

# Feature Vectors

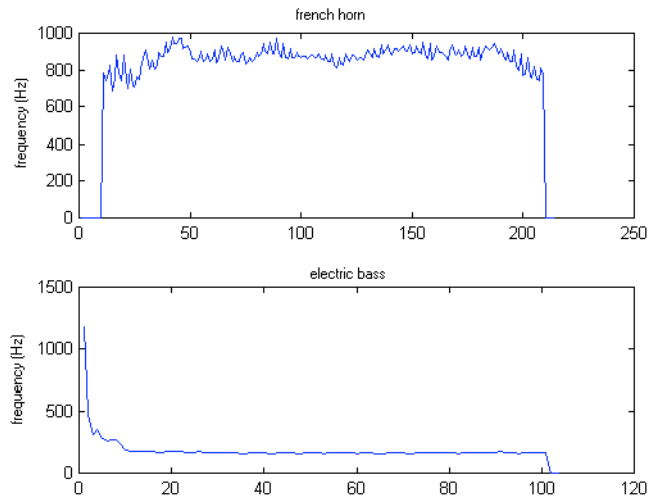
## Spectral low-level features

- Spectral low-level features aim at describing the structure of (frame or) sound spectra using a single quantity.
- They can be extracted in linear or logarithmic frequency domain, using spectral amplitudes, power values, logarithmic values, etc.
- The most common of these features is the spectral centroid (SC):

$$SC = \frac{\sum_{k=0}^{N/2} f_k |X(k)|^2}{\sum_{k=0}^{N/2} |X(k)|^2}$$

- It characterizes the centre of gravity of the (power) spectra.
- It is usually associated to the timbral sharpness of the sound and even to the concept of brightness.

## Spectral Centroid



Temporal evolution of the Spectral Centroid for 2 instrumental sounds

## Spectral Spread

- It is a measure of the average spread of the spectrum in relation to its centroid

$$SS = \sqrt{\frac{\sum_{k=0}^{N/2} (f_k - SC)^2 |X(k)|^2}{\sum_{k=0}^{N/2} |X(k)|^2}}$$

- For noisy sounds you would expect SS to be high, while tonal, less broadband, sounds will show a lower SS

## Spectral flatness

- It reflects the flatness properties of the power spectrum
- It is calculated as the ratio between the geometric and arithmetic mean

$$SF_b = \frac{\sqrt[N_b]{\prod_{k_b} |X(k_b)|^2}}{\frac{1}{N_b} \sum_{k_b} |X(k_b)|^2}$$

- It is calculated per spectral band, so that  $k_b$  in the above formula goes from the lower ( $k_l$ ) to the upper ( $k_u$ ) edge of the band, and  $N_b = k_u - k_l + 1$ . The flatness for the whole spectrum (SF) is thus the average of the sub-band flatness values

## Using harmonic data

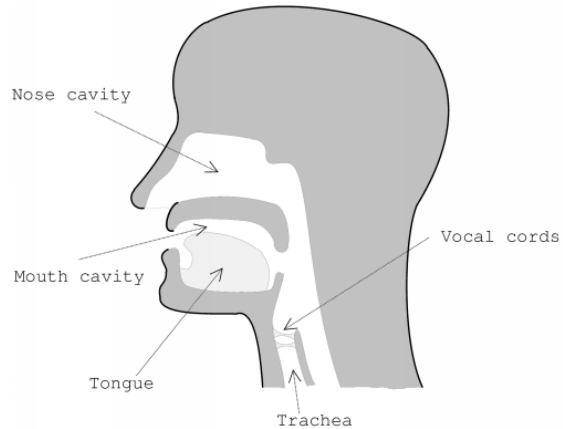
- Since SMS and its variations provide methods for harmonic peak information and fundamental frequency to be known, we can also use these low-level descriptors to characterize harmonic spectra.
- Assuming the frequency and amplitude of harmonic peaks to be known, we can (for example) calculate the harmonic spectral centroid (HSC) as the amplitude-weighted mean of the harmonic peaks of the spectrum:

$$HSC = \frac{\sum_{h=1}^{N_h} f_h A_h}{\sum_{h=1}^{N_h} A_h}$$

- Where  $f_h$  and  $A_h$  are, respectively, the frequency and amplitude of the  $h^{th}$  harmonic. The average of HSC over the duration of the signal results in a single summarizing quantity describing this property for the whole sound.

