9T1: Spectral-based audio features

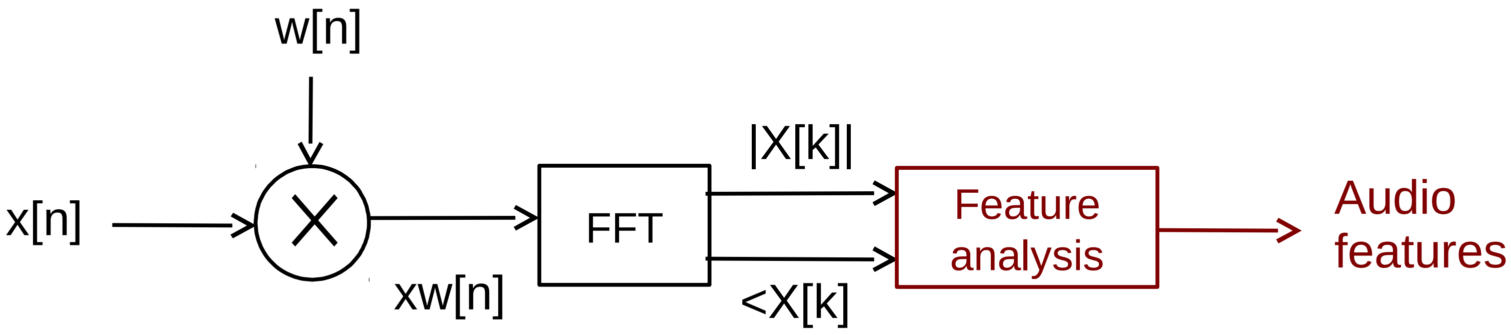
Xavier Serra

Universitat Pompeu Fabra, Barcelona

Index

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

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:

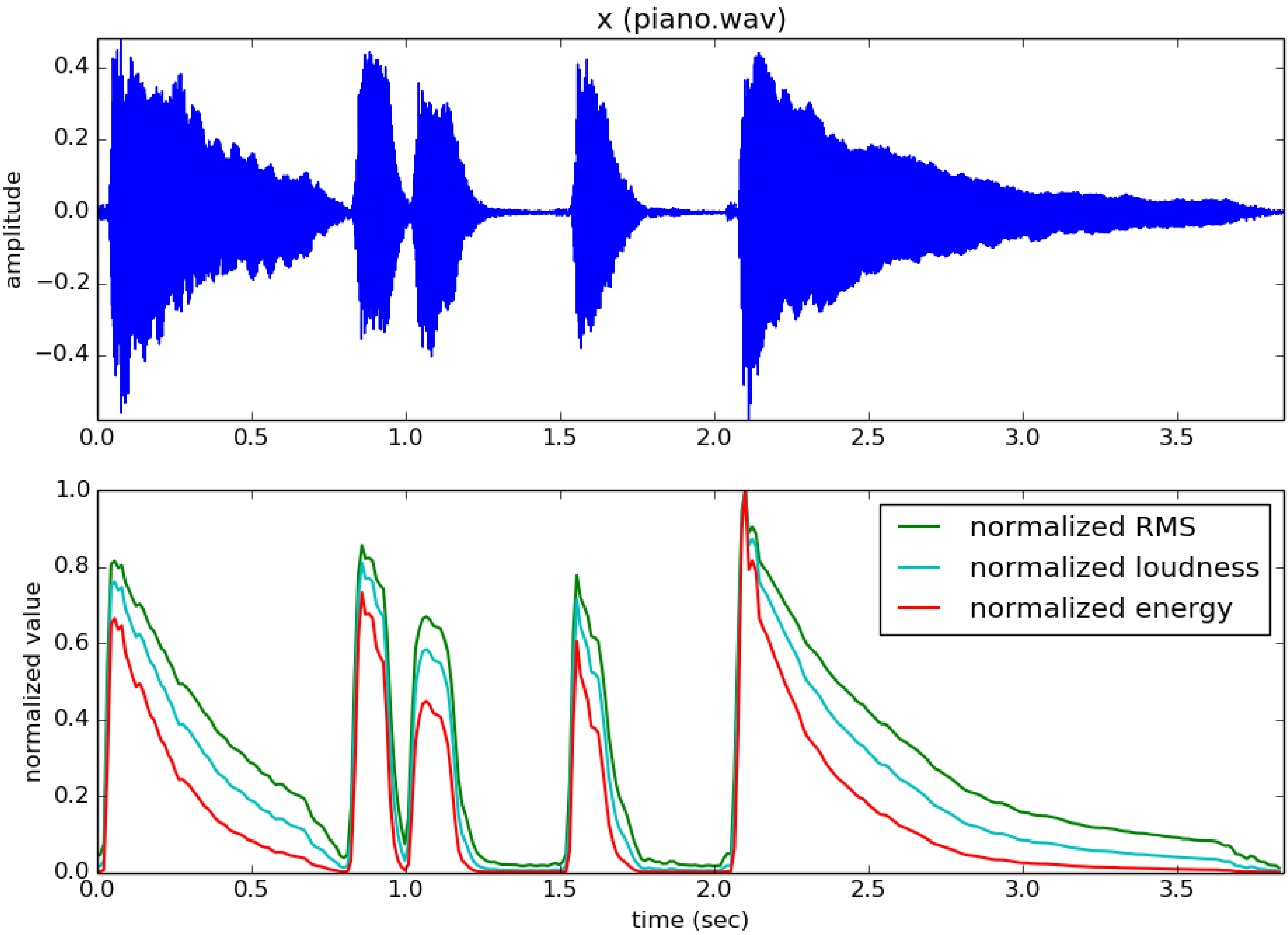
$$energy_l = \sum_{k=0}^{N-1} |X_l[k]|^2$$

