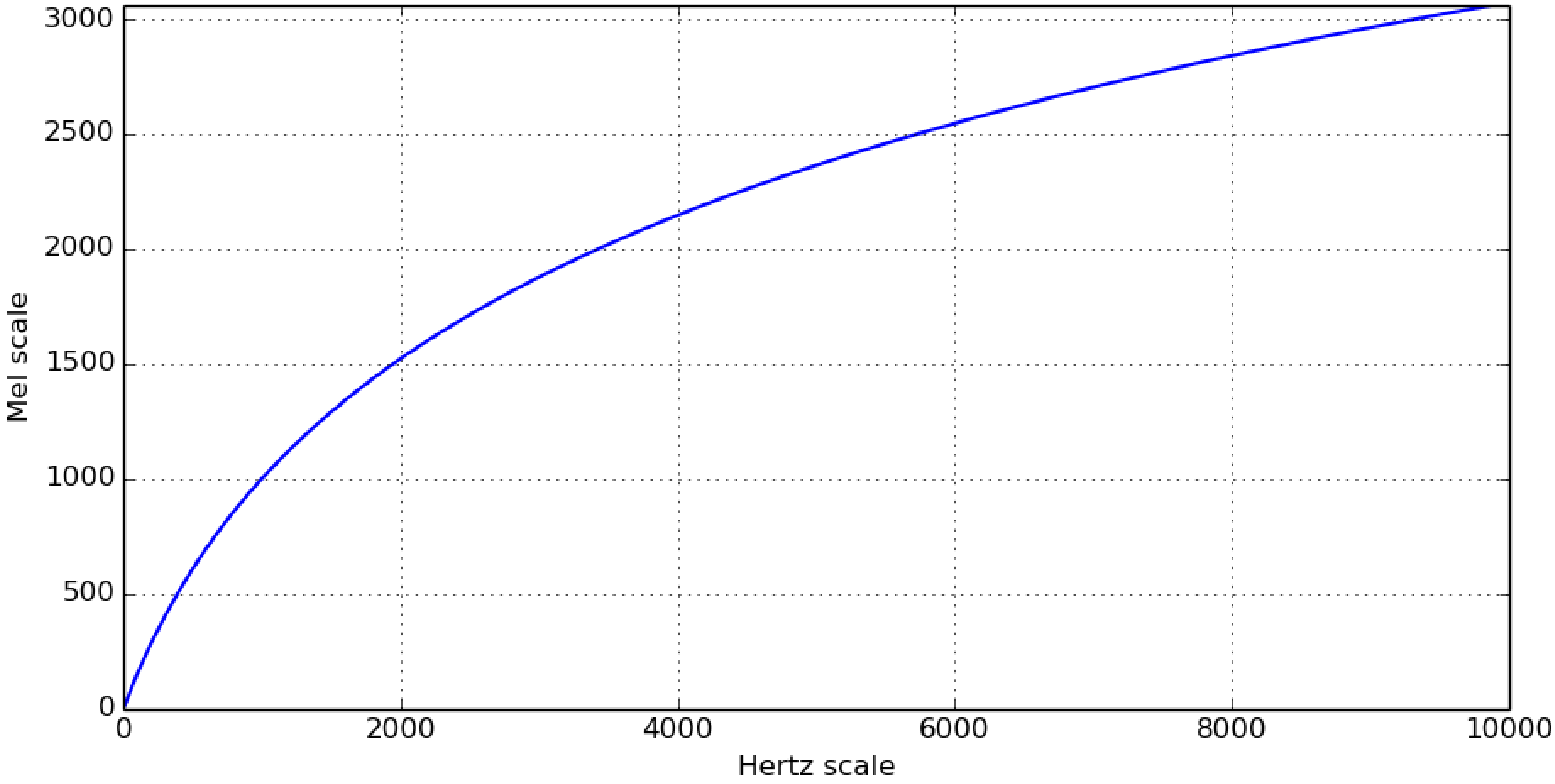
(usually) 12 coefficients in the cepstral domain

