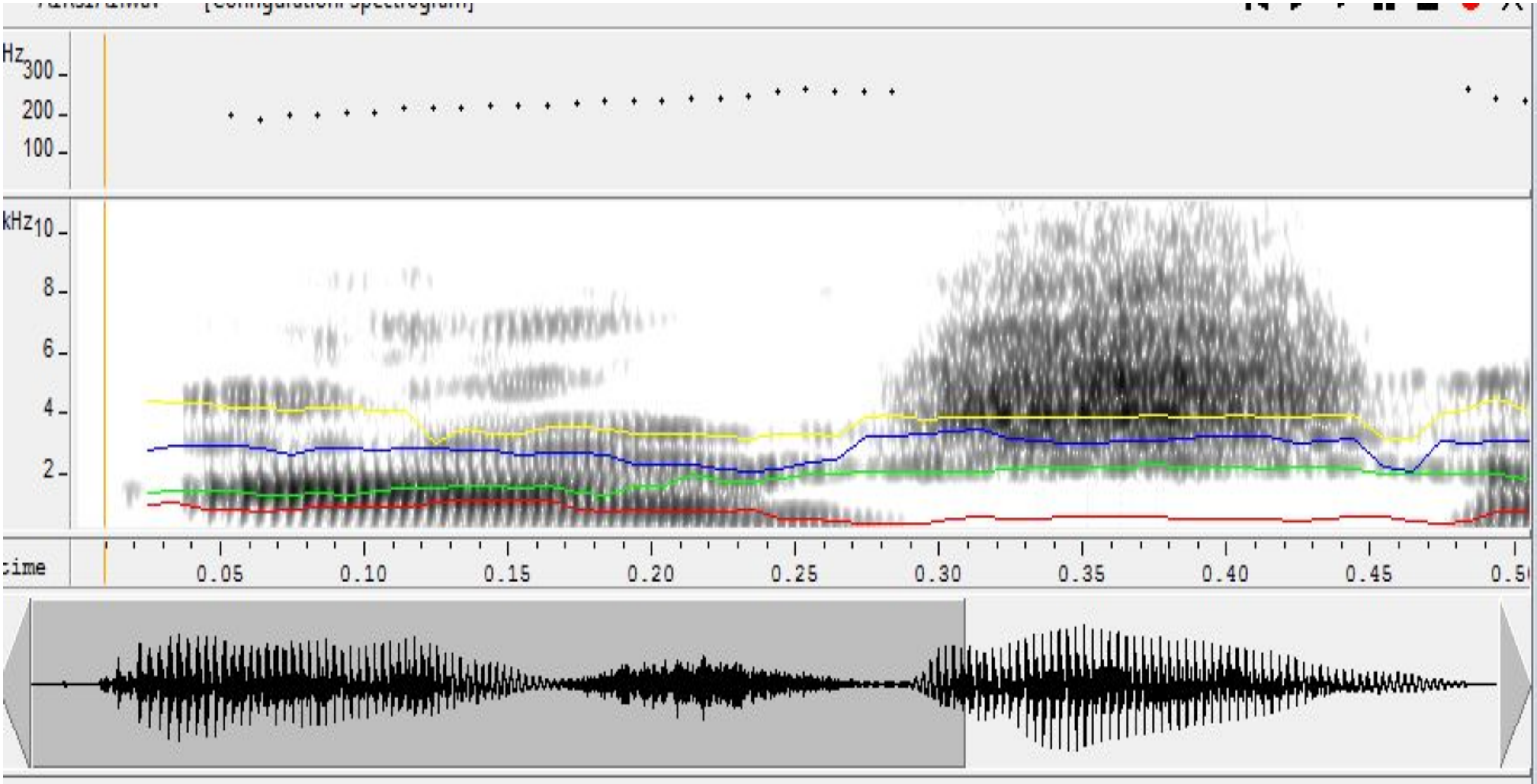
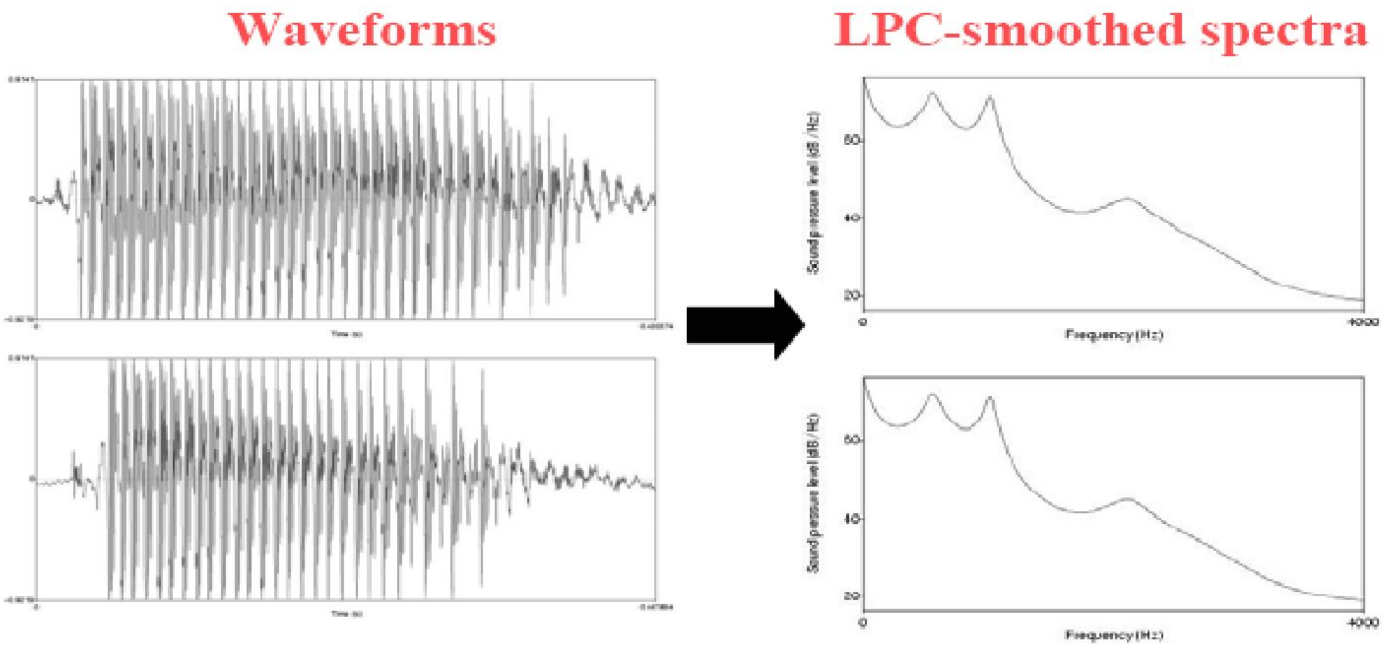
Features Extraction


Hz
300
200
100
kHz
10
8
6
4
2
time
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45

Why do we need feature extraction?

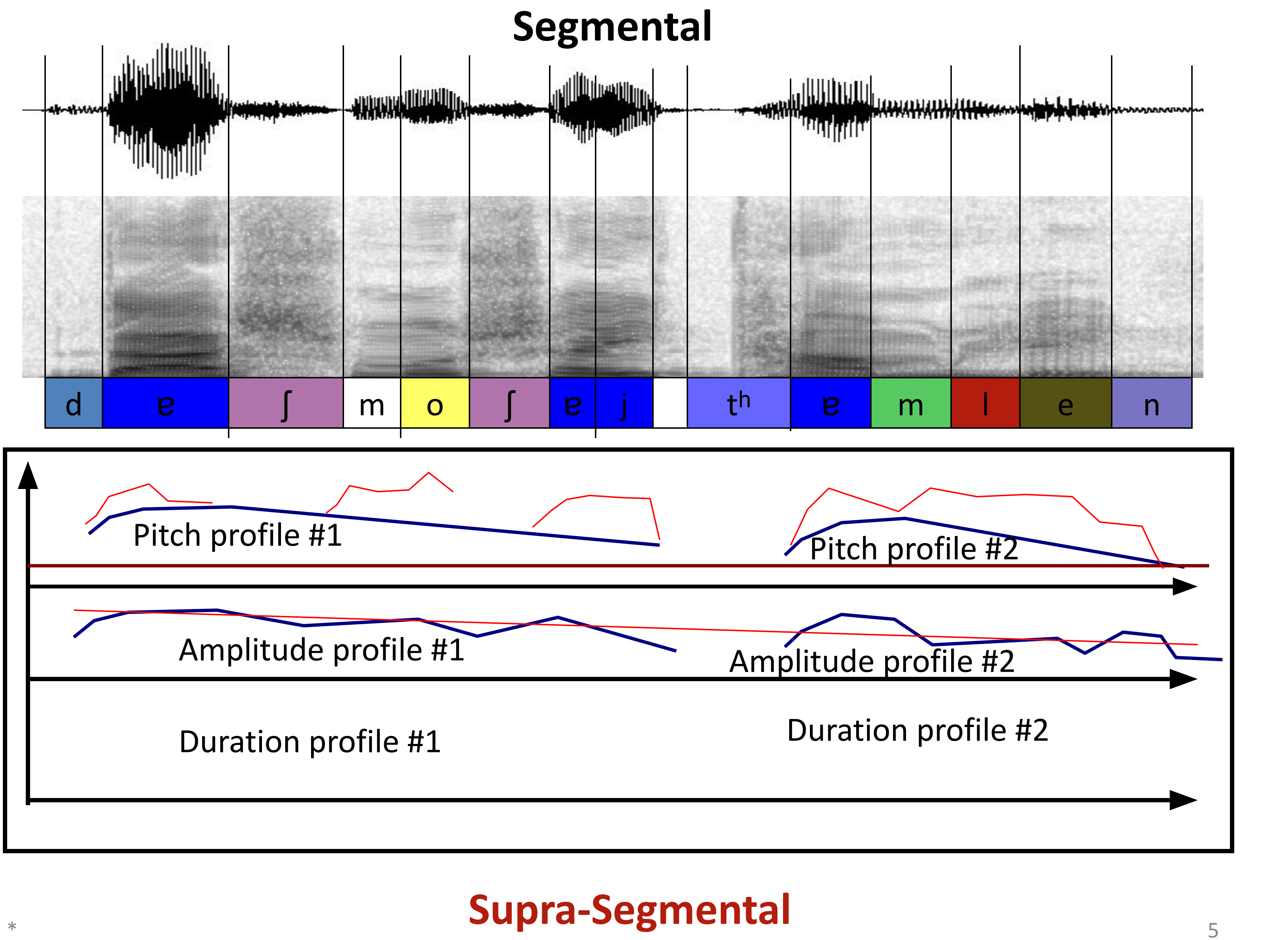
- Acoustic speech signal varies over time. Can't compare two waveforms

example: two instances of /a:/ vowel spoken in isolation, with time interval between repetitions < 1 second:



What is Features?