Root mean square:

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

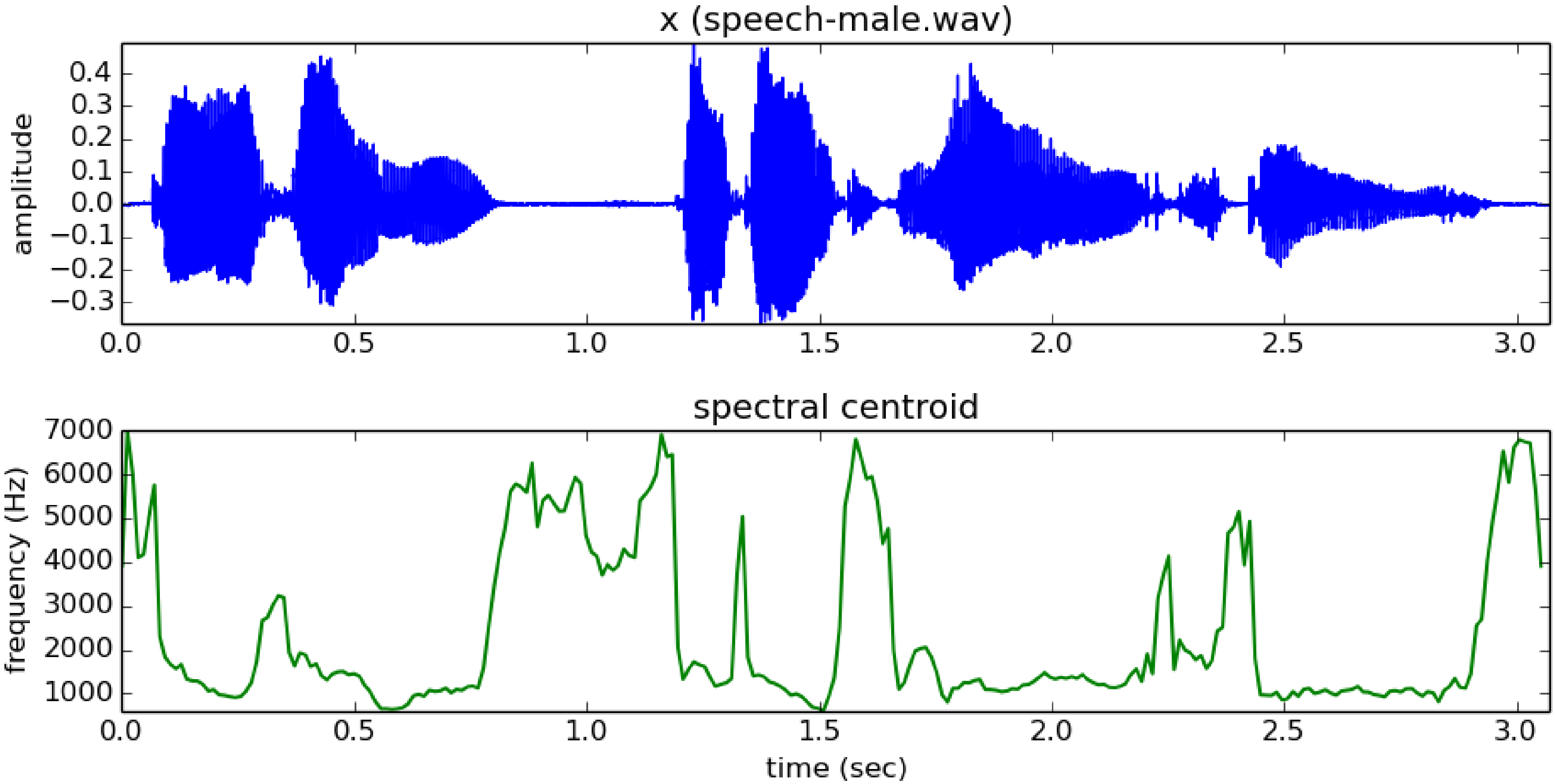
Steven's power law:

