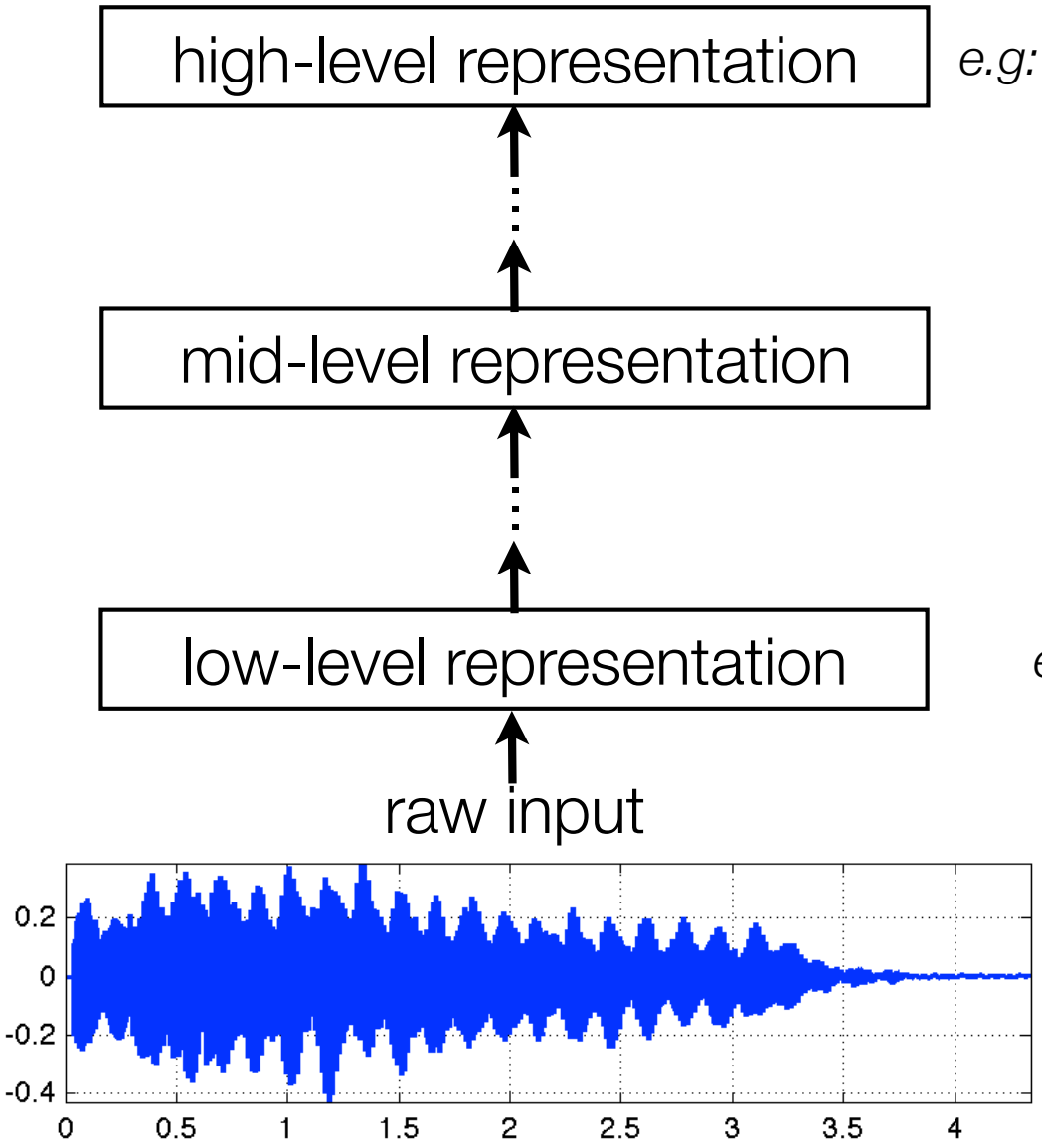
Low-level features and timbre

Juan Pablo Bello
EL9173 Selected Topics in Signal Processing: Audio Content Analysis
NYU Poly

Audio signal analysis



Low-level features

- The raw input data is often too large, noisy and redundant for analysis.
- Feature extraction: input signal is transformed into a new (smaller) space of variables that simplify analysis.
- Features: measurable properties of the observed phenomenon, usually containing information relevant for pattern recognition.
- They result from neighborhood operations on the input signal. If the operation produces a local decision -> feature detection.
- Usually one feature is not enough: combine several features into feature vectors, describing a multi-dimensional space.

Timbre

- Timbre: tonal qualities that define a particular sound/source. It can refer to, e.g., class (e.g. male, piano, truck), or quality (e.g. bright, rough, thin)

- Oftentimes defined comparatively: attribute that allows us to differentiate sounds of the same pitch, loudness, duration and spatial location (Grey, 75)

- Timbre spaces: empirically measure the perceived (dis)similarity between sounds and project to a low-dimensional space where dimensions are assigned a semantic interpretation (brightness, temporal variation, synchronicity, etc).

- Audio-based: recreate timbre spaces by extracting low-level features with similar interpretations (centroid, spectral flux, attack time, etc). Most of them describe the steady-state spectral envelope.