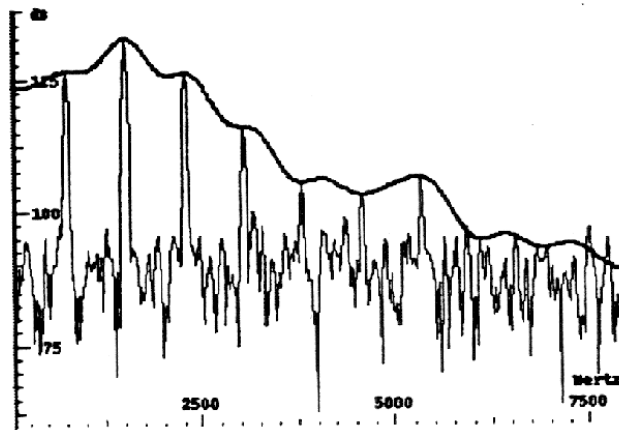
## The human speech system

- 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 ('ssshhh')
- We can produce pink noise by removing high frequencies



## What is the spectral envelope?

- It is a smoothing of the spectrum that preserves its general form while neglecting its spectral line structure

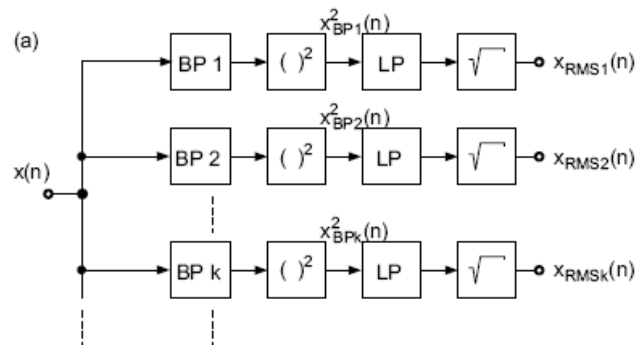


## Estimation of the spectral envelope

- There are a number of possible techniques:
  1. The channel vocoder: estimates the amplitude of the signal inside a few frequency bands
  2. Linear prediction: estimates the parameters of a filter that matches the spectrum of the sound.
  3. Cepstrum analysis: smoothes the logarithm of the spectrum and low-pass filters it to obtain the envelope.

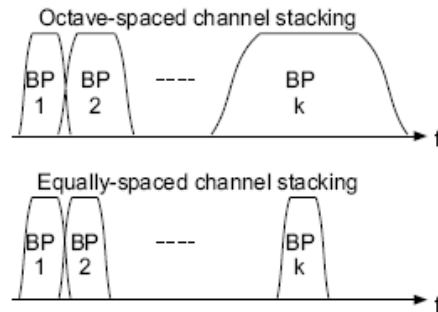
## Channel Vocoder (1)

- Filters the sound with a bank of BP filters
- Calculates RMS for each bandpass signal
- The more filters we use, the more frequency points of the spectrum we estimate



## Channel Vocoder (2)

- In the frequency domain: square root of the sum of the multiplication between FFT bin energies with the filter's frequency response.



- The filter bank can be defined on a logarithmic scale (e.g. constant-Q filter bank)
- Can be defined on a linear scale (equal bandwidth)

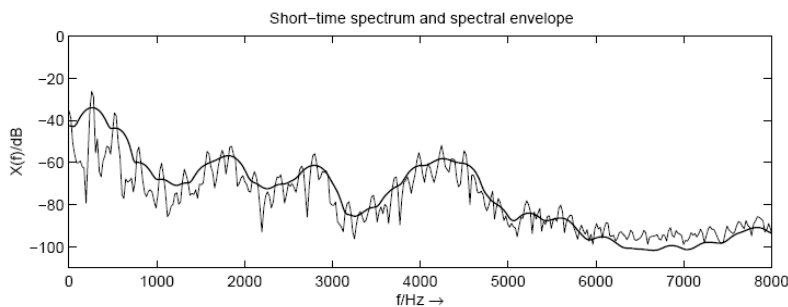
## Channel Vocoder (3)