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


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

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



x (speech-male.wav)

spectral centroid

Mel frequency cepstral coefficients

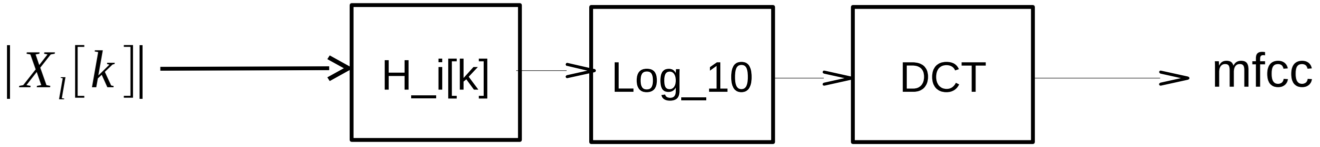
$$mfcc_l = DCT\left(\log_{10}\left(\sum_{k=0}^{N/2} |X_l[k]| H_i[k]\right)\right)$$

where

$|X[k]|$ is the positive magnitude spectrum

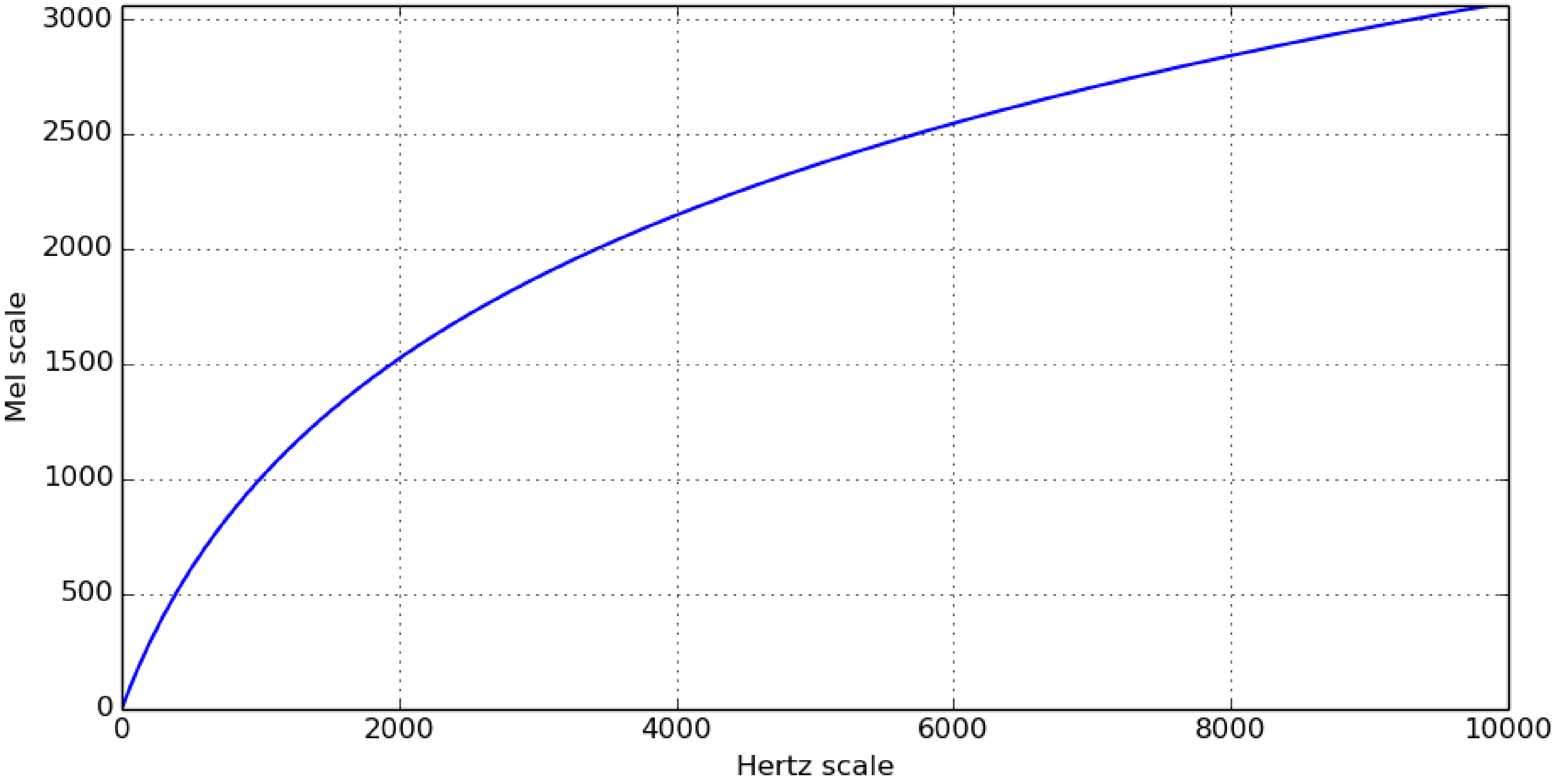
$H_i[k]$ is the mel scale filter bank for each filter i