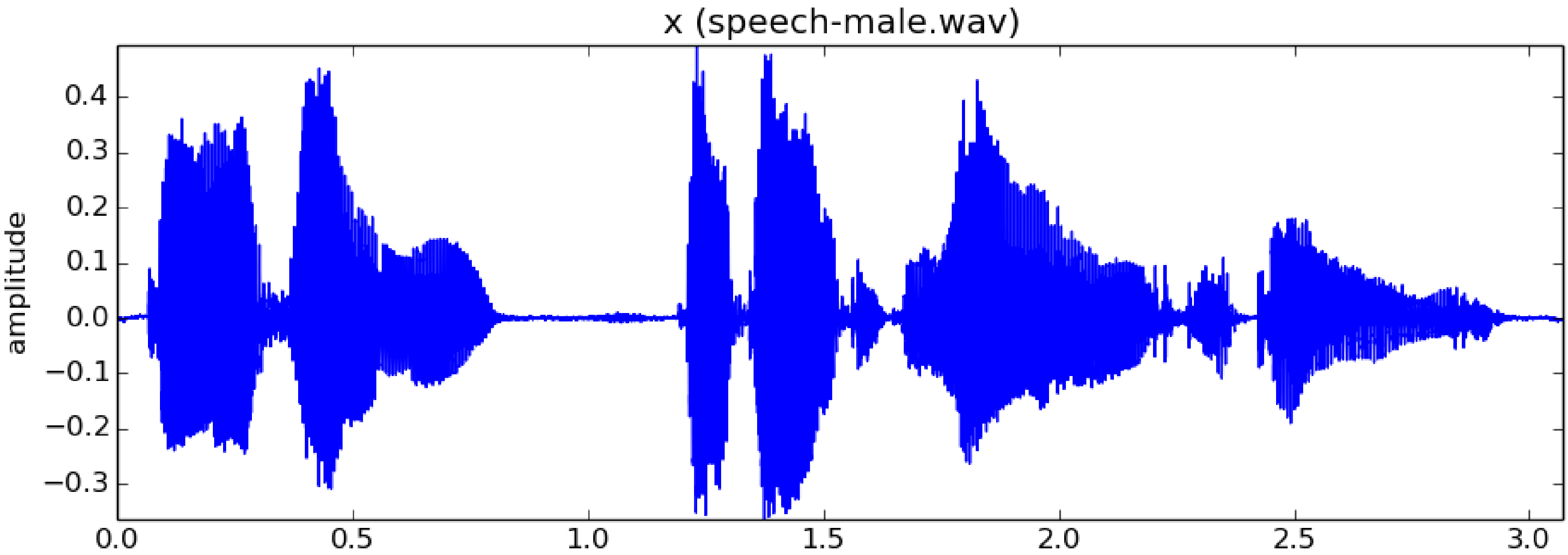
The DFT of a spectrum means comparing sine waves with different periods (0, 1, 2, ecc. times the number of frequency bins in the spectrum) against the SHAPE OF THE SPECTRUM

MFCC: Mel scale

Perceptually relevant logarithmic scale

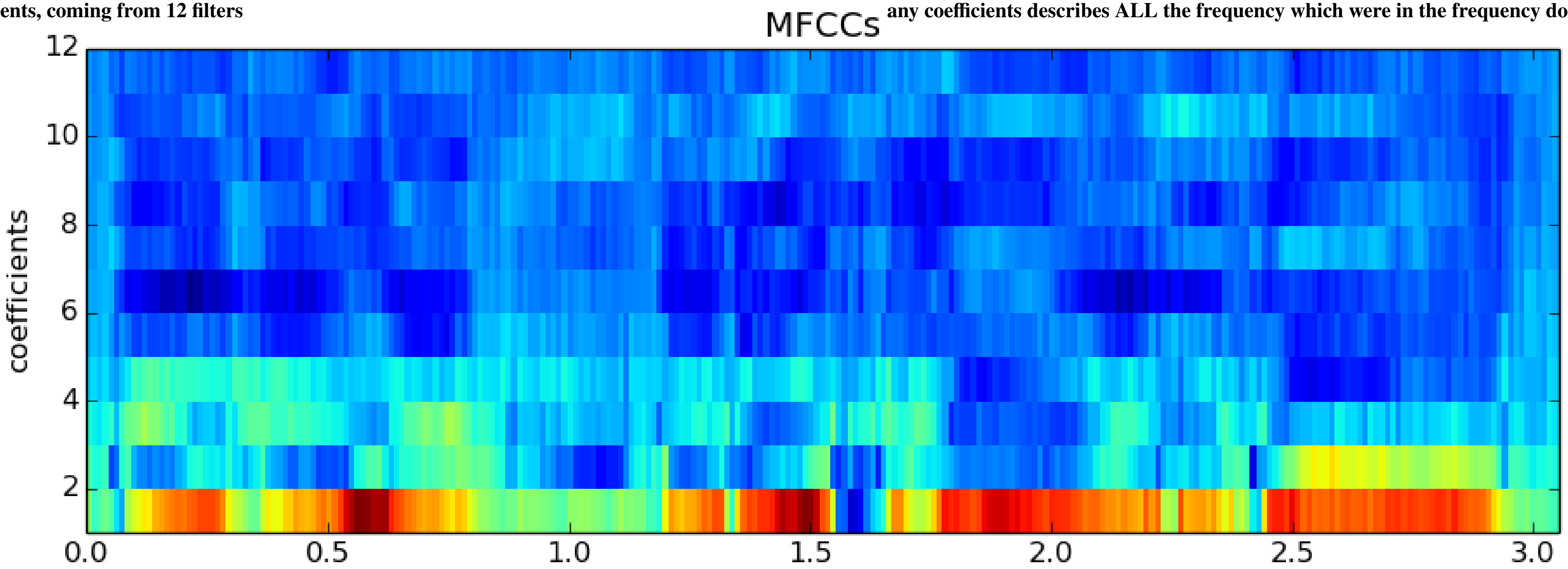
$$mel = 2595 \cdot \log_{10}(1 + \frac{f}{700})$$





