COM4511/COM6511 - Speech Technology

L2
Front-ends for speech technology

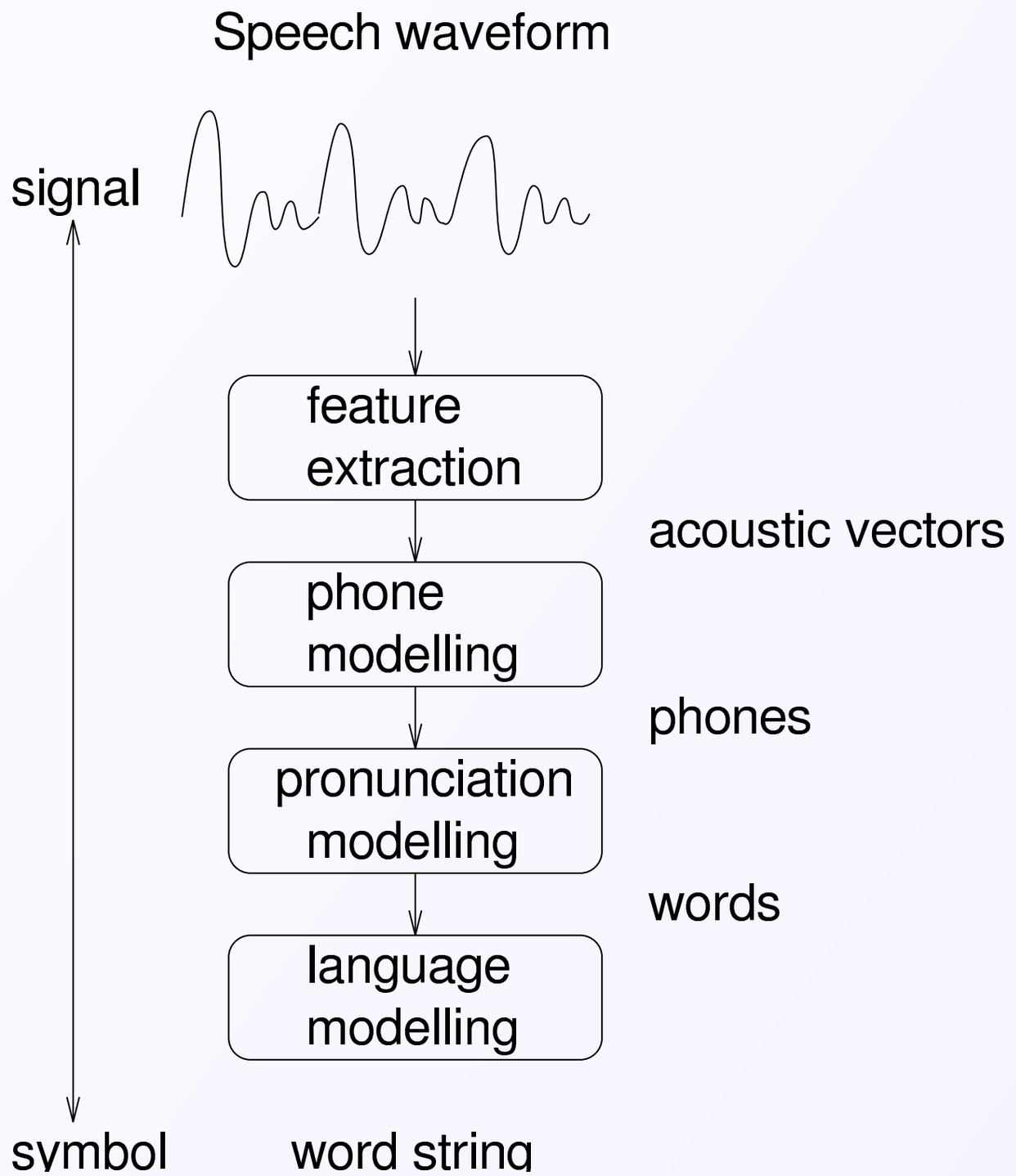
Thomas Hain
t.hain@sheffield.ac.uk
Spring Semester