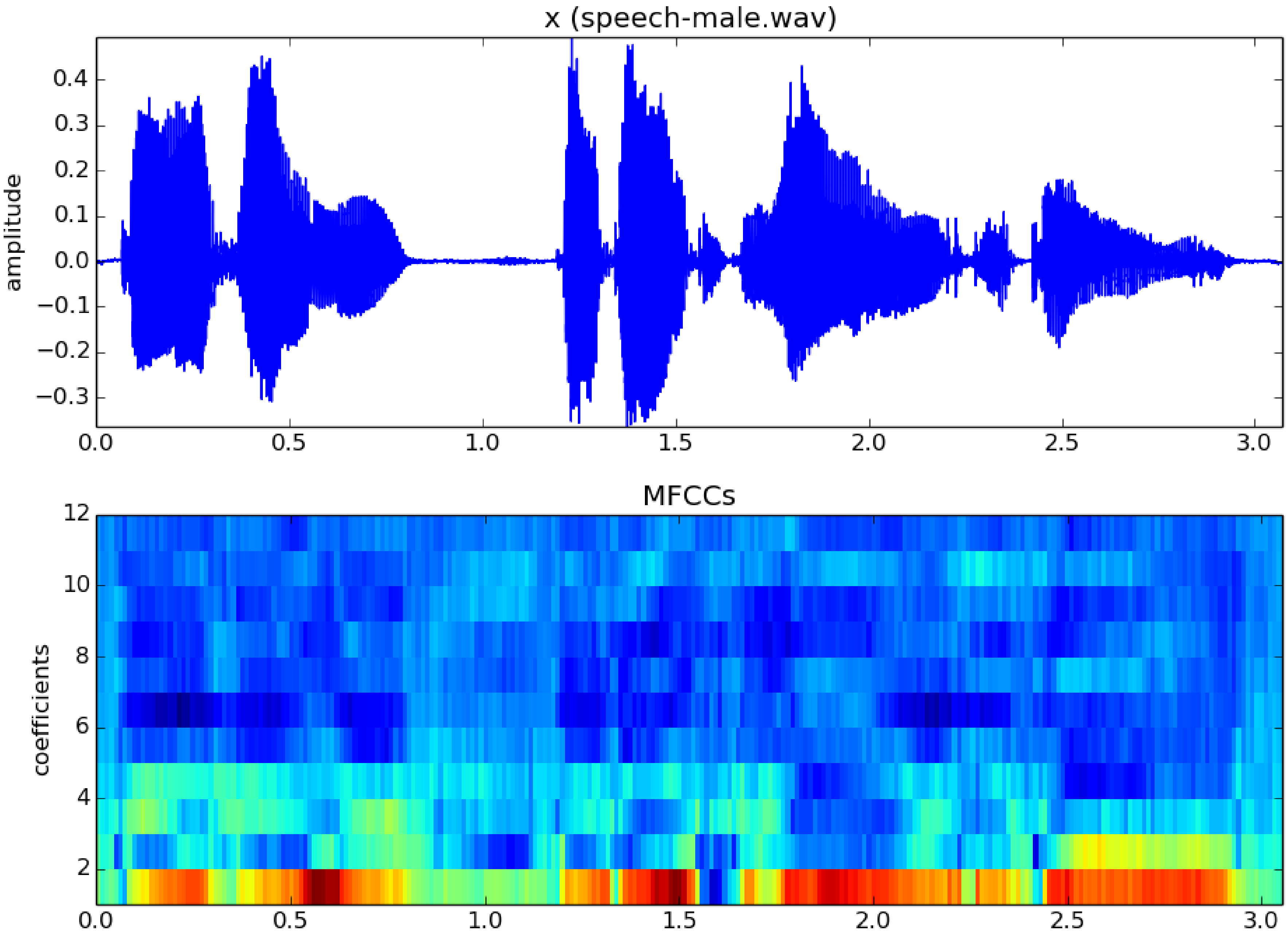
$$DCT[m](\text{Discrete Cosine Transform}) = \sum_{n=0}^{N-1} f[n]\cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)m\right)$$