- For a linearly spaced filterbank we can perform a circular convolution:

$$Y(k) = \sqrt{|X(k)|^2 * w(k)}$$

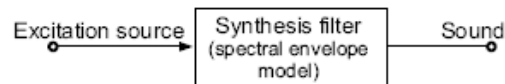
- A very quick implementation uses FFT/IFFT to perform the circular convolution

$$Y(k) = \sqrt{\Re\left(\text{IFFT}\left(\text{FFT}\left(|X(k)|^2\right) \cdot \text{FFT}(w(k))\right)\right)}$$



## 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 (2)

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

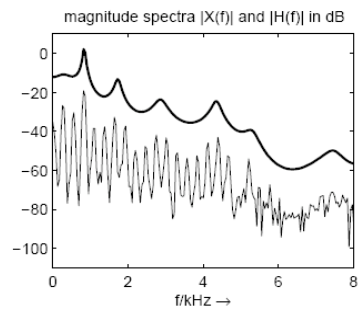
$$x(n) \approx \hat{x}(n) = \sum_{k=1}^p a_k x(n-k)$$

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

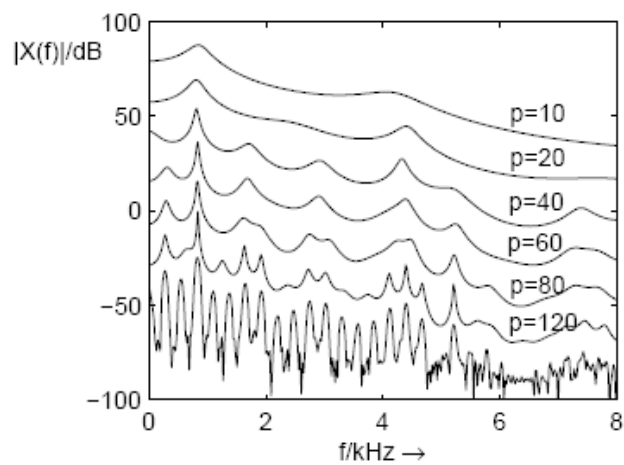
$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{k=1}^p a_k x(n-k)$$

## Linear Predictive Coding (3)

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



## LPC order and residual





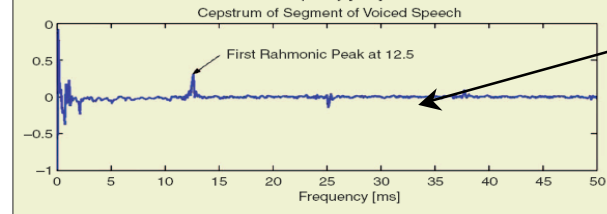
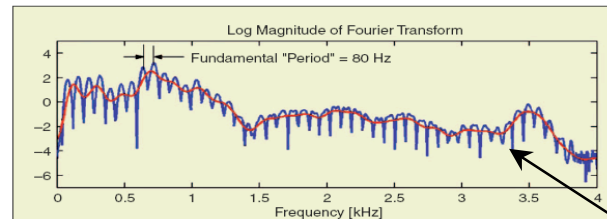
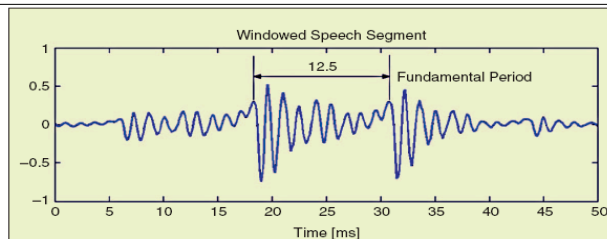
# Cepstrum