Temporal features

- The root-mean-square (RMS) level coarsely approximates loudness:

$$\mathrm{RMS}(m) = \sqrt{\frac{1}{N} \sum_{n=-N/2}^{N/2} \left(x(n+mh)\right)^2 w(n)}$$

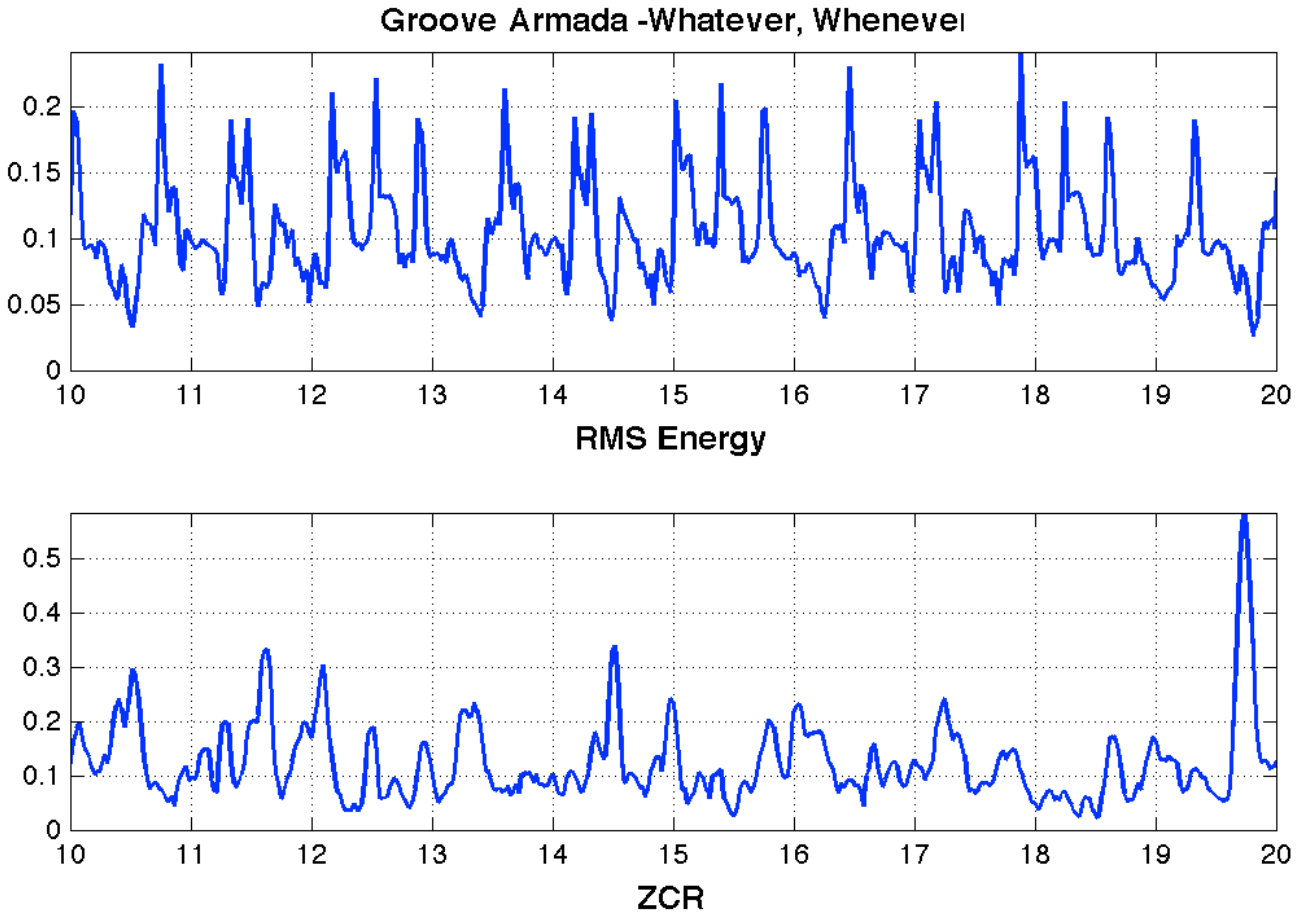
- Zero-crossing rate (ZCR) is a weighted measure of the number of times the signal changes sign in a frame:

$$\mathrm{ZCR}(m) = \tfrac{1}{2N} \textstyle\sum_{n=-N/2}^{N/2} |sgn(x(n+mh)) - sgn(x(n+mh-1))|$$

$$sgn(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

Temporal features

- ZCR is high for noisy (unvoiced) sounds and low for tonal (voiced) sounds
- For simple periodic signals, it is roughly related to the fundamental frequency