This lecture

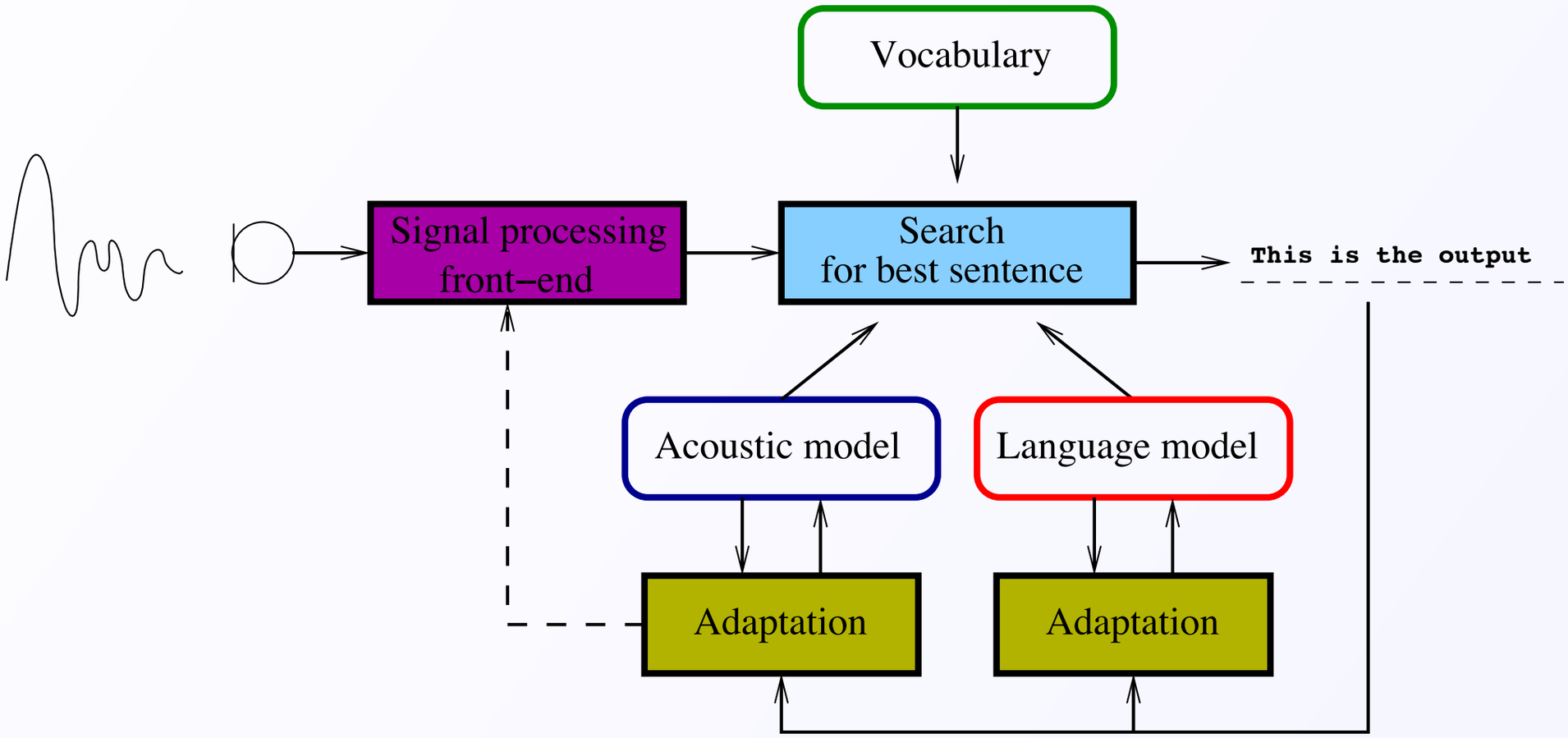
- Extraction features from speech signals
  - Content
  - Prosody

Levels of representation

Speech processing is mot often concerned with transformation of representations at different levels.