- Is the result of taking the Fourier Transform of the log Spectrum as if it were a signal
- It measures the rate of change in the different spectral bands
- The name cepstrum, first introduced by Bogert et al (1963) is an anagram of spectrum (they also introduced the terms quefrency, liftering, saphe, alanalysis, etc)
- For a real signal  $x(n)$ , the Cepstrum is calculated as:

$$c(n) = \text{IFFT}(\log|X(n,k)| + j\varphi(k))$$

- The real spectrum ignores the complex component and is therefore:

$$c_r(n) = \text{IFFT}(\log|X(n,k)|)$$



By low-pass "liftering" the cepstrum we obtain the spectral envelope of the signal

# Cepstrum

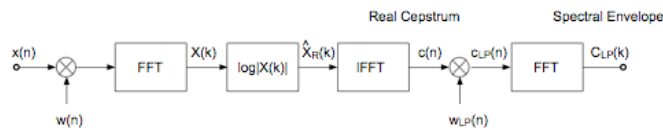
- The real cepstrum can be weighted using a low-pass window of the form:

$$\omega_{LP}(n) = \begin{cases} 1 & n = 0, N_1 \\ 2 & 1 \leq n \leq N_1 \\ 0 & N_1 < n \leq N-1 \end{cases}$$

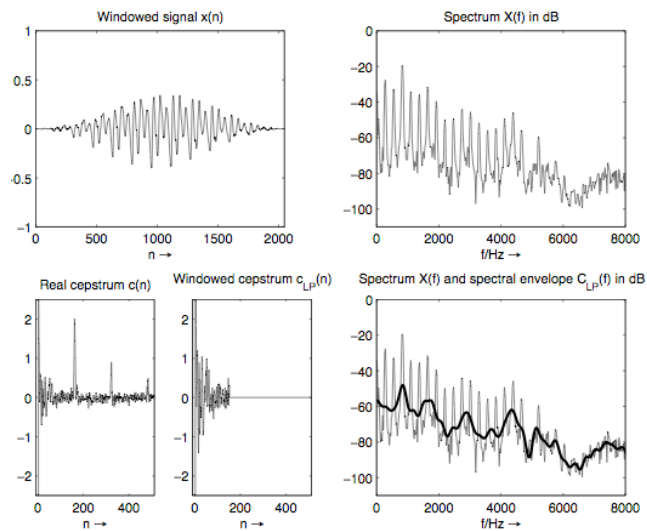
- With  $N_1 \leq N/2$ , such that the low-pass filtered spectrum and the spectral envelope can be obtained by:

$$c_{LP}(n) = c_r(n) \cdot \omega_{LP}(n)$$

$$C_{LP}(k) = FFT[c_{LP}(n)]$$

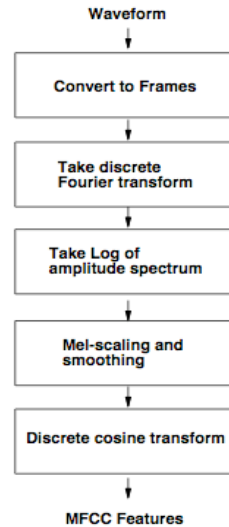


# Cepstrum



## MFCC

- Mel-Frequency Cepstral Coefficients are an interesting variation on the linear cepstrum, which are widely used in speech and music analysis.
- They are the most widely used features in speech recognition, mainly due to their ability to compactly represent the audio spectrum (only ~13 coefficients)
- The steps performed on their computation are motivated by perceptual or computational considerations (Logan, 2000)

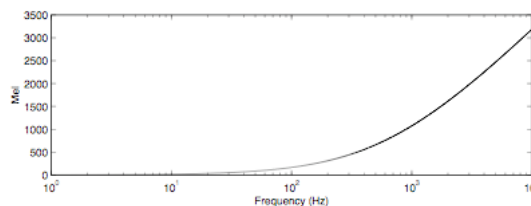


## MFCC

- The Mel scale is a non-linear perceptual scale of pitches judged to be equidistant.
- The scale is approximately linear below 1kHz and logarithmic above, the reference point is a 1kHz tone, which is equated to 1000 Mel
- A tone perceived to be half as high is defined to have 500 Mel, while a tone twice as high is defined to have 2000 Mel