12 coefficients, coming from 12 filters

any coefficients describes ALL the frequency which were in the frequency domain

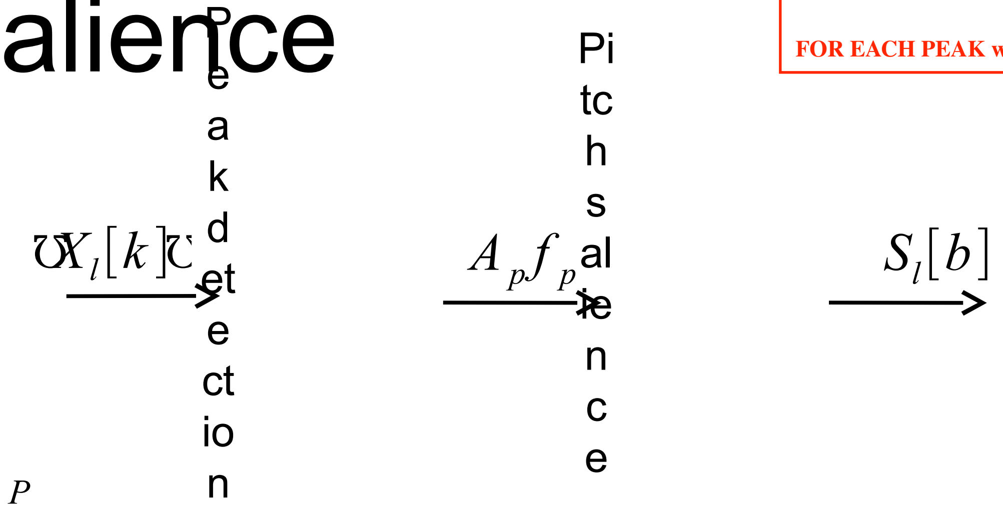


coefficient n. 0 is an offset, DC, is the AMPLITUDE basically

Pitch salience

Statistical evaluation of the spectral peaks

FOR EACH PEAK we have a probability of it being the F0



$$S[b]=\sum_{h=1}^{H}\sum_{p=1}^{P} e(A_p)\,g(b,h,f_p)\,(A_p)^{\beta}$$

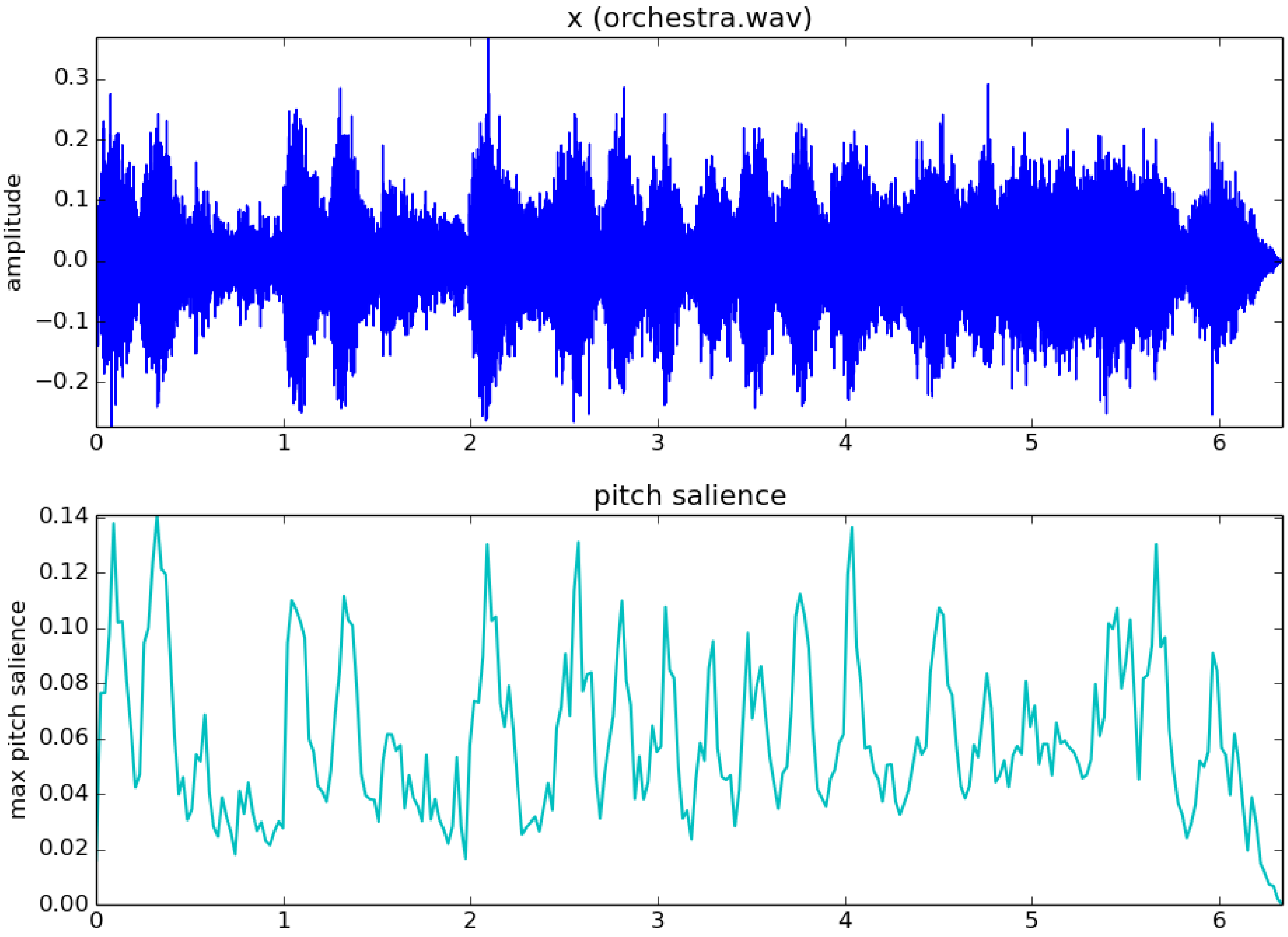
where

$S[b]=$ salience at bin frequency b (b expressed in cent scale)

$e()=$ magnitude threshold function

$g()=$ weighting function applied to peak p

$\beta=$ magnitude compression value


x (orchestra.wav)
amplitude
0.3
0.2
0.1
0.0
−0.1
−0.2
0
1
2
3
4
5
6
pitch salience
max pitch salience
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
1
2
3
4
5
6

Chroma (Harmonic Pitch Class Profile)