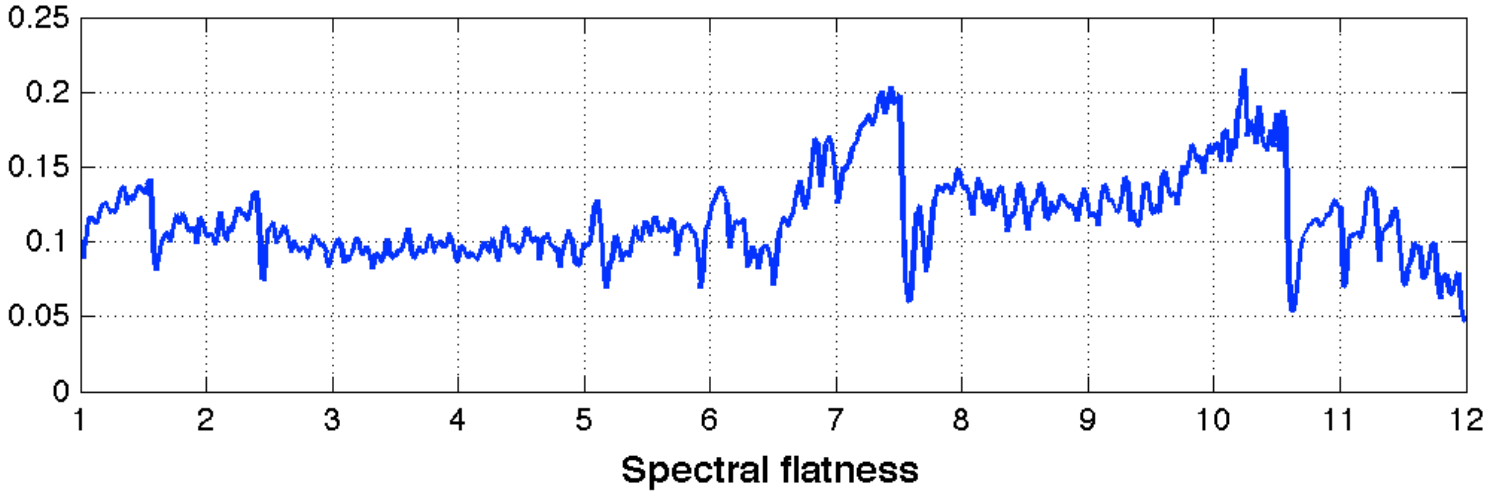
Spectral flatness

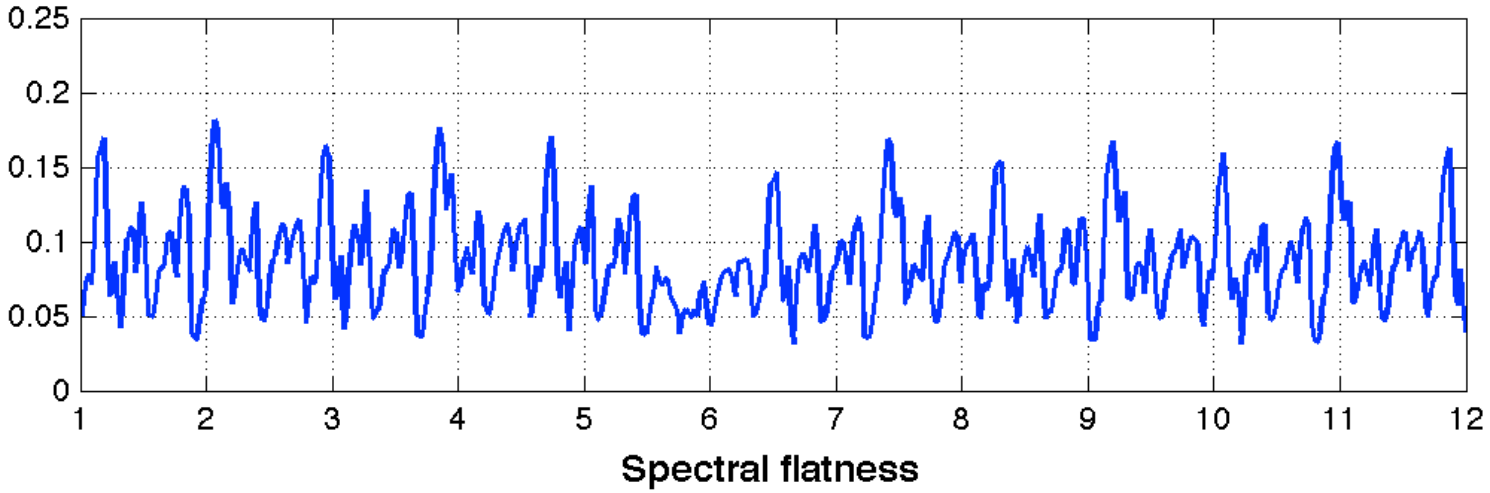
- Spectral flatness is a measure of the noisyness of the magnitude spectrum.
- It is the ratio between the geometric and arithmetic means:

$$\mathrm{SF}(m) = \frac{\left(\prod_k |X(m,k)|\right)^{\frac{1}{K}}}{\frac{1}{K}\sum_k |X(m,k)|}$$

- Different filterbanks can be used for pre-processing, s.t. k refers to band number and K to total number of bands.
- It is often used as a “tonality” coefficient (in dB)