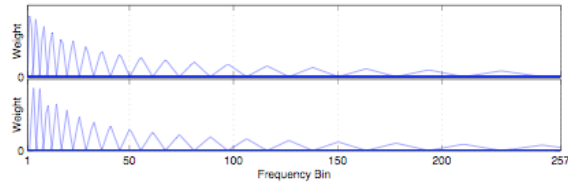
$$m = 1127.01048 \log(1 + f / 700)$$

$$f = 700(e^{m/1127.01048} - 1)$$

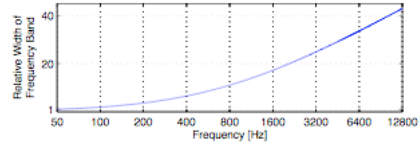


## MFCC

- To convert a linear spectrum to Mel we can use a filterbank of overlapping triangular windows:



- Such that the width  $d$  of each window increases according to the Mel scale, and the height of each triangle is  $2/d$

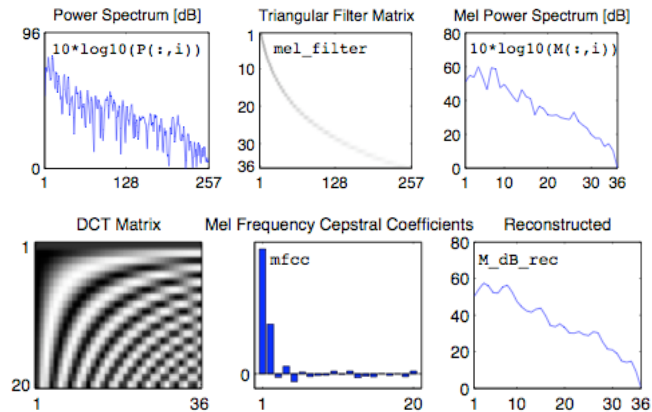


## MFCC

- The resulting Mel spectral vectors are highly correlated with each other, i.e. highly redundant
- Thus a more efficient representation of the log-spectrum can be obtained by applying a transform that decorrelates those vectors (see Rabiner and Juang, 93).
- This decorrelation, which can be achieved using Principal Component Analysis (PCA) is commonly approximated by means of the Discrete Cosine Transform (DCT).
- The DCT is similar to a DFT but only for real numbers. It has the property that most of its energy is concentrated on a few initial coefficients (thus effectively compressing the spectral info)

$$X_{DCT}(k) = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} x(n) \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right]$$

## MFCC



- MFCC roughly model certain characteristics of human audition: the non-linear perception of loudness and frequency and spectral masking (Pampalk, 2006)

## Useful references

- Zölzer, U. (Ed). "DAFX: Digital Audio Effects". John Wiley and Sons (2002)
  - Chapter 10: Amatriain, X., Bonada, J., Loscos, A. and Serra, X. "Spectral Processing".
- Kim, H-G., Moreau, N. and Sikora, T. "MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval". John Wiley & Sons (2005)
  - Chapter 2: "Low-Level Descriptors"
- The MPEG-7 standard: <http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>
- Zölzer, U. (Ed). "DAFX: Digital Audio Effects". John Wiley and Sons (2002)
  - Chapter 8: Arfib, D., Keiler, F. and Zölzer, U., "Source-filter Processing".
  - Good read, Chapter 2: Dutilleux, P. and Zölzer, U. "Filters"
- Pampalk, E. "Computational Models of Music Similarity and their Application in Music Information Retrieval". PhD Thesis, Vienna University of Technology, Vienna, Austria (2006). PDF available at: <http://staff.aist.go.jp/elias.pampalk/mir-phds/>
- Logan, B. "Mel Frequency Cepstral Coefficients for Music Modeling", Proceedings of the ISMIR International Symposium on Music Information Retrieval, Plymouth, MA (2000).
- [http://en.wikipedia.org/wiki/Discrete\\_cosine\\_transform](http://en.wikipedia.org/wiki/Discrete_cosine_transform)