$$hpcp[k]=\sum_{p=1}^{P} w(k,f_p)\,A_p^{\,2}$$

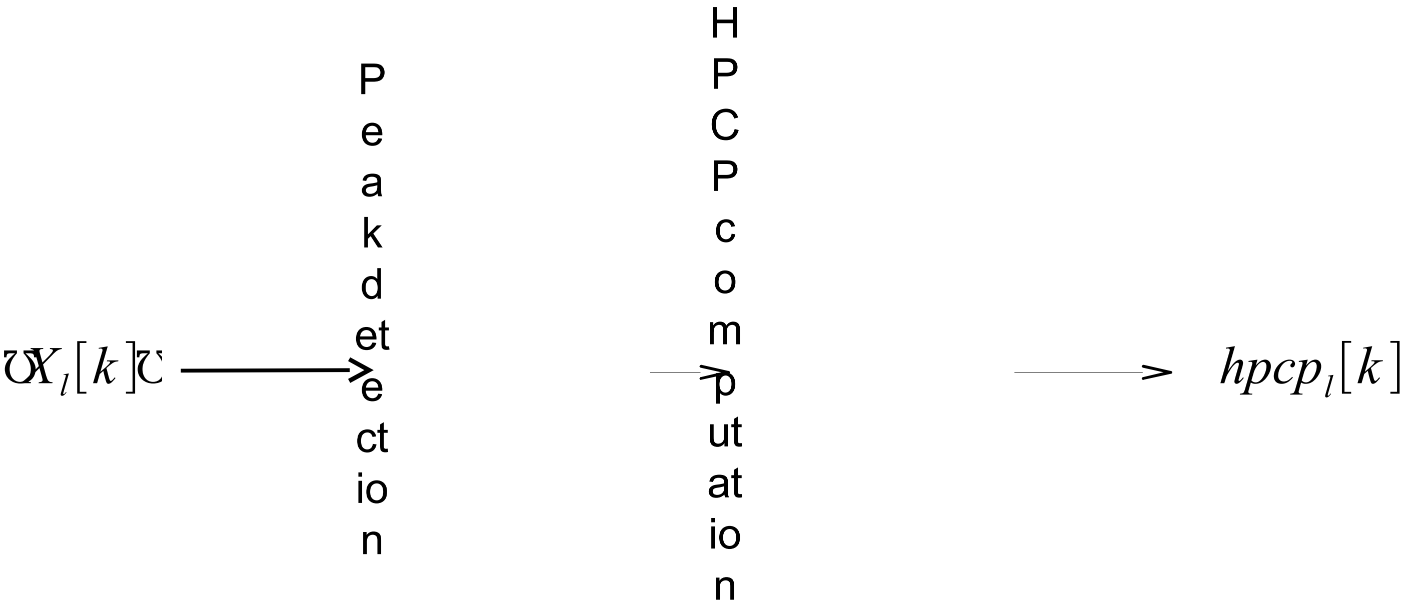
Probabilistically based F0 algorithm (a probability for each pitch class) WRAPPED around the octave