Spectral flatness





Spectral features

- The most common is the spectral centroid (SC):

$$\mathrm{SC}(m) = \frac{\sum_k f_k |X(m,k)|}{\sum_k |X(m,k)|}$$

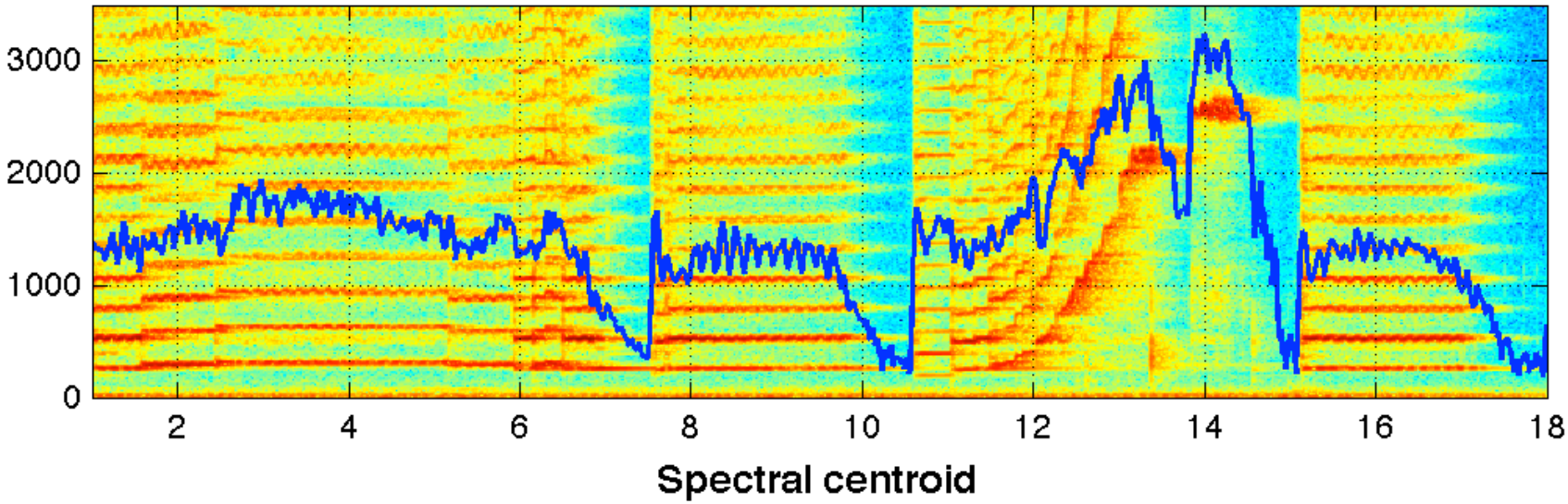
- It is usually associated with the sound's "brightness"

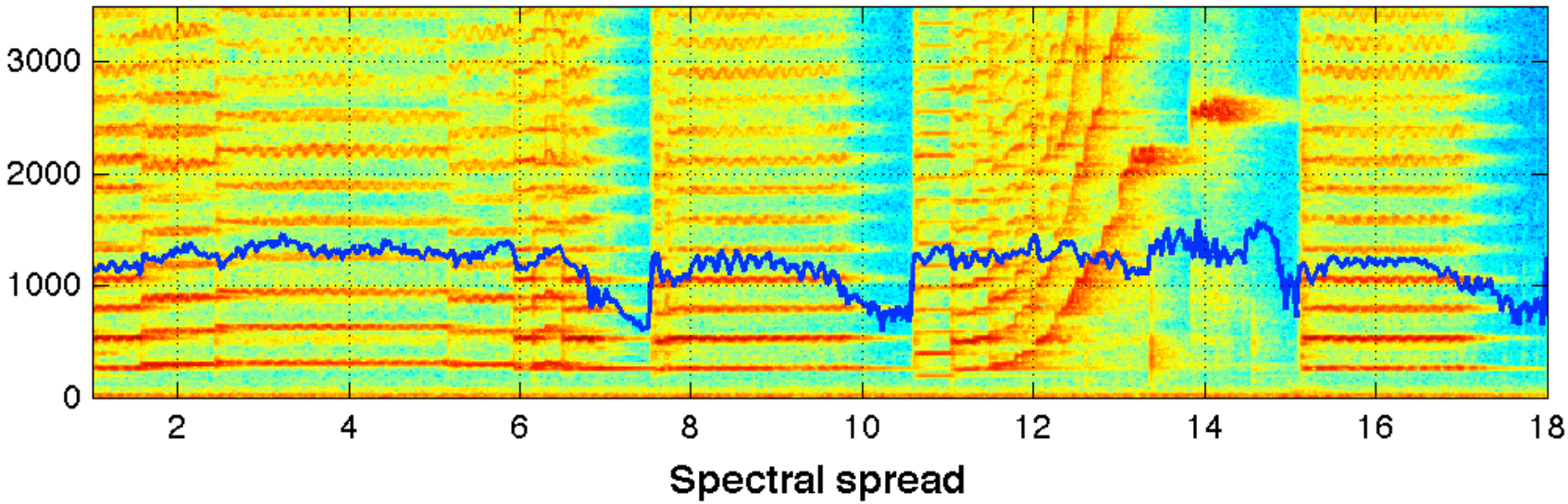
- Spectral spread (SS) is a measure of the bandwidth of the spectrum:

$$\mathrm{SS}(m) = \frac{\sum_k (f_k - SC(m))^2 |X(m,k)|}{\sum_k |X(m,k)|}$$

- Higher-order moments can be used to characterize the asymmetry and peakedness of the distribution