- Feature = a measure of a property of the speech waveform
- Reasons for feature extraction:
  - Redundancy and harmful information is removed
  - Reduced computation time
  - Easier modeling of the feature distribution
- Speech has many "natural" (Acoustic-phonetic) features:
  - Fundamental frequency (F0), formant frequencies, formant bandwidths, spectral tilt, intensity, phone durations, articulation, etc
- Not-so-natural features:
  - Cepstrum, linear predictive coefficients, line spectral frequencies, vocal tract area function, delta and double-delta coefficients, etc


Segmental
d
ɐ
ʃ
m
o
ʃ
ɐ
j
tʰ
ɐ
m
l
e
n
Pitch profile #1
Pitch profile #2
Amplitude profile #1
Amplitude profile #2
Duration profile #1
Duration profile #2
Supra-Segmental

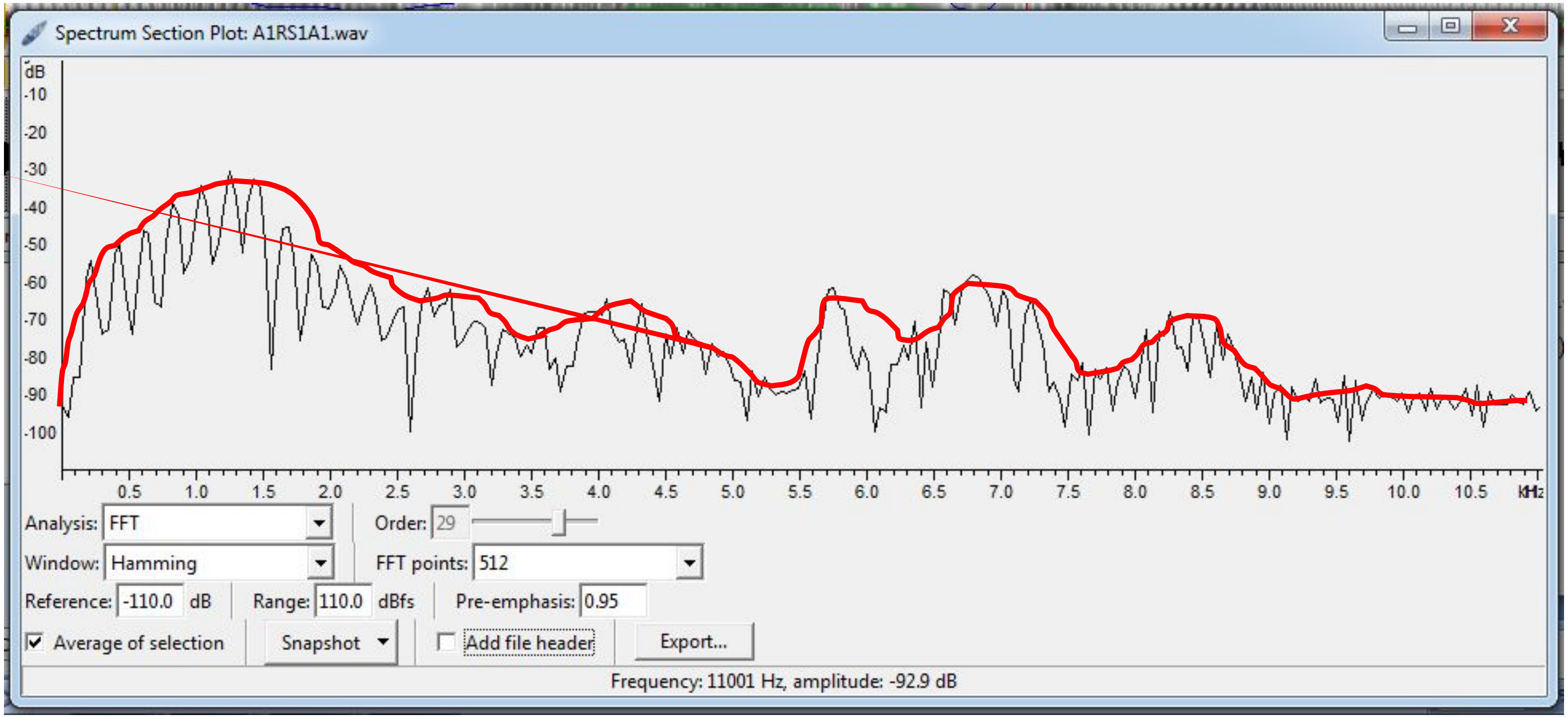
Speech Events

Segmental

Supra-segmenta
l

Supra-segmental features and Prosody

- Intonation, pause, duration, stress together are called prosodic or supra-segmental features and may be considered as the melody, rhythm, and emphasis of the speech at the perceptual level.
- The prosody of a sentence is important for naturalness and for conveying the correct meaning of a sentence.



- Peaks denote dominant frequency components in the speech signal
- Peaks are referred to as formants
- Formants carry the identity of the sound

Parameter / Feature Classification

Frequency Domain Parameters

- Filter Bank Analysis
- Short-term spectral analysis
- Cepstral Transfer Coefficient (CC)
- Formant Parameters
- MFCC, Delta MFCC, Delta-Delta MFCC

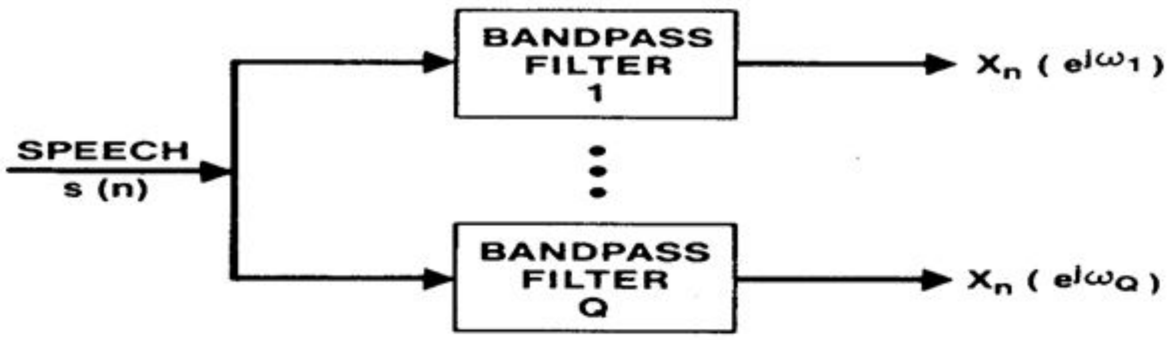
Time Domain Parameters

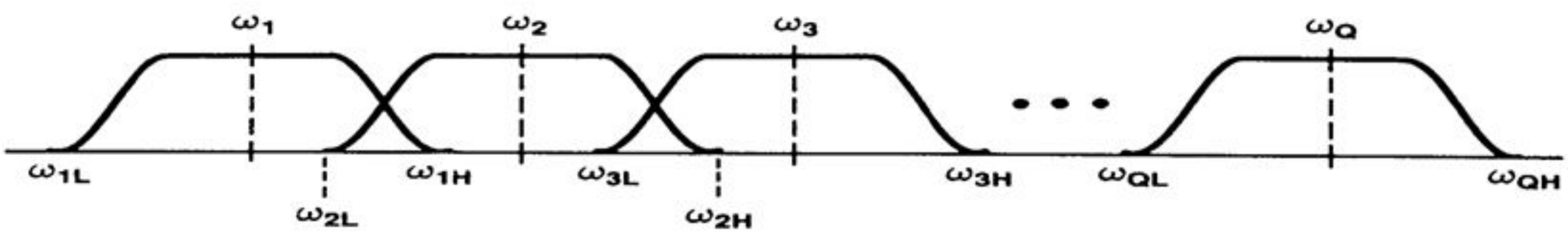
- LPC
- Shape Parameters

Time- Frequency Domain Parameters

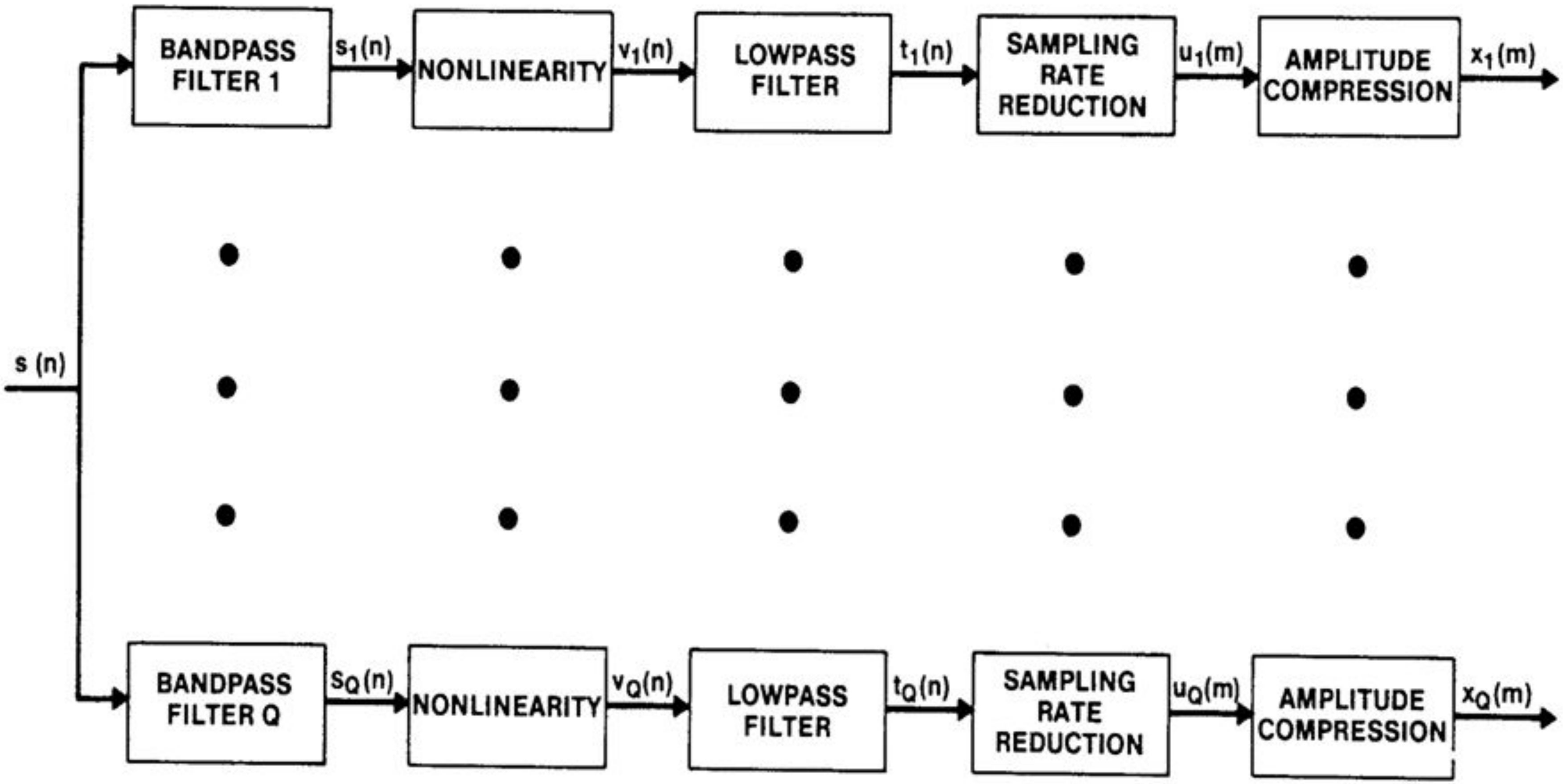
- Perceptual Linear Prediction (PLP):
- Wavelet Analysis