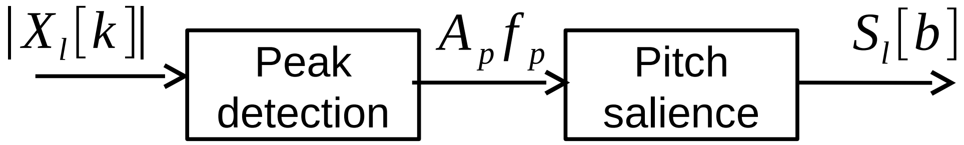
MFCC: Mel scale

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




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
MFCCs
coefficients
12
10
8
6
4
2
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Pitch salience



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

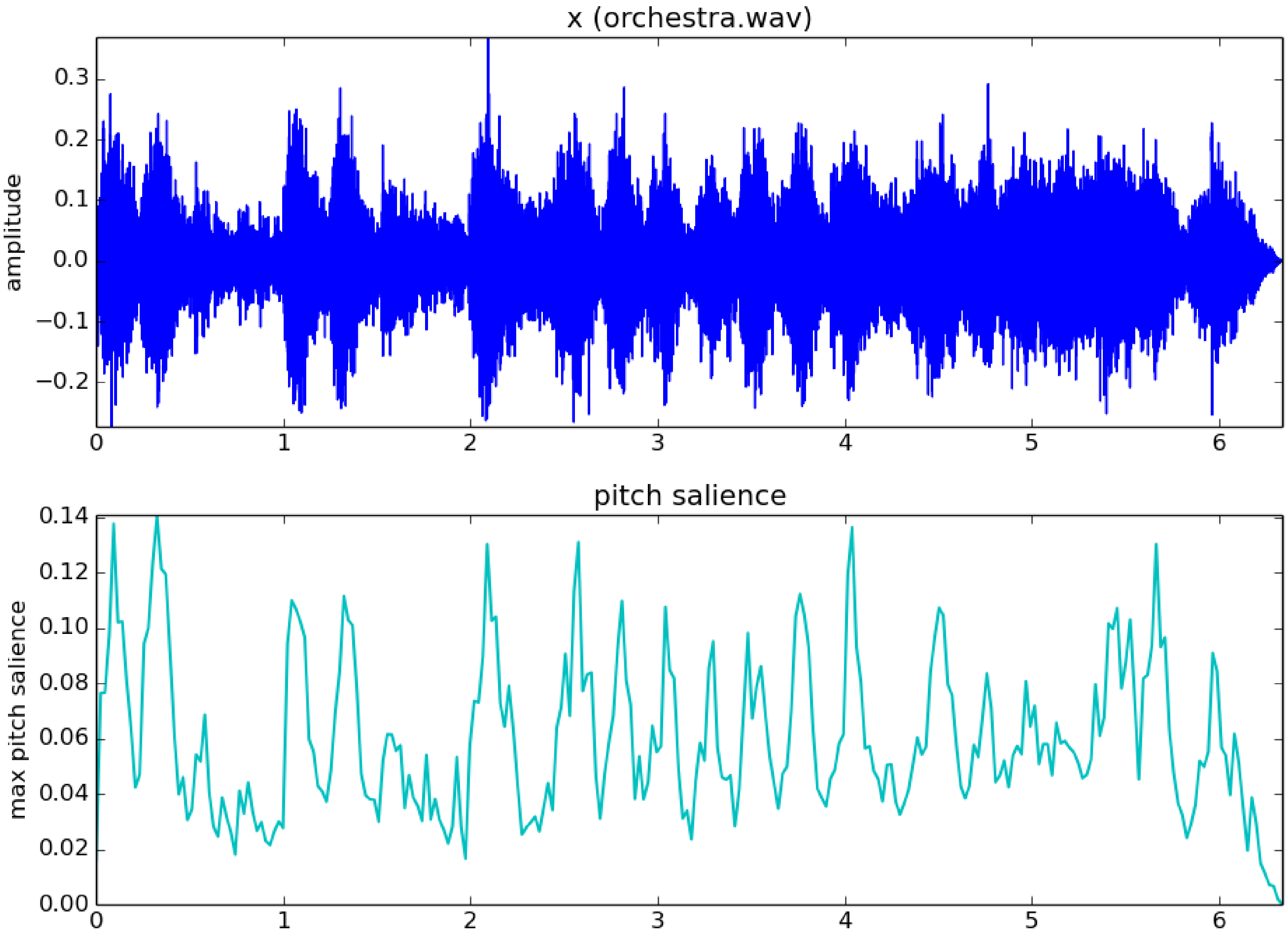
where

$S[b]=$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)