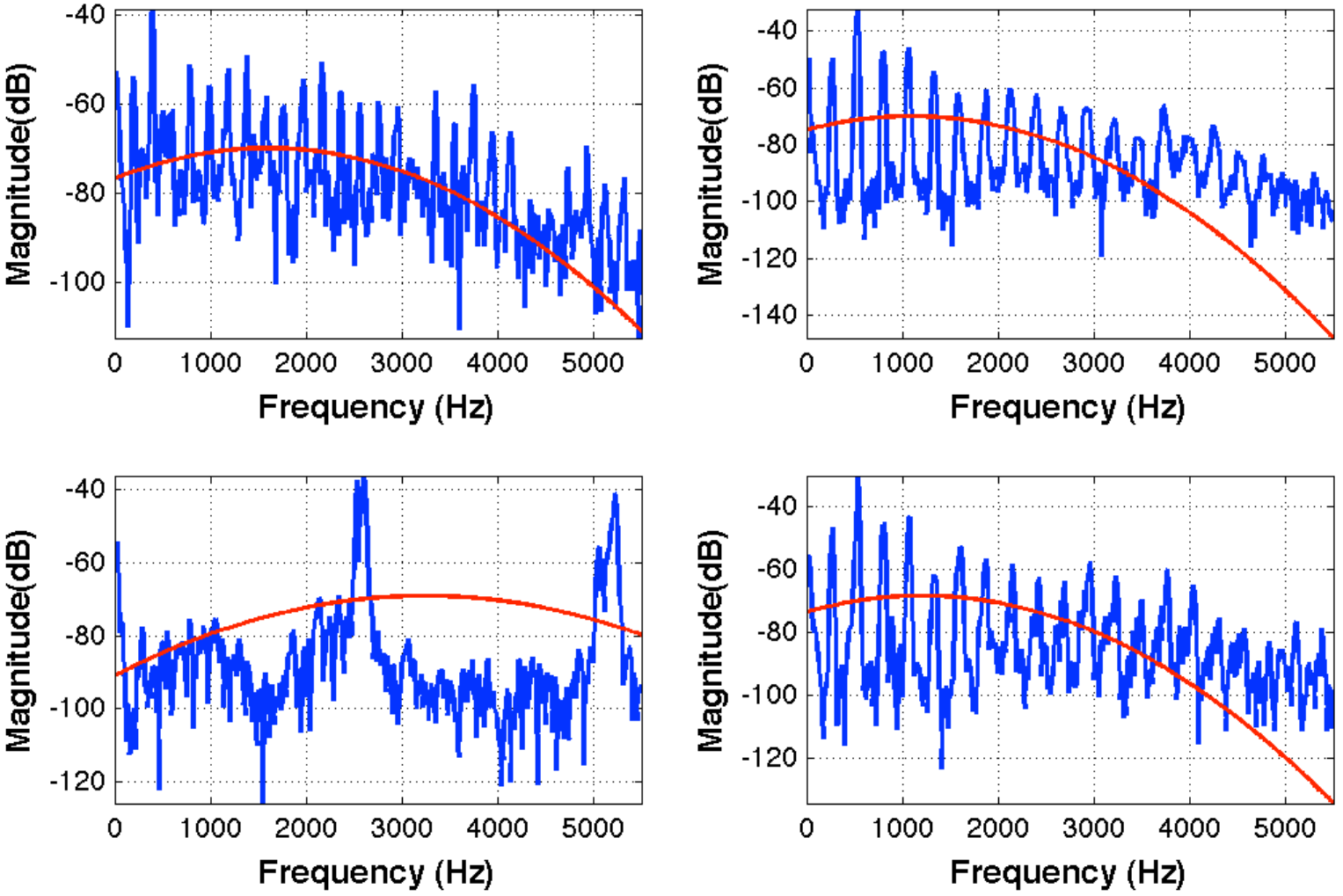
Spectral features





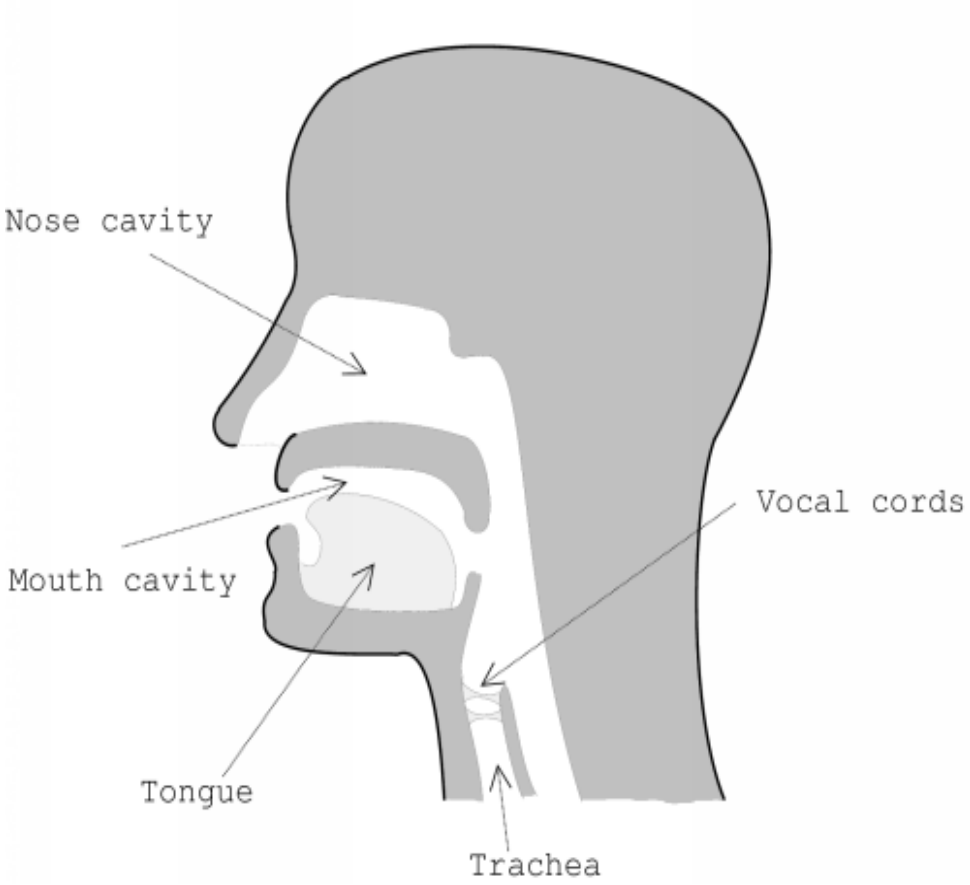
Spectral envelope

- SC and SS define a coarse (unimodal) model of the spectral envelope

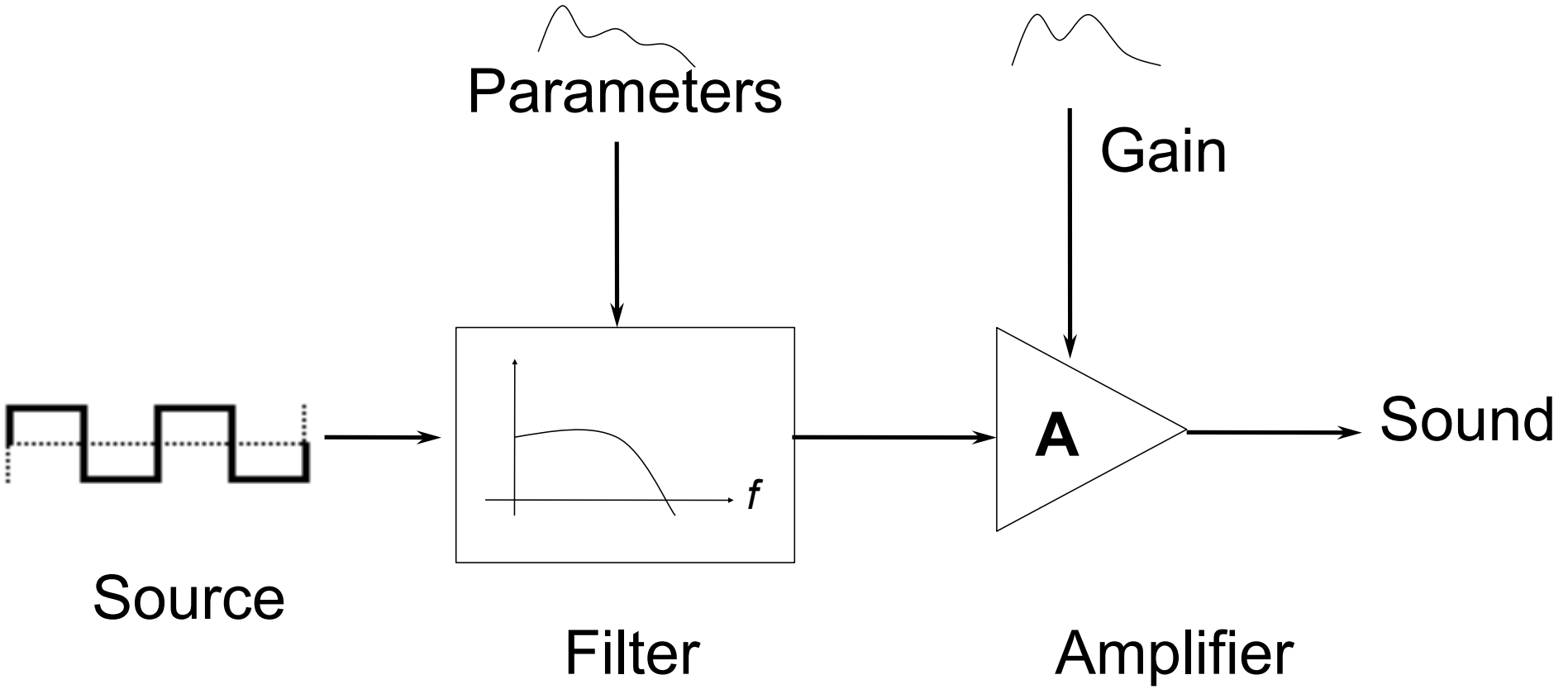


The human speech system

- The vocal chords act as an oscillator
- The mouth cavity, tongue and throat act as filters
- We can shape a tonal sound (‘oooh’ vs ‘aaah’)
- We can whiten the signal (‘sssshhh’)
- We can produce pink noise by removing high frequencies