Filter Bank Analysis





Complete Filter Bank Analysis Model

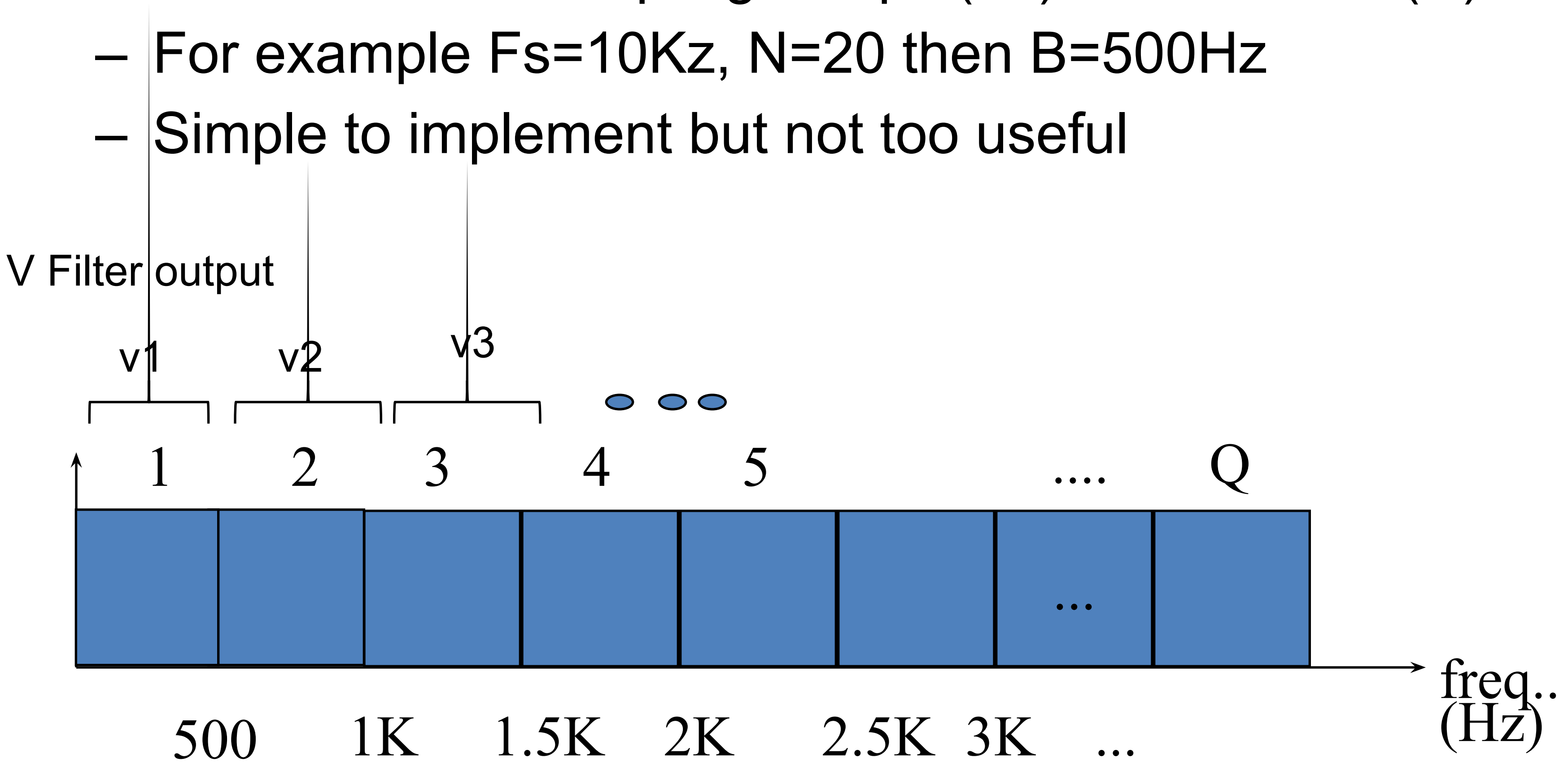


How to determine filter band ranges

- Uniform filter banks
- Log frequency banks
- Mel filter bands

Uniform Filter Banks

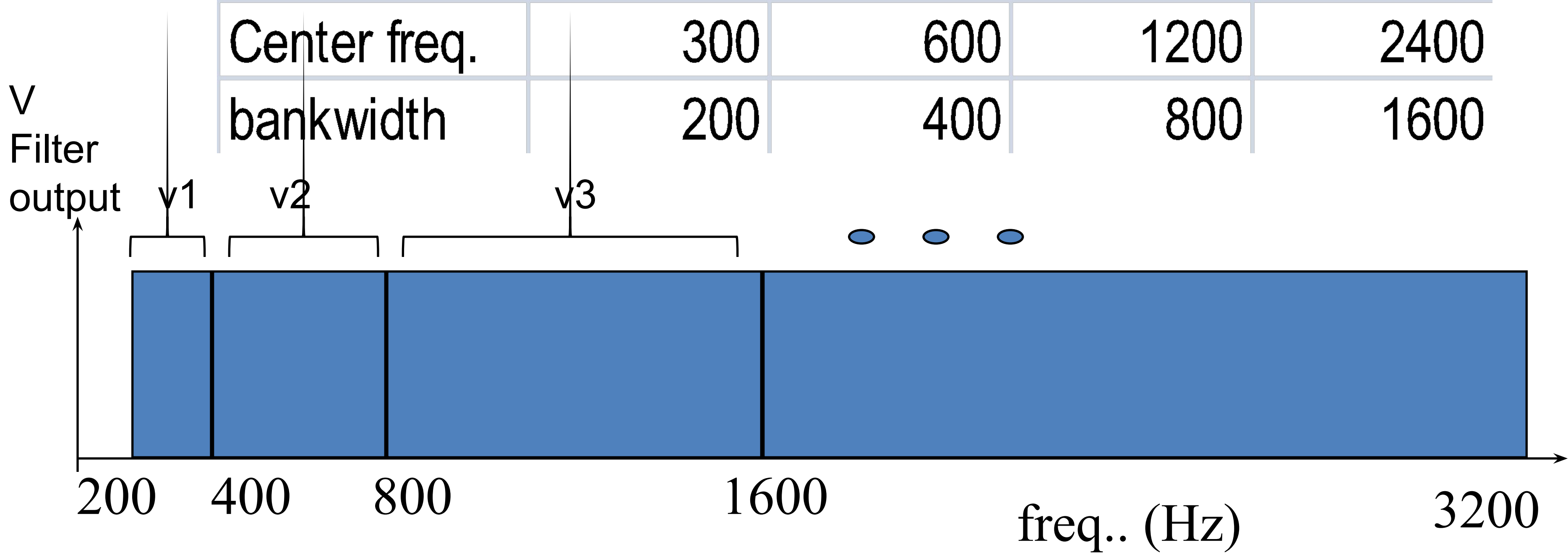
- Uniform filter banks
  - bandwidth B= Sampling Freq... (Fs)/no. of banks (N)
  - For example Fs=10Kz, N=20 then B=500Hz
  - Simple to implement but not too useful



Non-uniform filter banks: Log frequency

- Log. Freq... scale : close to human ear

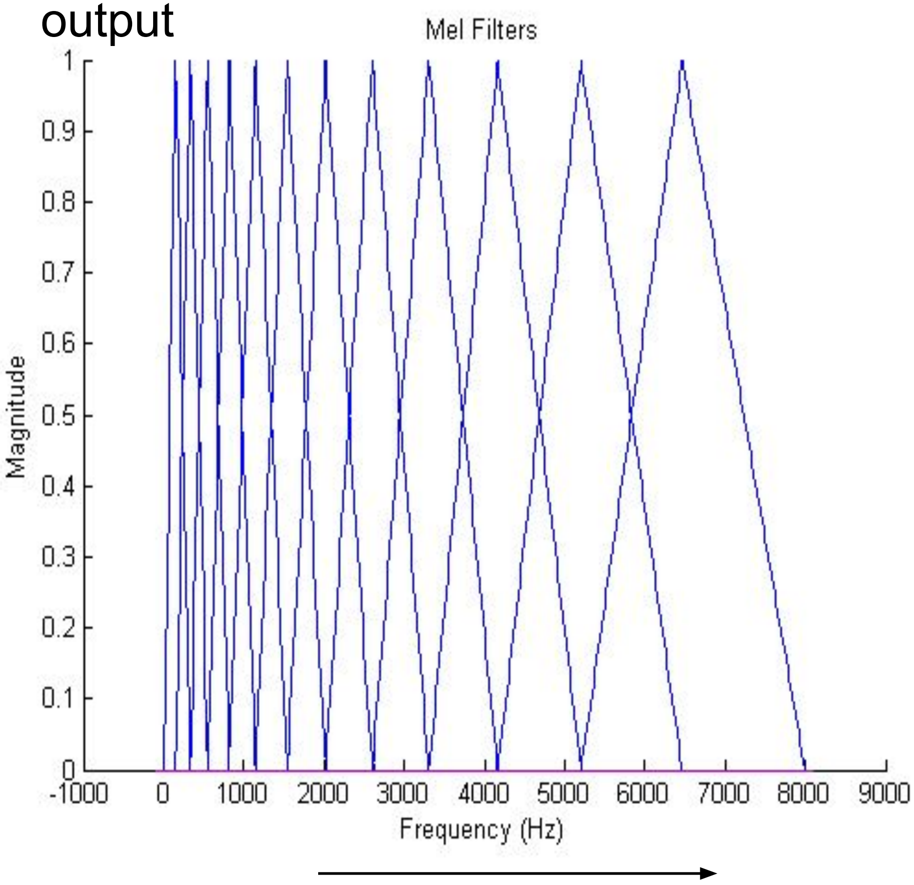
	filter 1	filter 2	filter 3	filter 4
Center freq.	300	600	1200	2400
bankwidth	200	400	800	1600