where

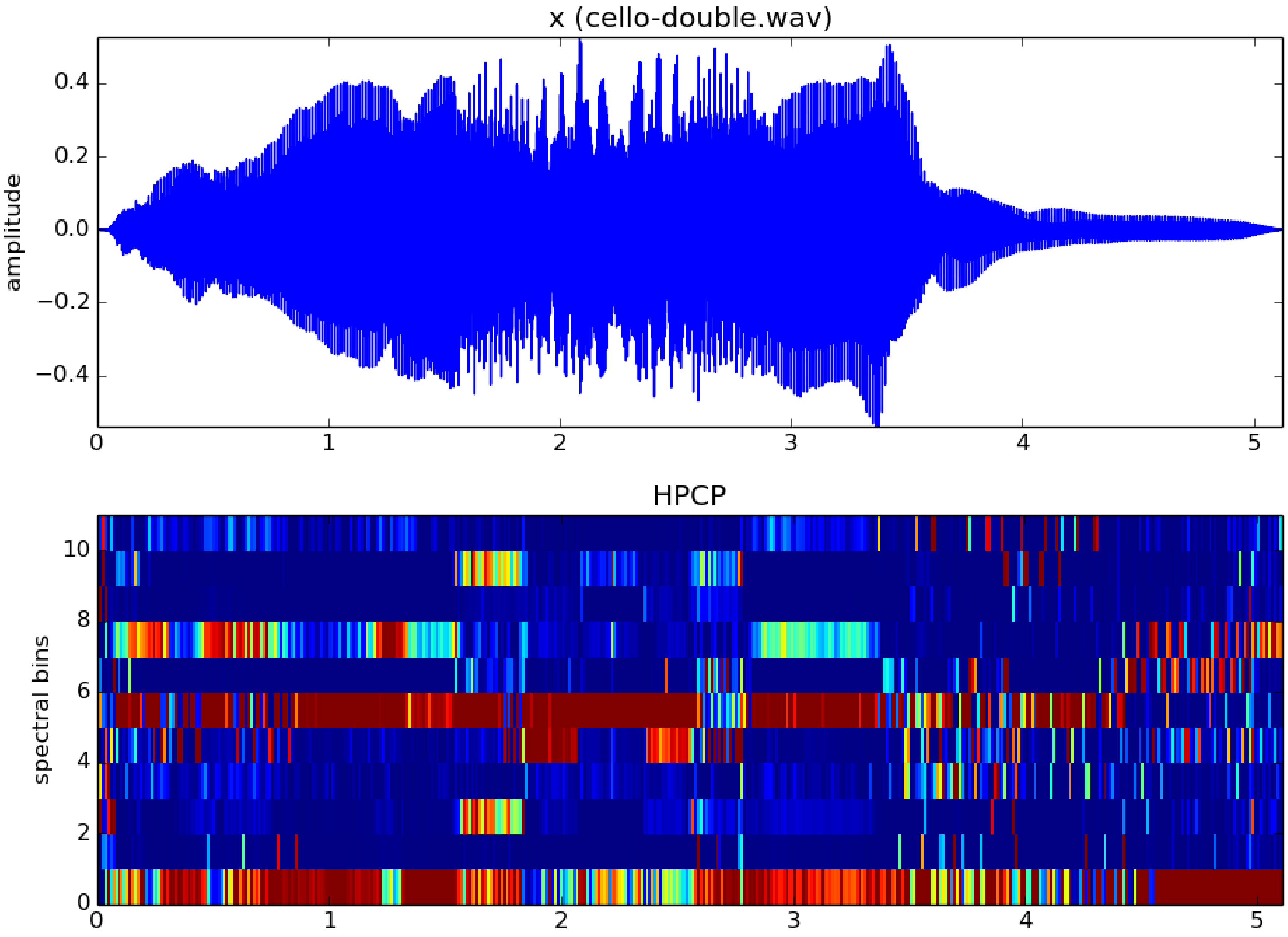
$A_p =$ amplitude of spectral peak p

$w(k,f_p) =$ weight of the peak frequency f_p for bin k

$k =$ spectral bin locations of the chosen HPCP frequencies



$|X_l[k]|$ → Peak detection → HPCP computation → $hpcp_l[k]$


x (cello-double.wav)
amplitude
0.4
0.2
0.0
−0.2
−0.4
0
1
2
3
4
5
HPCP