Source-filter model



Channel Vocoder

- Decomposes the sound using a bank of bandpass filters + sums magnitude for each bandpass signal

- For a set of L-long filters w overlapped by L-1 bins:

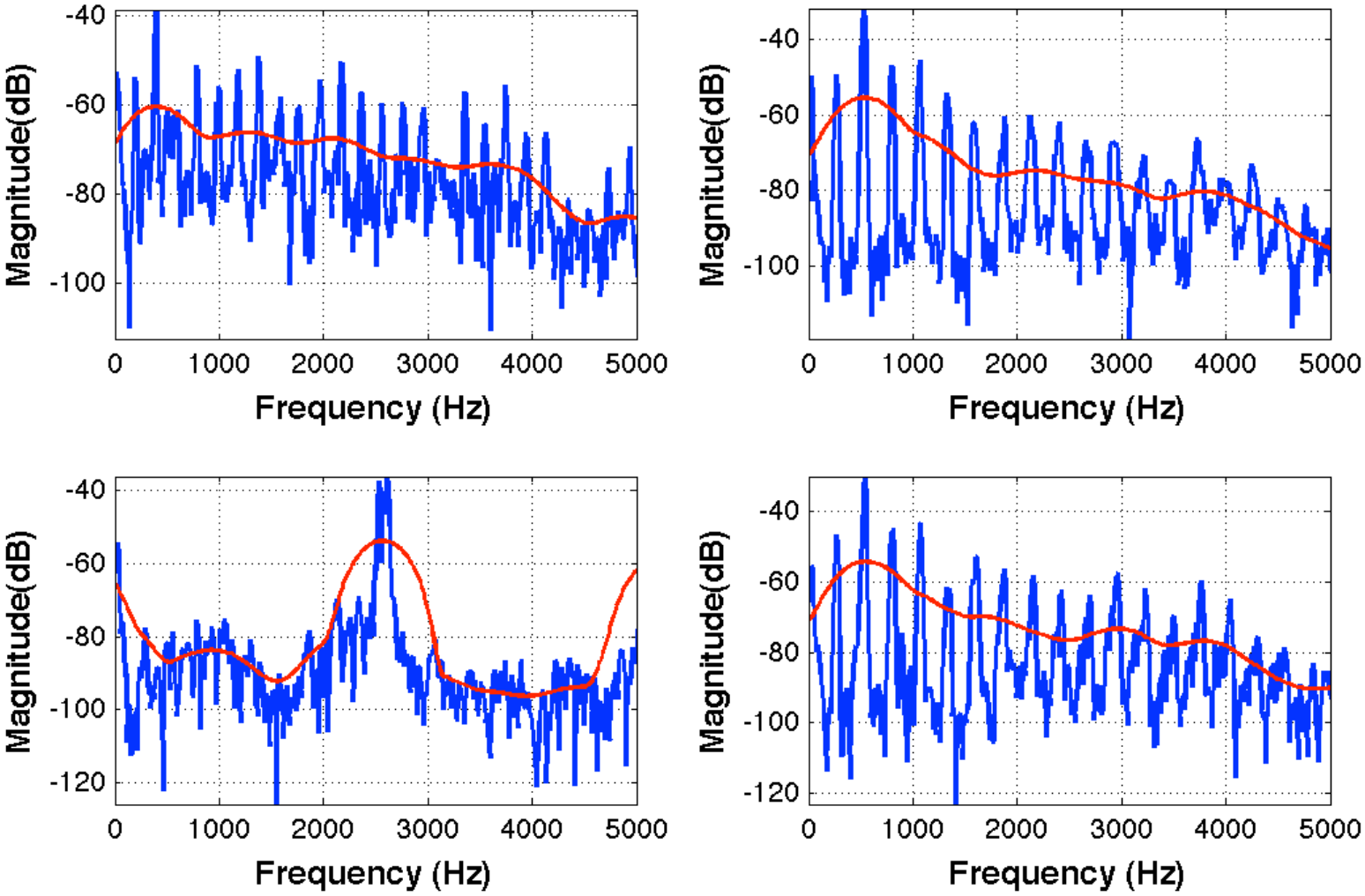
$$\mathrm{CV}(m) = |X(m,k)| * w(k)$$

$$\mathrm{CV}(m) = \mathbb{R}\left(\mathrm{IFFT}\left[\mathrm{FFT}(|X(m,k)|) \times \mathrm{FFT}(w(k))\right]\right)$$

- w(k) is normalized to unit sum, zero-padded to the length of X, and circularly shifted s.t. its center coincides with the first bin.