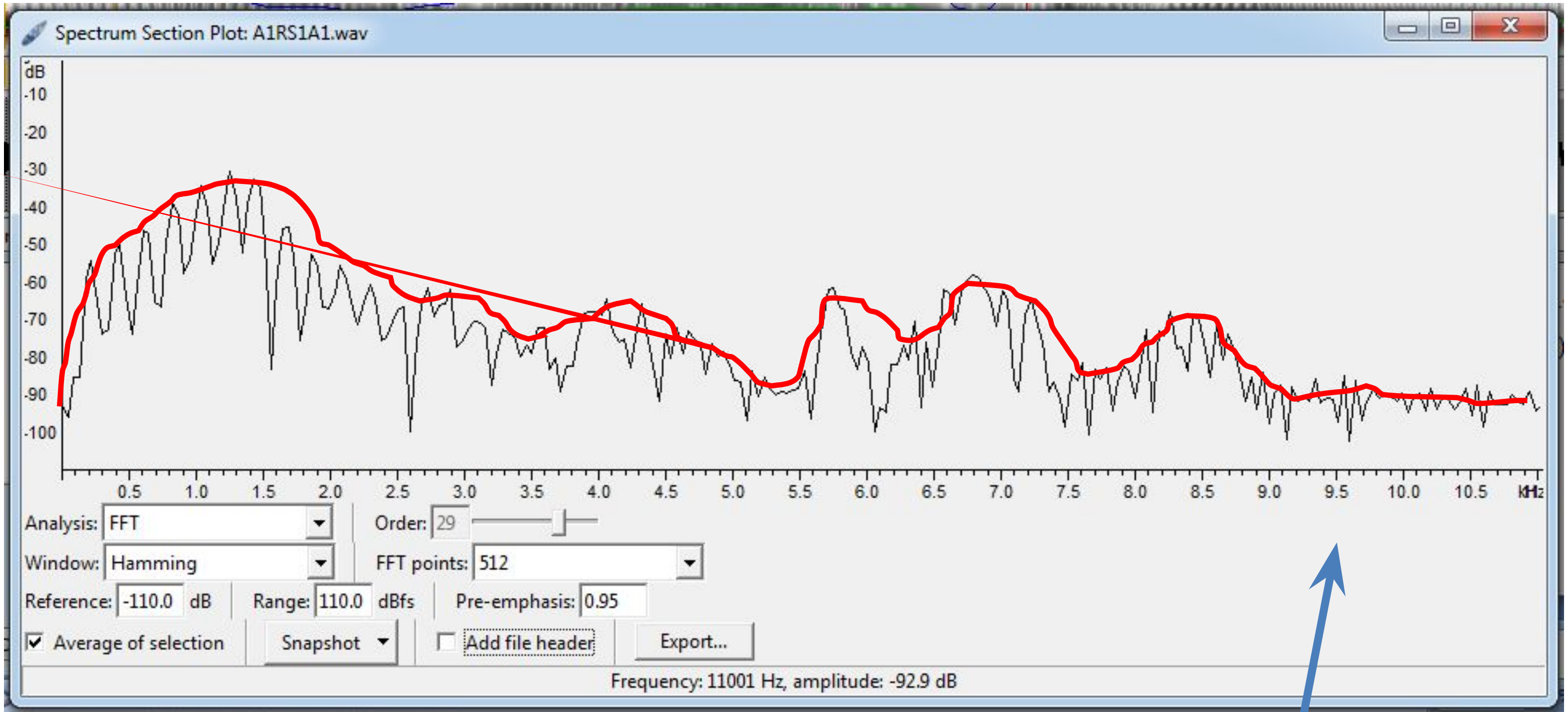


Mel filter bands

- Freq. lower than 1 KHz has narrower bands (and in linear scale)
- Higher frequencies have larger bands (and in log scale)
- More filter below 1KHz
- Less filters above 1KHz

Filter output

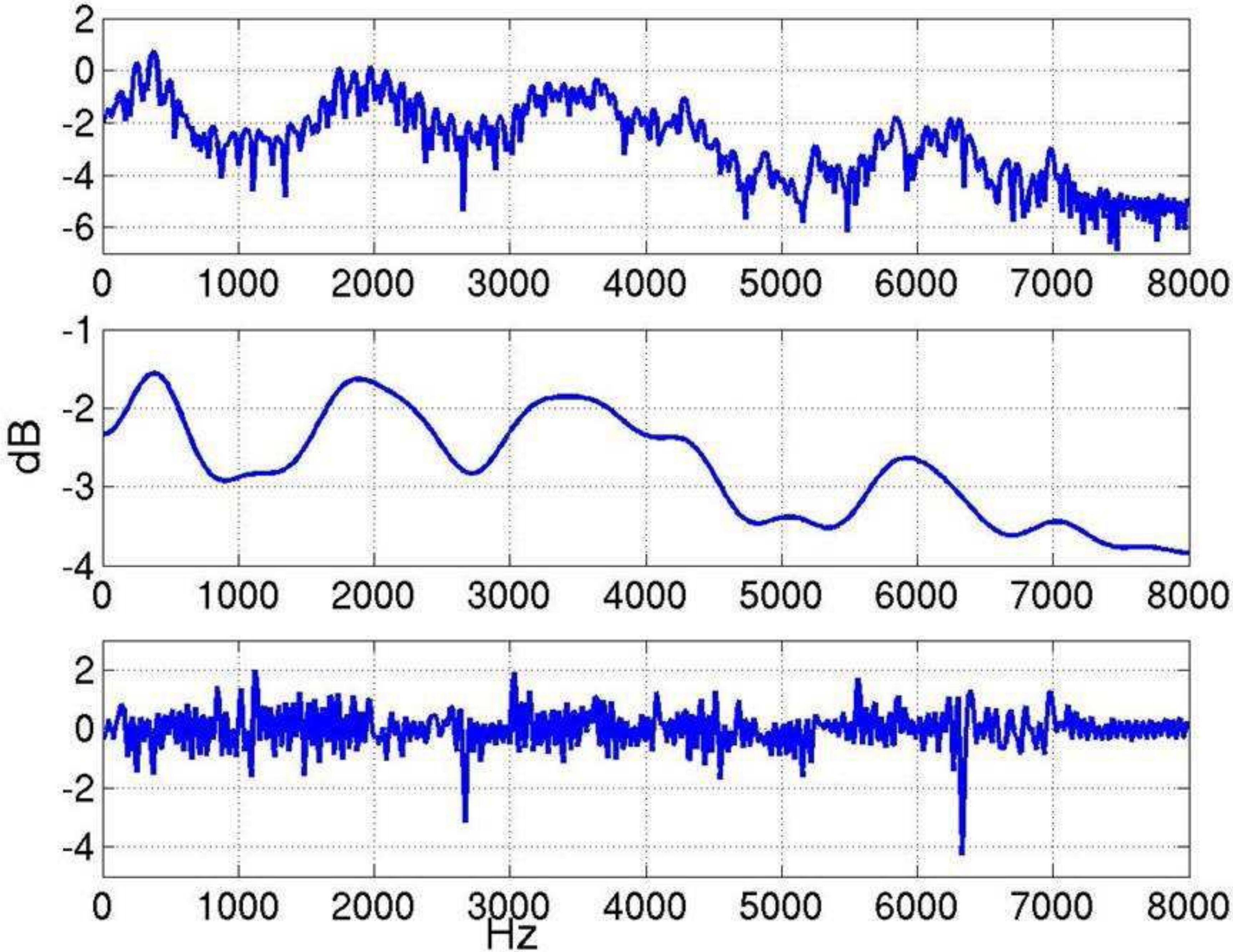




$$s[n] = h[n] * e[n]$$

DFT of s[n] $$S[k] = H[k]E[k]$$

$$\log(|S[k]|) = \log(|H[k]|) + \log(|E[k]|)$$



2
0
-2
-4
-6
0
1000
2000
3000
4000
5000
6000
7000
8000
-1
-2
-3
-4
dB
0
-2
-4
Hz

Cepstral analysis

- **Homomorphic speech processing**
  - Speech is modelled as the output of a linear, time varying system (linear time-invariant (LTI) in short seg.) excited by either quasi-periodic pulses or random noise.
  - The problem of speech analysis is to estimate the parameters of the speech model and to measure their variations with time.
  - Since the excitation and impulse response of a LTI system are combined in a convolutional manner, the problem of speech analysis can also been viewed as a problem in separating the components of a convolution, called ”deconvolution”.

$$y[n] = x[n] * h[n]$$

The principle of superposition for conventional linear systems:

$$
\begin{cases}
L[x(n)] = L[x_1(n) + x_2(n)] = L[x_1(n)] + L[x_2(n)] \\
\qquad = y_1(n) + y_2(n) = y(n) \\
L[ax(n)] = aL[x(n)] = ay(n)
\end{cases}
$$