spectral bins
10
8
6
4
2
0
0
1
2
3
4
5

Multiple-frames spectral features

- Event segmentation, onsets
- Predominant pitch
- Statistics of single-frame features

Event segmentation, onsets

- Spectral flux (used in segmentation)

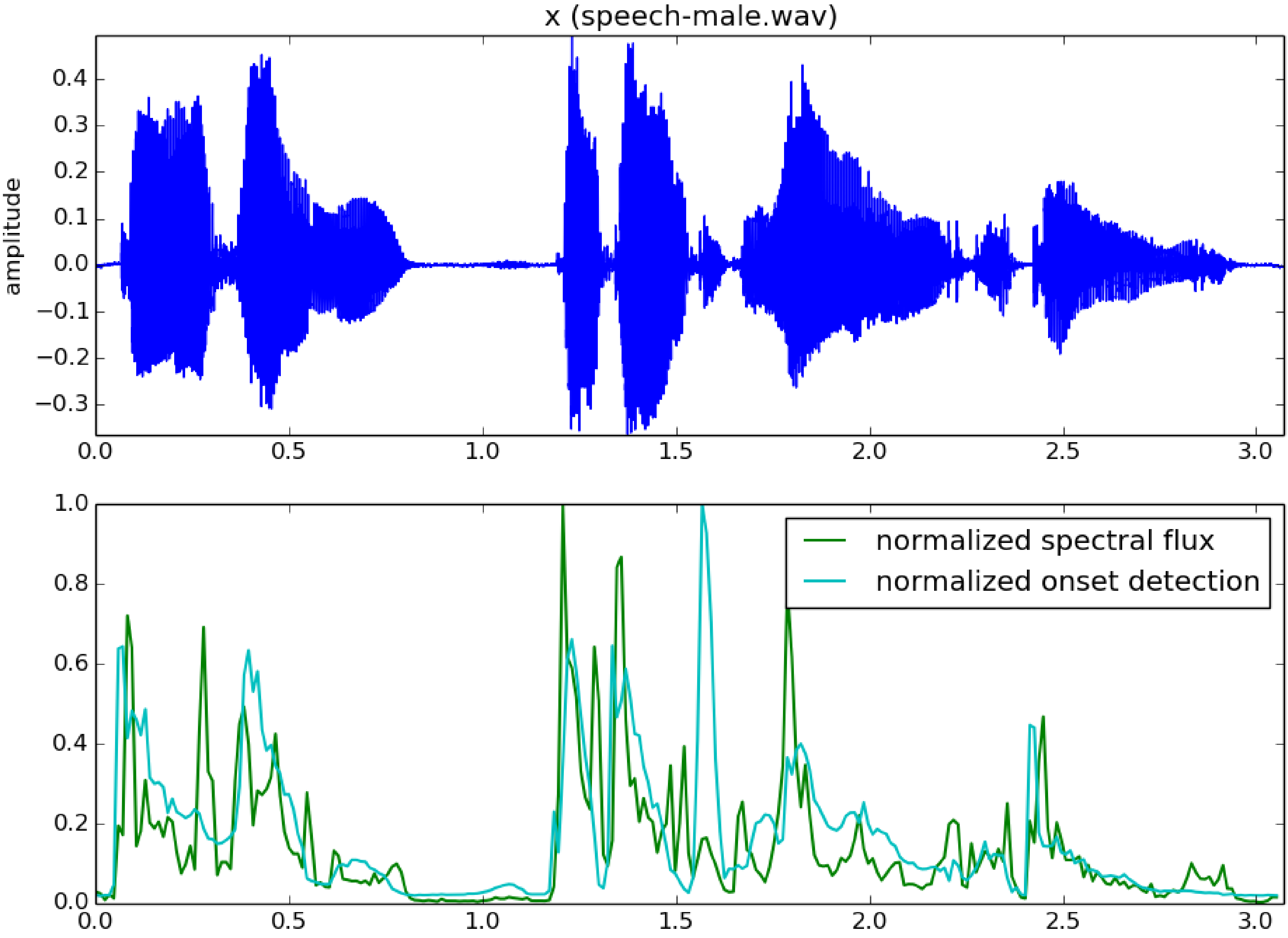
$$SF_l = \sum_{k=0}^{N/2} H\left(|X_l[k]| - |X_{(l-1)}[k]|\right)$$