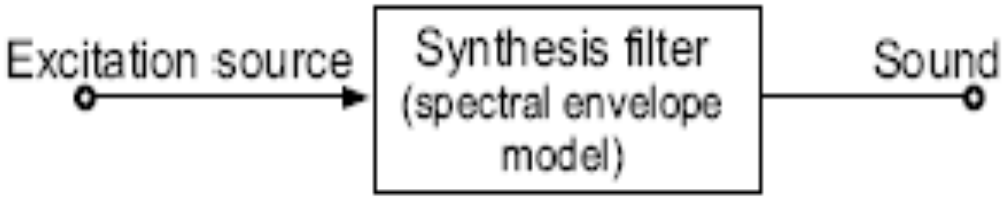
Channel Vocoder

- The spectral envelope approximation is coarser/finer depending on L



Linear Predictive Coding

- Linear predictive coding (LPC) is a source-filter analysis-synthesis methodology that approximates sound generation as an excitation (a pulse train or noise) passing through an all-pole resonant filter.



- Extensively used in speech and music applications. It reduces the amount of data to a few filter coefficients.

- It derives its name from the fact that output samples are predicted as a linear combination of filter coefficients and previous samples.

Linear Predictive Coding

- The input sample x(n) is extrapolated, i.e. approximated by a linear combination of past samples of the input signal:

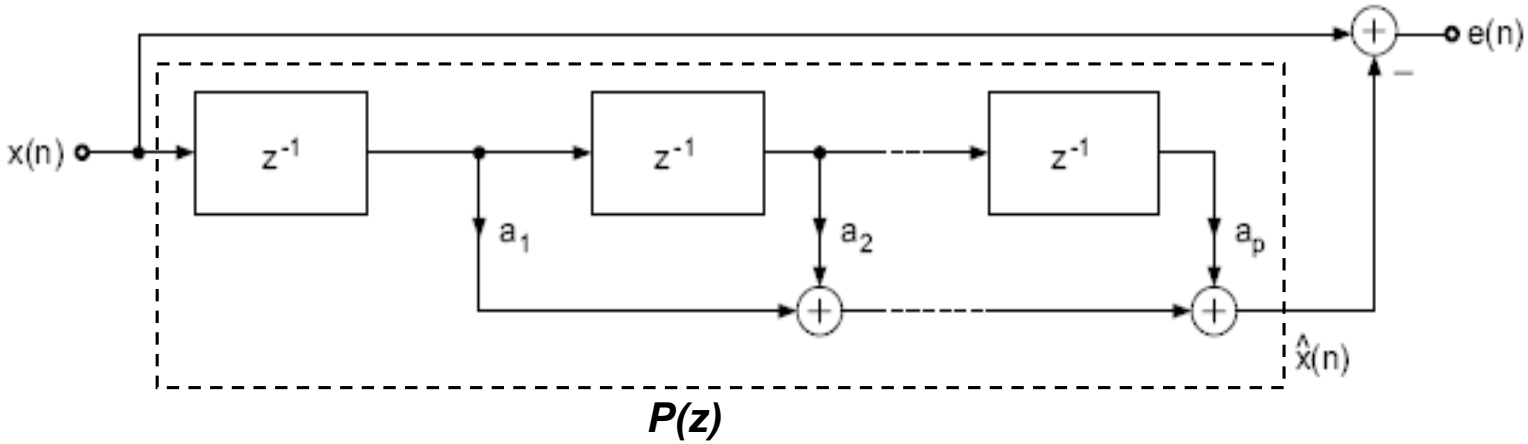
$$x(n) \approx \hat{x} = \sum_{k=1}^{p} a_k x(n-k)$$

- Because this is a prediction we always have a residual error:

$$e(n) = x(n) - \hat{x} = x(n) - \sum_{k=1}^{p} a_k x(n-k)$$

Linear Predictive Coding

- The prediction error calculation can be implemented by means of a FIR filter:



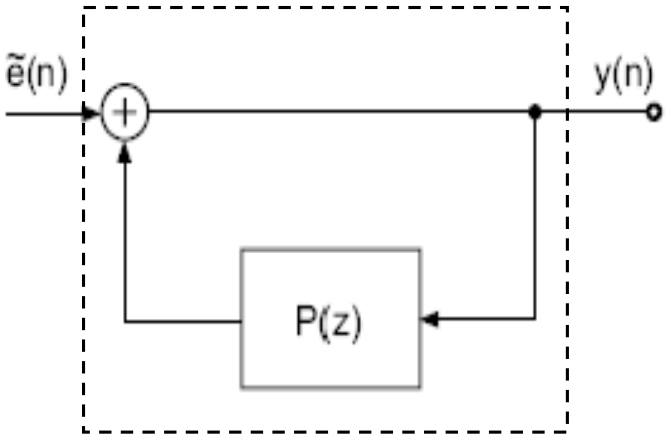
- The z-transform of the prediction filter is: $P(z) = \sum_{k=1}^{p} a_k z^{-k}$
- such that: $E(z) = Z(z)[1 - P(z)]$

Linear Predictive Coding

- The inverse filter can be defined as:

$$A(z) = 1 - P(z) = 1 - \sum_{k-1}^{p} a_k z^{-k} \quad \text{s.t.} \quad E(z) = X(z)A(z)$$

- For synthesis we use an approximation of the residual as the excitation used as input to the all-pole (LPC) filter, resulting on the model:



$$Y(z) = \tilde{E}(z)H(z)$$

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - P(z)}$$

Linear Predictive Coding

- The IIR filter H(z) is known as the LPC filter and represents the spectral model of x(n).
- With optimal coefficients, residual energy is minimized
- The higher the coefficient order p, the closer the approximation is to |X(k)|