If signals fall in non-overlapping frequency bands then they are separable

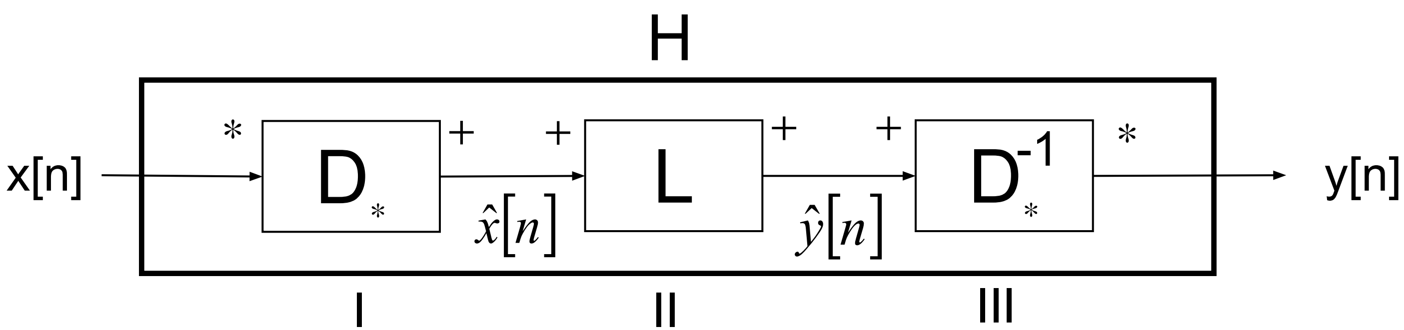
$x[n]=x_1[n]+x_2[n]$

$X_1(\omega)=\mathscr{F}\{x_1[n]\}$ & $X_1(\omega)$ $[0, \pi/2]$,

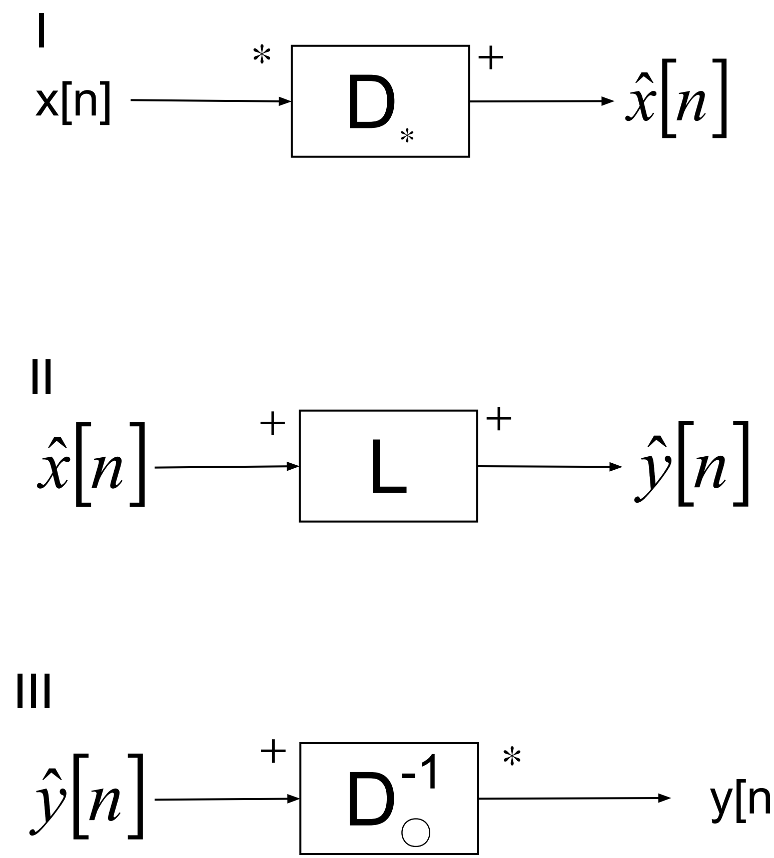
$X_2(\omega)=\mathscr{F}\{x_2[n]\}$ & $X_2(\omega)$ $[\pi/2, \pi]$,

Principles of Homomorphic Processing

- Importance of homomorphic systems for speech processing lies in their capability of transforming nonlinearly combined signals to additively combined signals so that linear filtering can be performed on them.
- Homomorphic systems can be expressed as a cascade of three homomorphic sub-systems □ referred to as the ***canonic representation***:



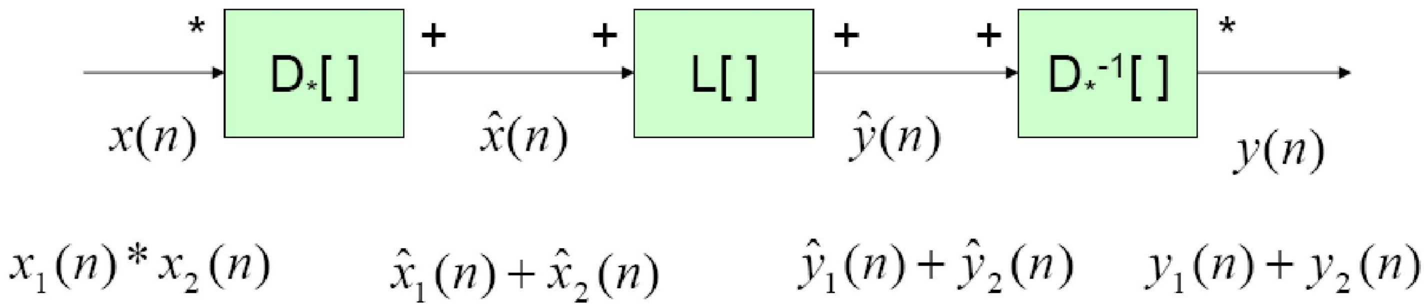
Canonic Representation of a Homomorphic System



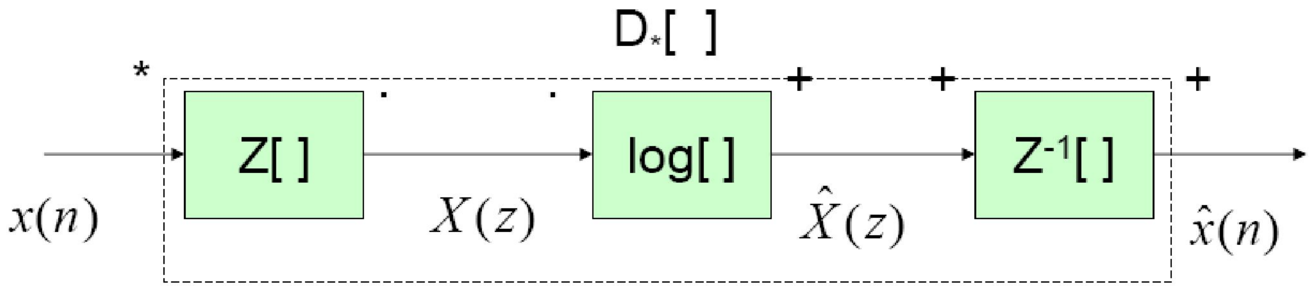
I. System takes inputs combined by convolution and transforms them into additive outputs
II. System is a conventional linear system
III. Inverse of first system--takes additive inputs and transforms them into convolution outputs

Cepstral analysis

❑ Canonic form for system for homomorphic deconvolution



❑ The characteristic system for homomorphic deconvolution



Cepstral analysis

Observation:

$$x[n] = x_1[n] * x_2[n] \Leftrightarrow X(z) = X_1(z)X_2(z)$$

taking logarithm of X(z), then

$$\log\{X(z)\} = \log\{X_1(z)\} + \log\{X_2(z)\}$$

$$\text{i.e., } \hat{X}(z) = \hat{X}_1(z) + \hat{X}_2(z)$$

$$\hat{x}[n] = \hat{x}_1[n] + \hat{x}_2[n]$$ in the cepstral domain

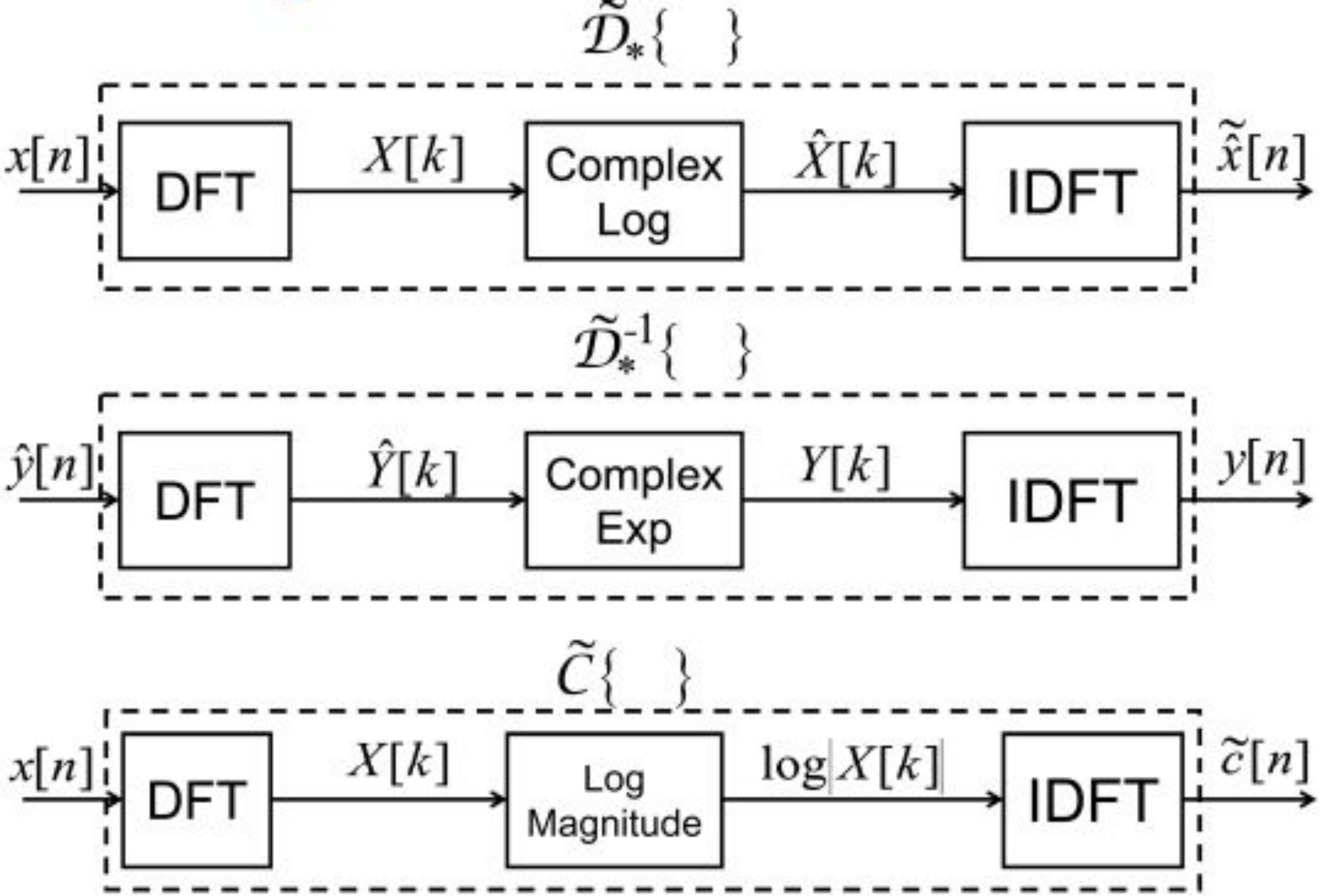
- So, the two convolved signals are additive in the cepstral domain