$$\text{where } H(x) = \frac{x + |x|}{2}$$

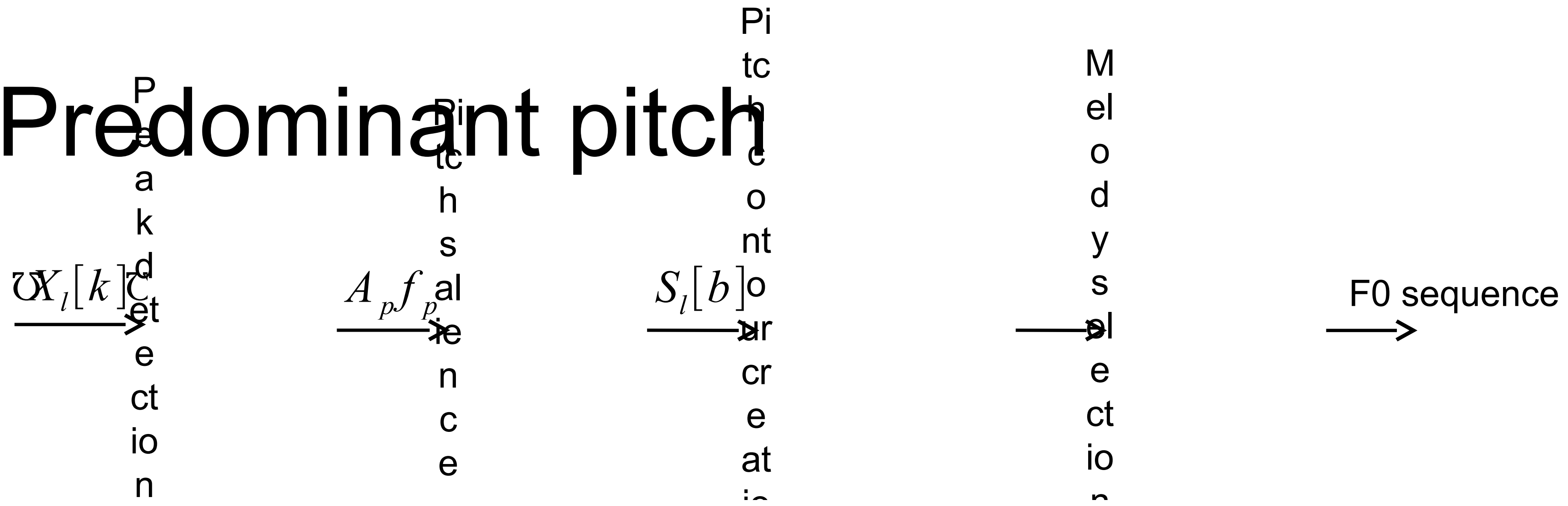
- Onset detection based on high-frequency content