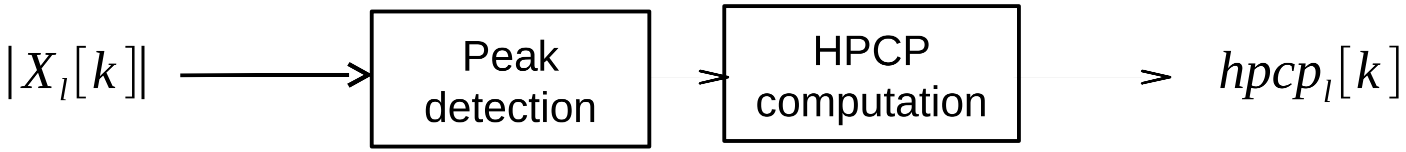
$$hpcp[k] = \sum_{p=1}^{P} w(k, f_p) A_p^2$$

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

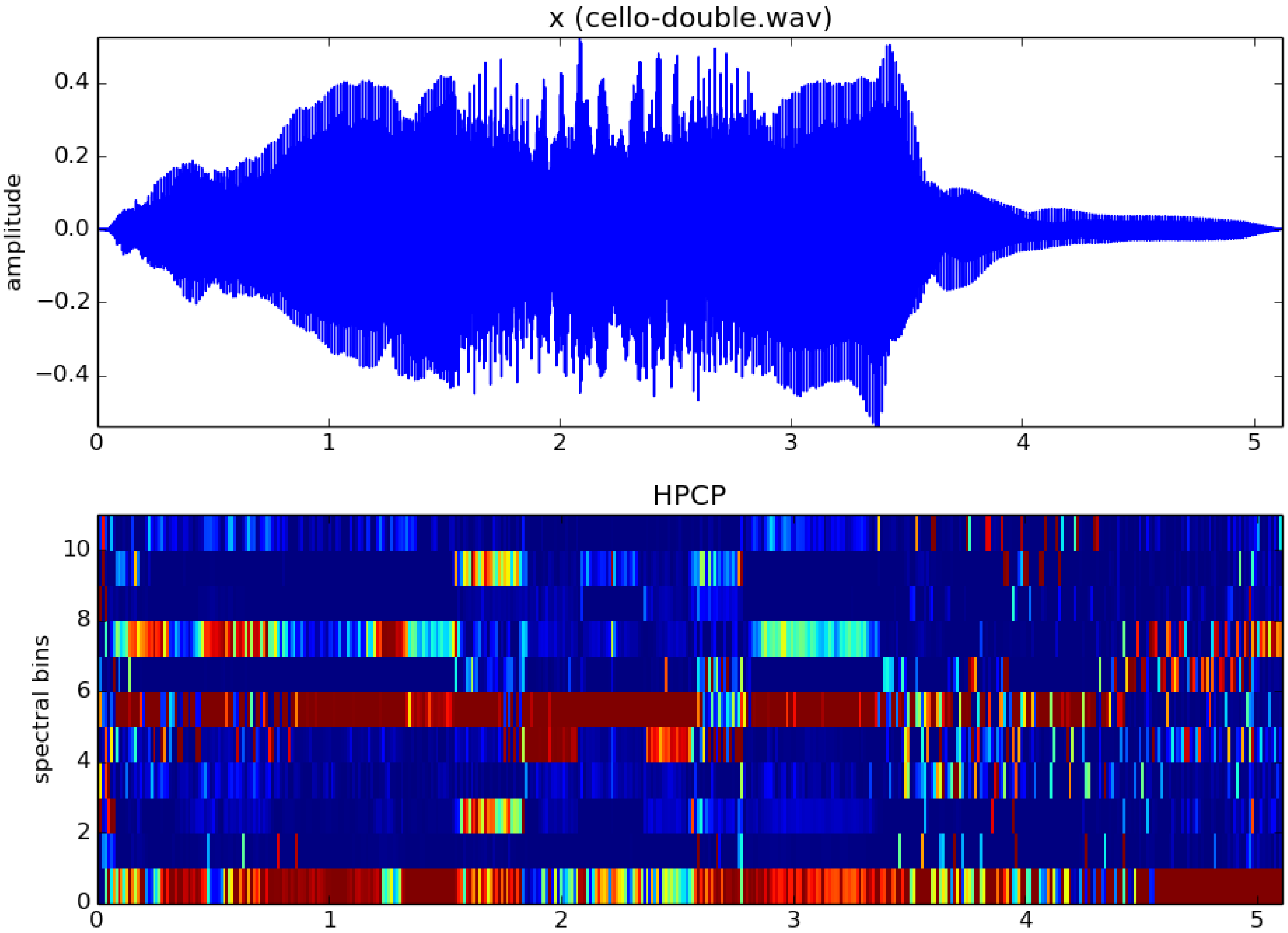