Cepstral analysis

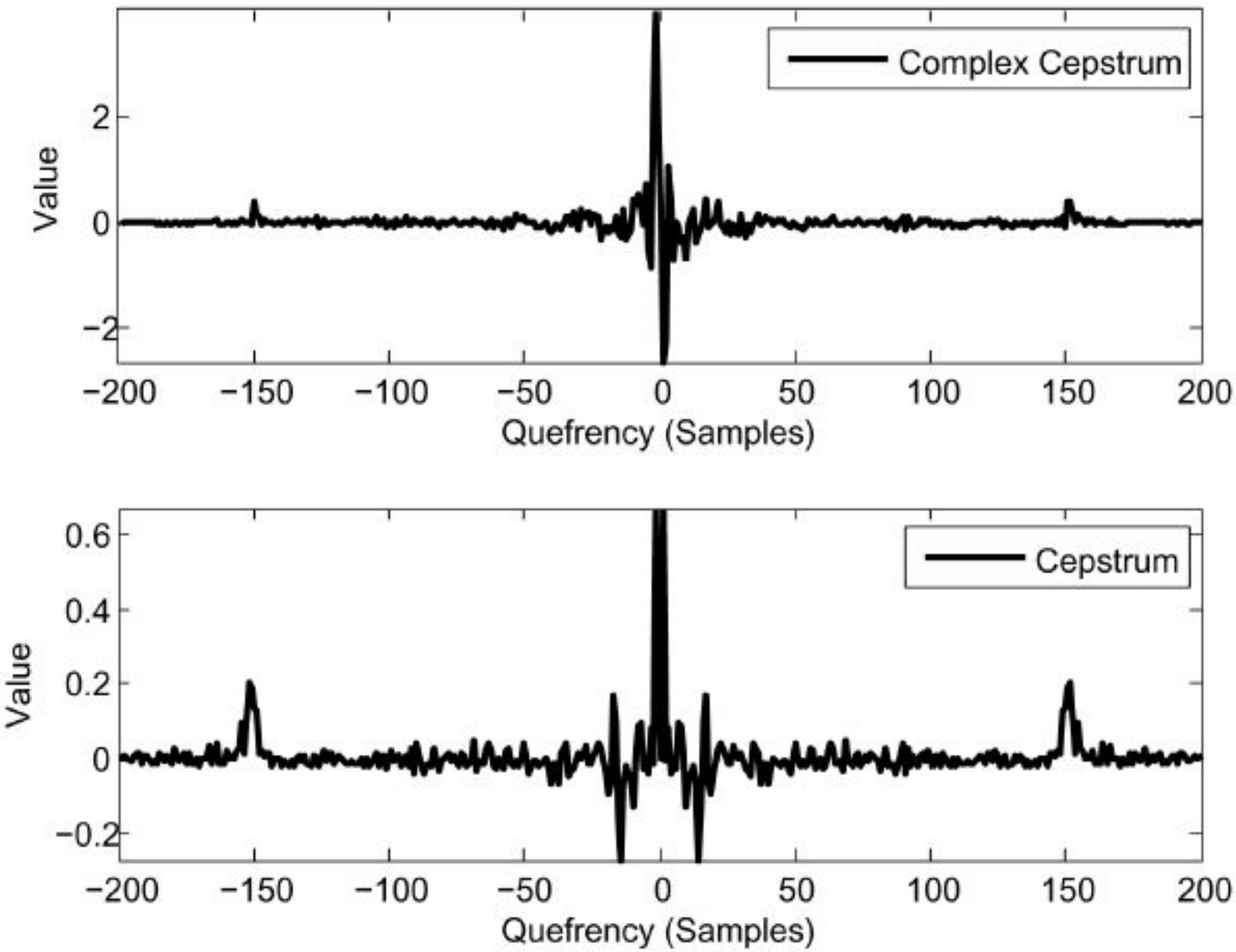
Real cepstrum *c[n]* is the even part of $\hat{x}[n]$

$$\begin{cases} \hat{x}[n] = \dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi} \hat{X}(e^{jw}) e^{jwn} dw \\ \qquad = \dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi} \log\{X(e^{jw})\} e^{jwn} dw & \text{complex cepstrum} \\ c[n] = \dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi} \log|X(e^{jw})| e^{jwn} dw & \text{cepstrum} \end{cases}$$

Computational Considerations



Voiced Speech Example



Basic Speech processing steps for Frequency Parameter

Signal: **S(t) = x[n]**

Pre-emphasis

Framing

Basic Signal Processing Block

Windowing

DFT

Power Spectrum

y[n] = x[n] – α * x[n-1] *(where α = 0.95)*

x(n) = x(n)*W(n)

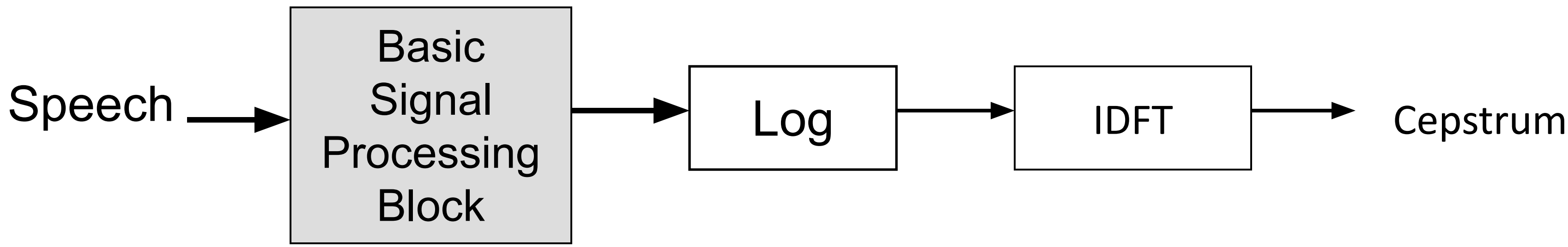
Hamming : 0.54 – 0.46 cos(2*pi*n/(N-1))

Hanning : 0.5(1- cos(2*pi*n/(N-1)))

Cosine : sin(pi*n/(N-1))

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0, \dots, N-1$$

Cepstral Transform Coefficients (CC)



$$\text{Cepstrum} = \text{IDFT}(\log(\text{DFT}(S(n))))$$

LPC Cepstrum

The LPC vector is defined by [a_0,a_1,a_2,...a_p] and the CC vector is defined by [$c_0 c_1 c_2 ... c_p ... c_{n-1}$

LPC Cepstrum (c_m)	
$$\begin{aligned} c_0 &= \log G^2 \\ c_m &= a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \leq m \leq p \\ c_m &= \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad m > p \end{aligned}$$	$$\begin{aligned} G &= e^{c_0/2} \\ a_m &= c_m - \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \leq m \leq p \end{aligned}$$