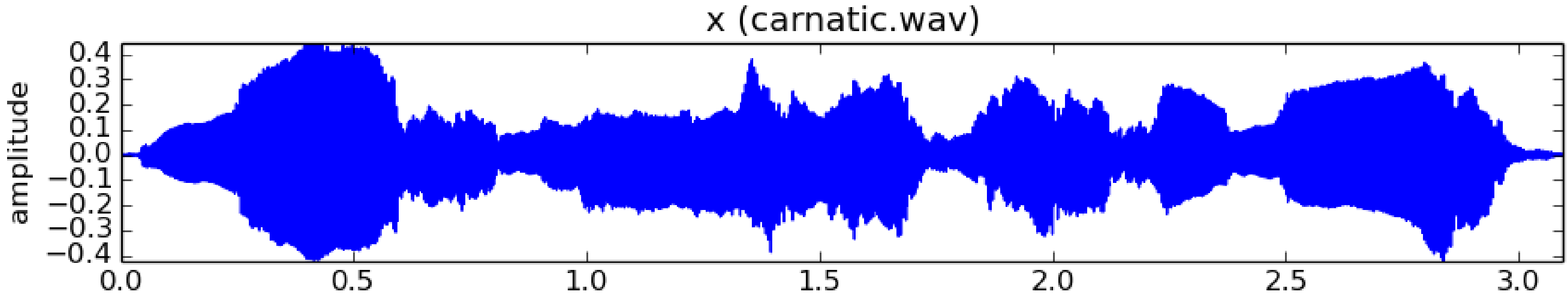
$$\text{Onset detection function} = HFC_l - HFC_{(l-1)}$$

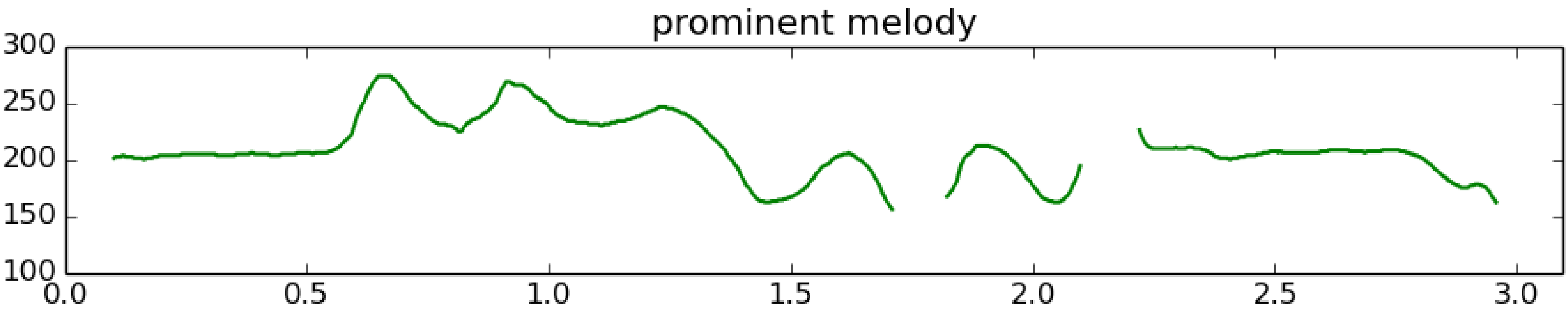
$$\text{where} \quad HFC_l = \sum_{k=1}^{N/2} |X_l[k]|^2 k$$


x (speech-male.wav)
amplitude
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1.0
0.8
0.6
0.4
0.2
0.0
normalized spectral flux
normalized onset detection

Predominant pitch







Statistics of single frame features

- Arithmetic mean (first moment)

$$mean=\frac{1}{N}\sum_{i=0}^{N-1} y[i]$$

- Variance (second moment)

$$variance=\frac{1}{N}\sum_{i=0}^{N-1} \left(y[i]-mean\right)^2$$

- Skewness (third moment)

$$skewness=\frac{\frac{1}{N}\sum_{i=0}^{N-1} \left(y[i]-mean\right)^3}{\left[\frac{1}{N-1}\sum_{i=0}^{N-1} \left(y[i]-mean\right)^2\right]^{3/2}}$$

References

- Essentia: http://essentia.upf.edu
- http://en.wikipedia.org/wiki/Spectral_centroid
- http://en.wikipedia.org/wiki/Mel-frequency_cepstrum
- http://en.wikipedia.org/wiki/Loudness
- http://en.wikipedia.org/wiki/Harmonic_pitch_class_profiles
- http://en.wikipedia.org/wiki/Onset_(audio)
- http://en.wikipedia.org/wiki/Moment_(mathematics)

9T1: Spectral-based audio features

Xavier Serra

rsitat Pompeu Fabra, Barcelona