$|X_l[k]\|$

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)

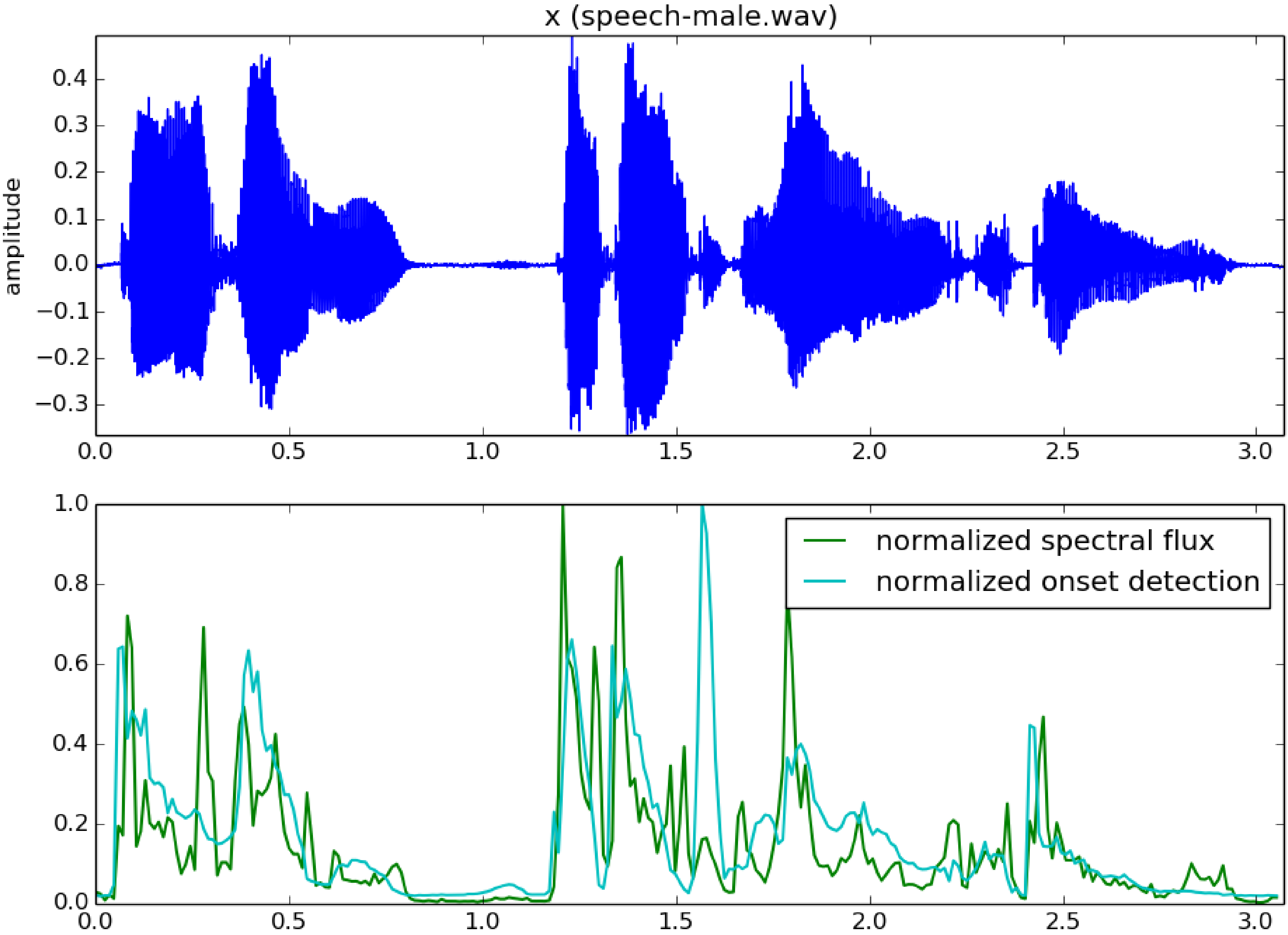
$$SF_l = \sum_{k=0}^{N/2} H\left(|X_l[k]| - |X_{(l-1)}[k]|\right)$$