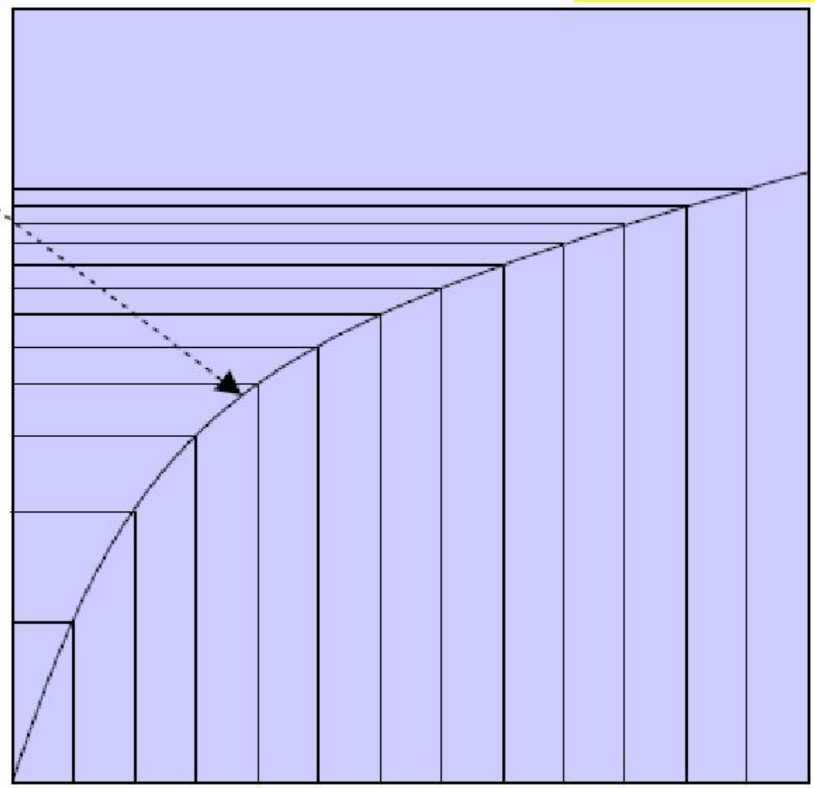
Warping frequency warping

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

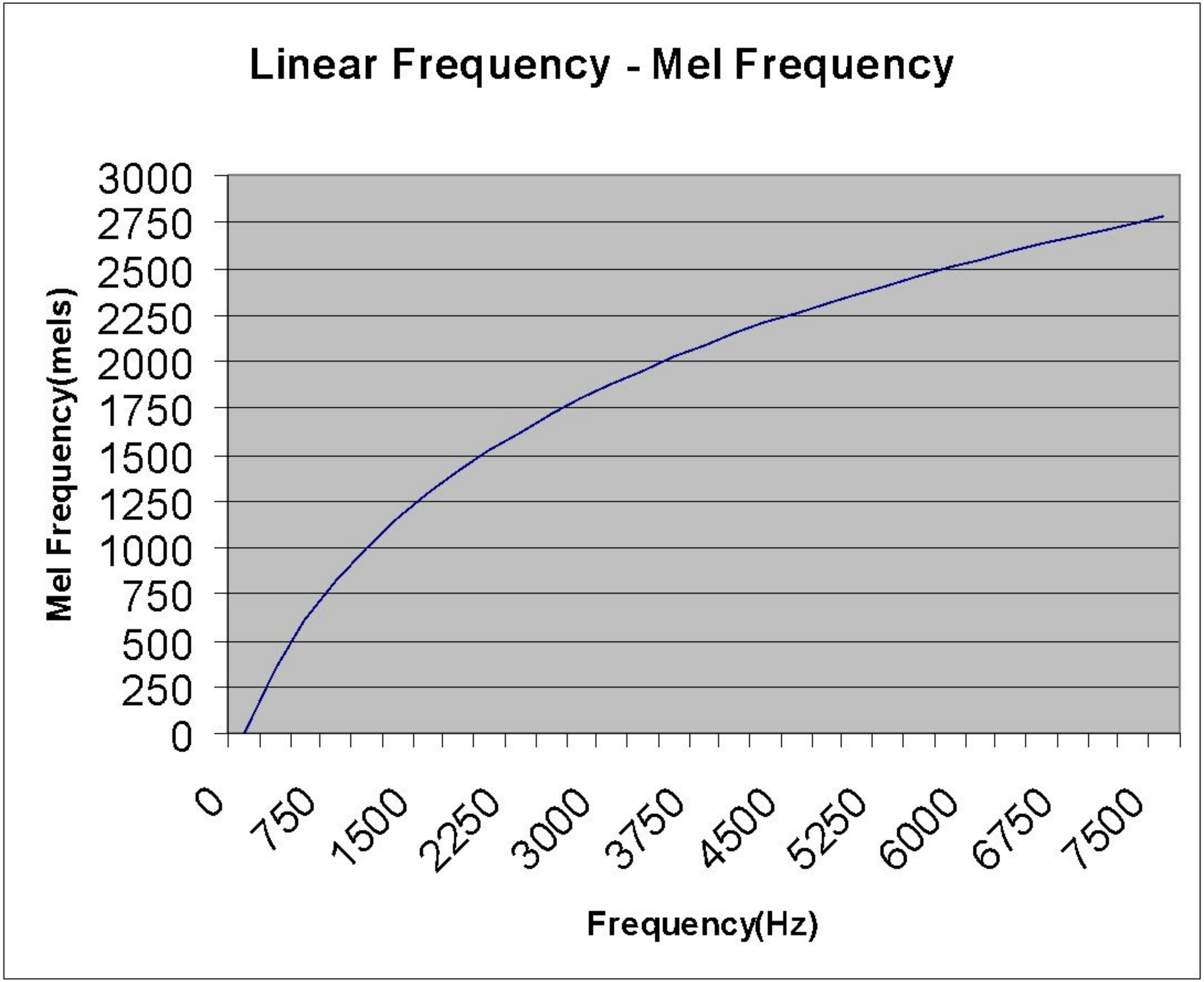
A standard warping function is the Mel warping function

Warping function (based on studies of human hearing)

Warped frequency axis: unequal increments of frequency at equal intervals or conversely, equal increments of frequency at unequal intervals

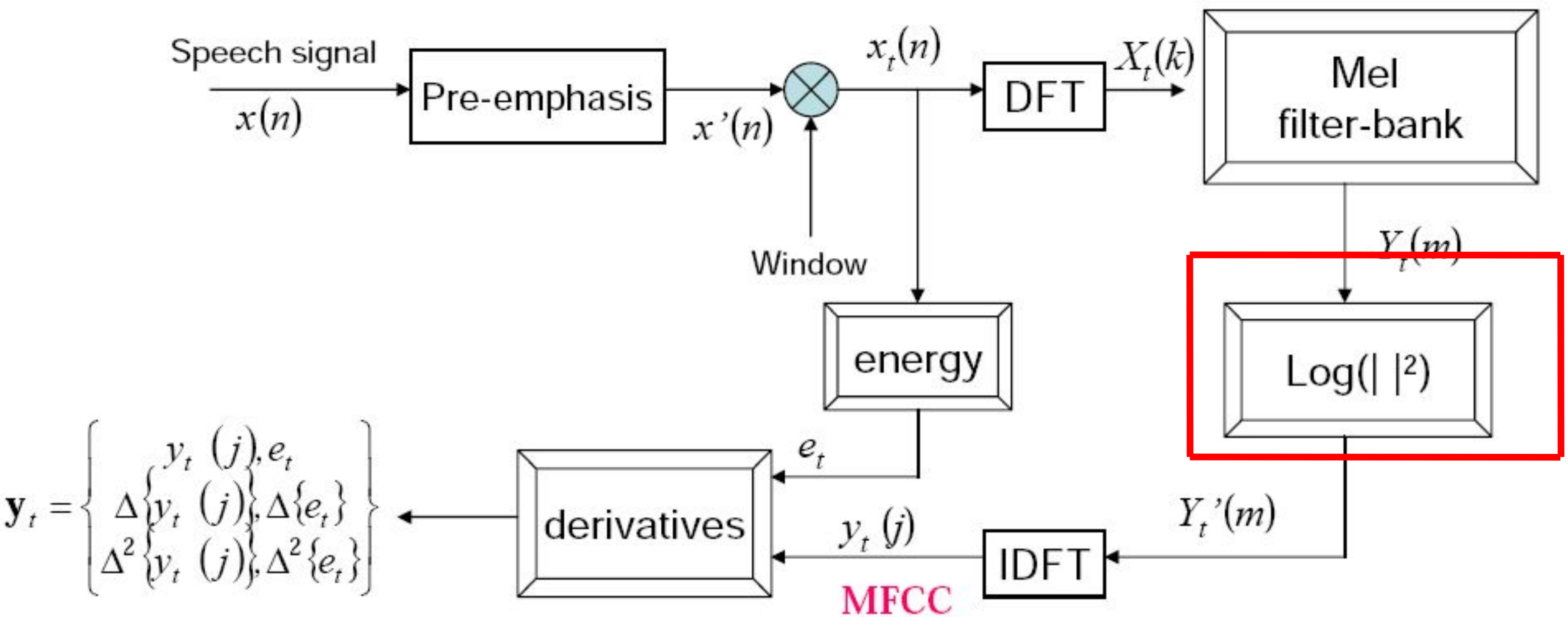


Linear frequency axis: Sampled at uniform intervals by an FFT



Linear Frequency - Mel Frequency
Mel Frequency(mels)
3000
2750
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
0
750
1500
2250
3000
3750
4500
5250
6000
6750
7500
Frequency(Hz)

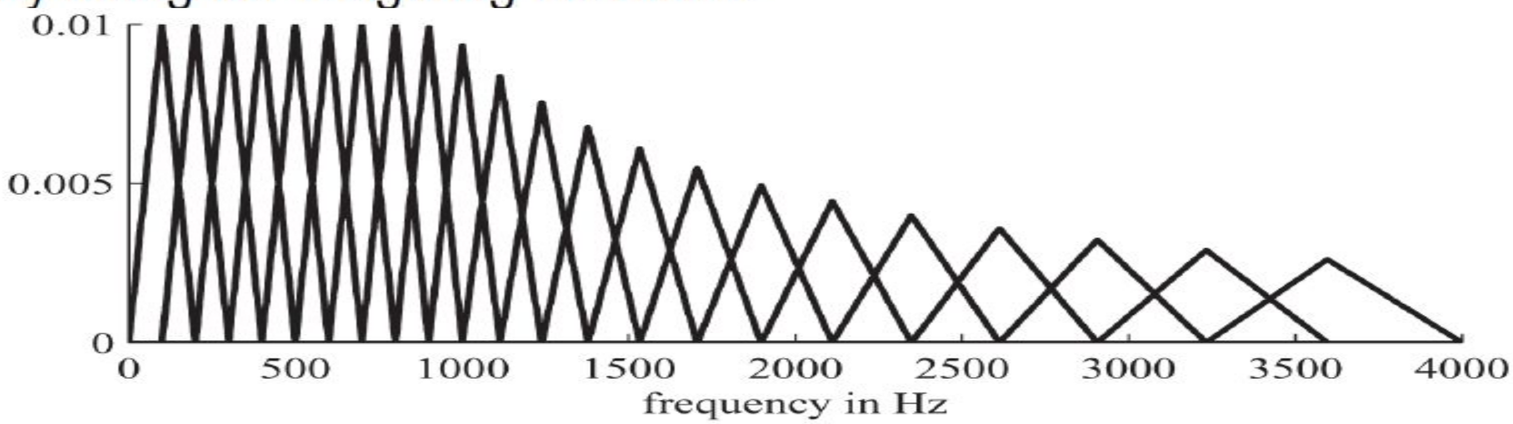
MFCC



Speech signal
$x(n)$
Pre-emphasis
$x'(n)$
Window
$x_t(n)$
DFT
$X_t(k)$
Mel filter-bank
$Y_t(m)$
Log(| |²)
$Y_t'(m)$
IDFT
$y_t(j)$
MFCC
energy
e_t
derivatives
$\mathbf{y}_t = \begin{Bmatrix} y_t(j), e_t \\ \Delta\{y_t(j)\}, \Delta\{e_t\} \\ \Delta^2\{y_t(j)\}, \Delta^2\{e_t\} \end{Bmatrix}$

Mel Frequency Cepstral Coefficients

- Basic idea is to compute a frequency analysis based on a filter bank with approximately critical band spacing of the filters and bandwidths. For 4 kHz bandwidth, approximately 20 filters are used.
- First perform a short-time Fourier analysis, giving $X_m[k], k = 0,1,...,NF/2$ where m is the frame number and k is the frequency index (1 to half the size of the FFT)
- Next the DFT values are grouped together in critical bands and weighted by triangular weighting functions.



Mel Frequency Cepstral Coefficients

- The mel-spectrum of the m^{th} frame for the r^{th} filter $(r = 1, 2, \ldots, R)$ is defined as:

$$\text{MF}_m[r] = \frac{1}{A_r} \sum_{k=L_r}^{U_r} |V_r[k] X_m[k]|^2$$

where $V_r[k]$ is the weighting function for the r^{th} filter, ranging from DFT index L_r to U_r, and

$$A_r = \sum_{k=L_r}^{U_r} |V_r[k]|^2$$

is the normalizing factor for the r^{th} mel-filter. (Normalization guarantees that if the input spectrum is flat, the mel-spectrum is flat).

- A discrete cosine transform of the log magnitude of the filter outputs is computed to form the function mfcc$[n]$ as:

$$\text{mfcc}_m[n] = \frac{1}{R} \sum_{r=1}^{R} \log(\text{MF}_m[r]) \cos\left[\frac{2\pi}{R}\left(r + \frac{1}{2}\right)n\right], \quad n = 1, 2, \ldots, N_{\text{mfcc}}$$

- Typically $N_{\text{mfcc}} = 13$ and $R = 24$ for 4 kHz bandwidth speech signals.

Delta Cepstrum

- The set of mel frequency cepstral coefficients provide perceptually meaningful and smooth estimates of speech spectra, over time
- Since speech is inherently a dynamic signal, it is reasonable to seek a representation that includes some aspect of the dynamic nature of the time derivatives (both first and second order derivatives) of the short-term cepstrum
- The resulting parameter sets are called the delta cepstrum (first derivative) and the delta-delta cepstrum (second derivative).
- The simplest method of computing delta cepstrum parameters is a first difference of cepstral vectors, of the form:

$$\Delta \text{mfcc}_m[n] = \text{mfcc}_m[n] - \text{mfcc}_{m-1}[n]$$

- The simple difference is a poor approximation to the first derivative and is not generally used. Instead a least-squares approximation to the local slope (over a region around the current sample) is used, and is of the form:

$$d_i = \frac{\sum_{n=1}^{N} n(c_{n+i} - c_{n-i})}{2\sum_{n=1}^{N} n^2}$$

Perceptual Linear Prediction

- PLP parameters are the coefficients that result from standard all-pole modeling or linear predictive analysis, of a specially modified, short-term speech spectrum.
- In PLP the speech spectrum is modified by a set of transformations that are based on models of the human auditory system
- The spectral resolution of human hearing is roughly linear up to 800 or 1000Hz, but it decreases with increasing frequency above this linear range

Perceptually motivated analyses

- **Critical-band spectral resolution**: PLP incorporates critical-band spectral-resolution into its spectrum estimate by remapping the frequency axis to the Bark scale and integrating the energy in the critical bands to produce a critical-band spectrum approximation.
- **Equal-loudness pre-emphasis**: At conversational speech levels, human hearing is more sensitive to the middle frequency range of the audible spectrum. PLP incorporates the effect of this phenomenon by multiplying the critical-band spectrum by an equal loudness curve that suppresses both the low- and high-frequency regions relative to the midrange from 400 to 1200 Hz.
- **Intensity-loudness power law:** There is a nonlinear relationship between the intensity of sound and the perceived loudness. PLP approximates the power-law of hearing by using a cube-root amplitude compression of the loudness-equalized critical band spectrum estimate.

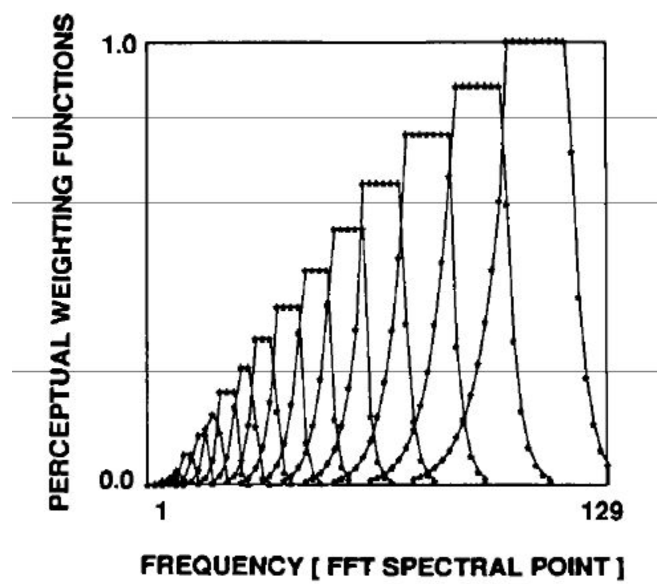
Perceptual LPC

(Hermansky, *J. Acoust. Soc. Am.*, 1990)

- First, warp the spectrum to a Bark scale:

$$\tilde{S}(b) = \sum_{k=0}^{N-1} |H_b(k)|^2 |X(k)|^2, \qquad b = 1, \ldots, K$$

- The filters, $H_b(k)$, are uniformly spaced in Bark frequency. Their amplitudes are scaled by the equal-loudness contour (an estimate of how loud each frequency sounds):



Perceptual LPC

- Second, compute the cube-root of the power spectrum
  - Cube root replaces the logarithm that would be used in MFCC
  - Loudness of a tone is proportional to cube root of its power

$$Y(b) = S(b)^{0.33}$$

- Third, inverse Fourier transform to find the "Perceptual Autocorrelation:"

$$\begin{aligned}\tilde{R}(m) &= \frac{1}{2K}\sum_{b=0}^{2K} Y(b)e^{\frac{j2\pi bm}{2K}} \\ &= \frac{1}{K}\sum_{b=1}^{K} Y(b)\cos\left(\frac{\pi bm}{K}\right) + \frac{(-1)^m}{2K}Y(K)\end{aligned}$$

Perceptual LPC

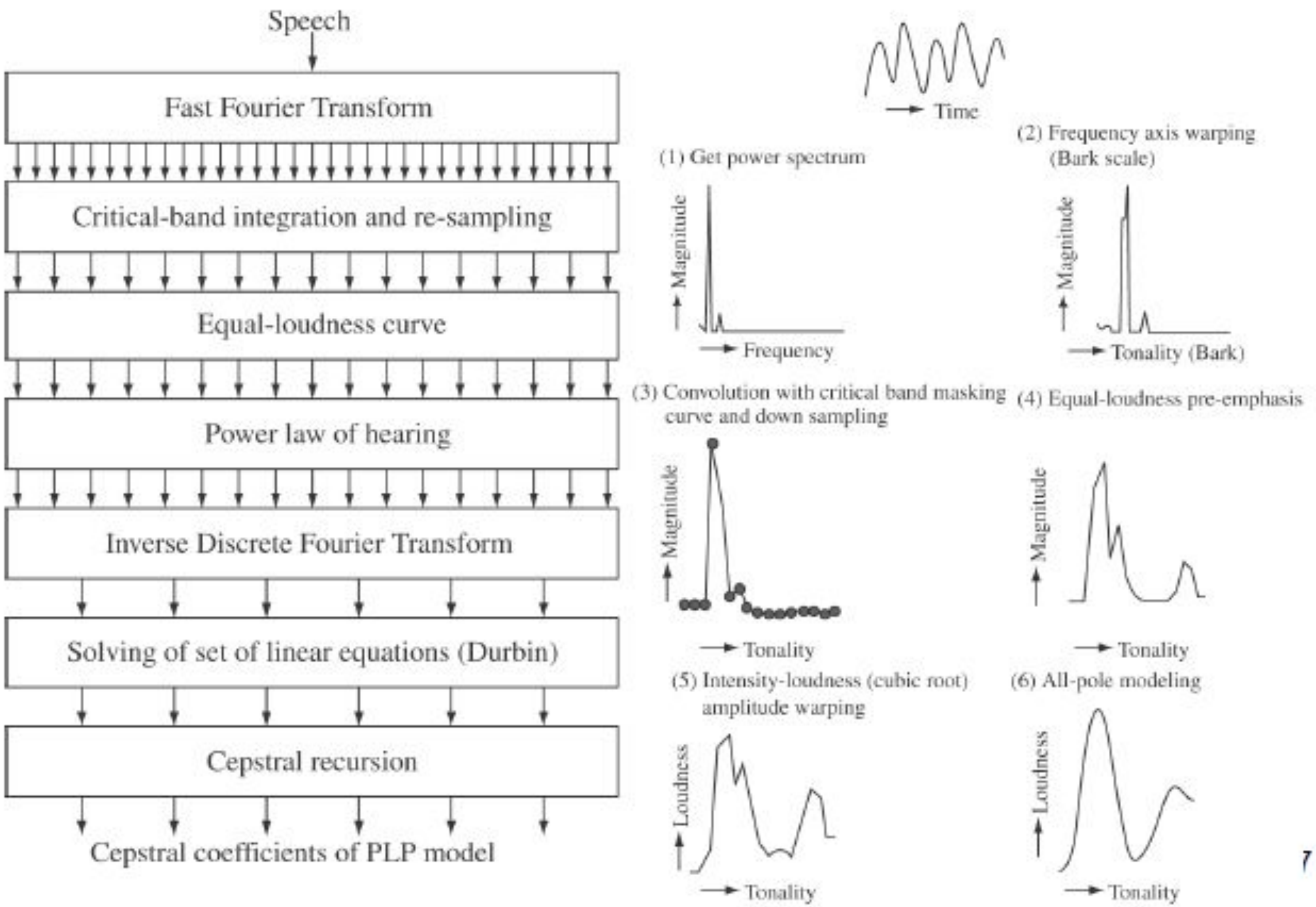
- Fourth, use Normal Equations to find the Perceptual LPC (PLP) coefficients:

$$\tilde{R}(m) = \sum_{k=1}^{p} \tilde{a}_k \tilde{R}(|m-k|)$$

- Fifth, use the LPC Cepstral recursion to find Perceptual LPC Cepstrum (PLPCC):

$$\tilde{c}(m) = \tilde{a}_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) \tilde{c}(k)\tilde{a}_{m-k}, \quad 1 \leq m \leq p$$

Perceptual Linear Prediction



RASTA(RelAtive SpecTrA)

- The rate of change of nonlinguistic components of speech and background noise environments often lies outside the typical rate-of-change of vocal-tract shapes in conversational speech
- Hearing is relatively insensitive to slowly varying stimuli
- The basic idea of RASTA filtering is to exploit these phenomena by suppressing constant and slowly varying elements in each spectral component of the short term auditory-like spectrum prior to computation of the linear prediction coefficients

RASTA (RelAtive SpecTral Amplitude)

(Hermansky, *IEEE Trans. Speech and Audio Proc.*, 1994)

- Modulation-filtering of the cepstrum is equivalent to modulation-filtering of the log spectrum:

  $c_t^*[m] = \Sigma_k h_k c_{t-k}[m]$

- RASTA is a particular kind of modulation filter:

$$H(z) = \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{10z^{-2}(1 - 0.98z^{-1})}$$