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

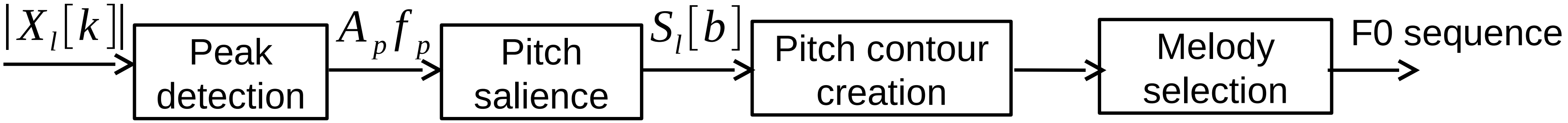
- Onset detection based on high-frequency content

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

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


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



$|X_l[k]|$

Peak detection

$A_p f_p$

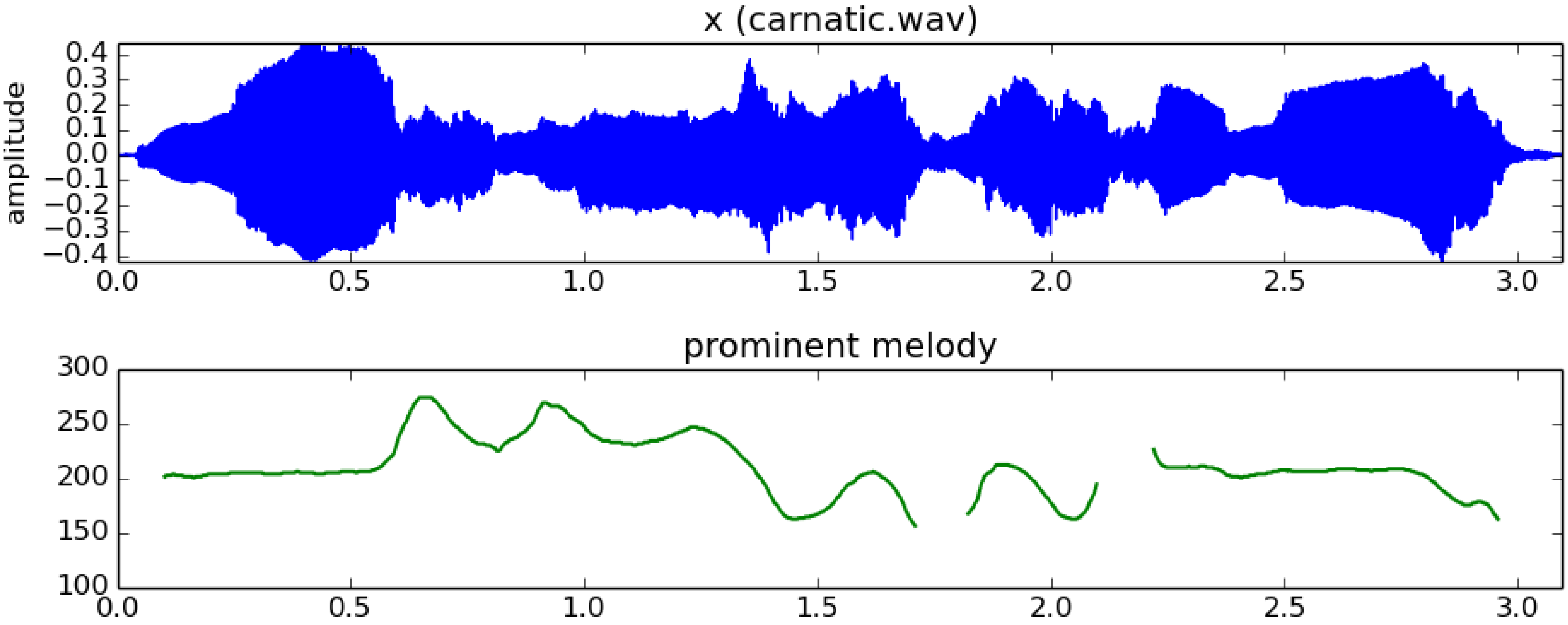
Pitch salience

$S_l[b]$

Pitch contour creation

Melody selection

F0 sequence



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

Universitat Pompeu Fabra, Barcelona