Generic Recognition Architecture

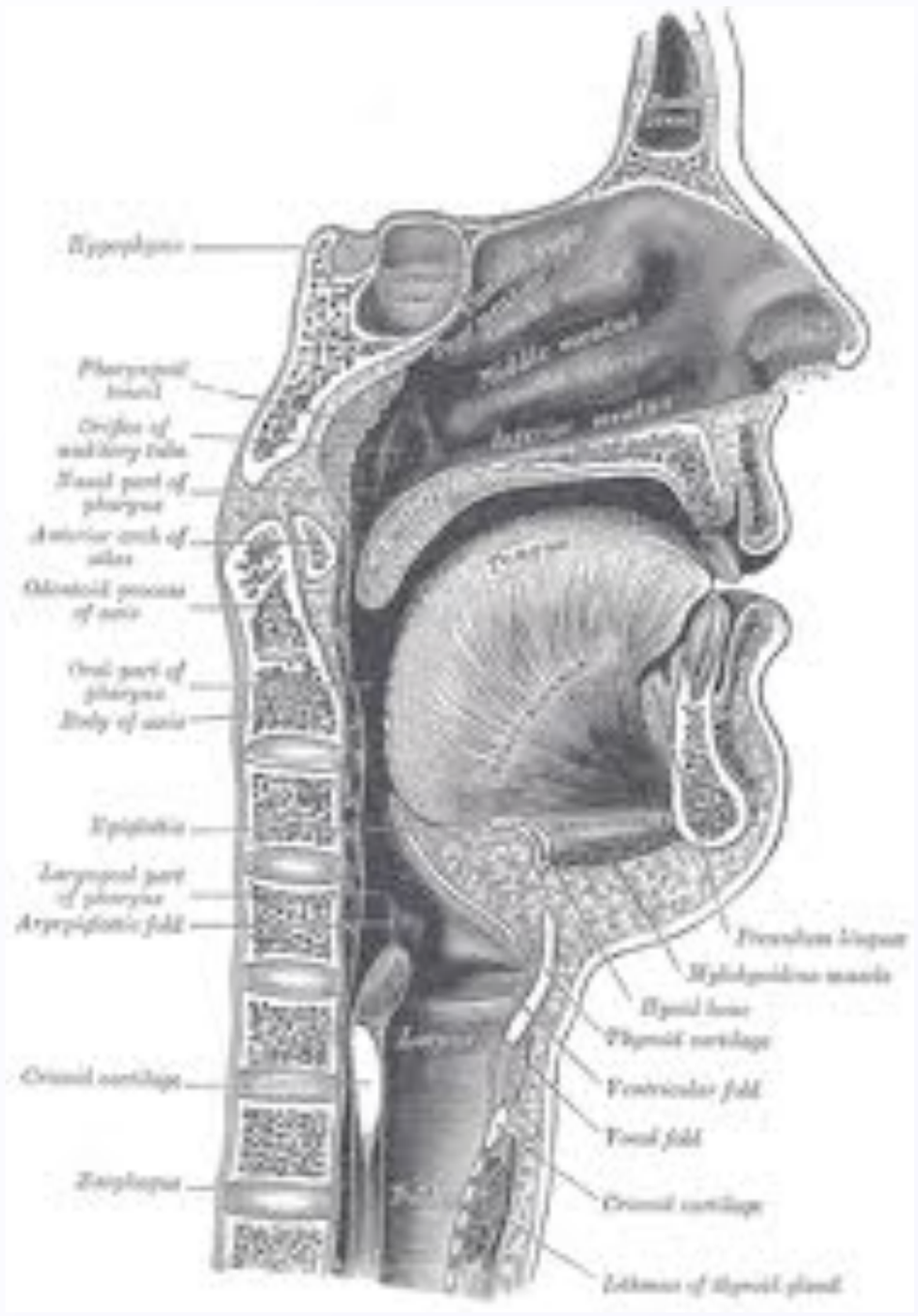


A search is made for the most likely word or sentence given the acoustic and language models (recognition,decoding). A finite set of words is defined in the vocabulary of the ASR system.

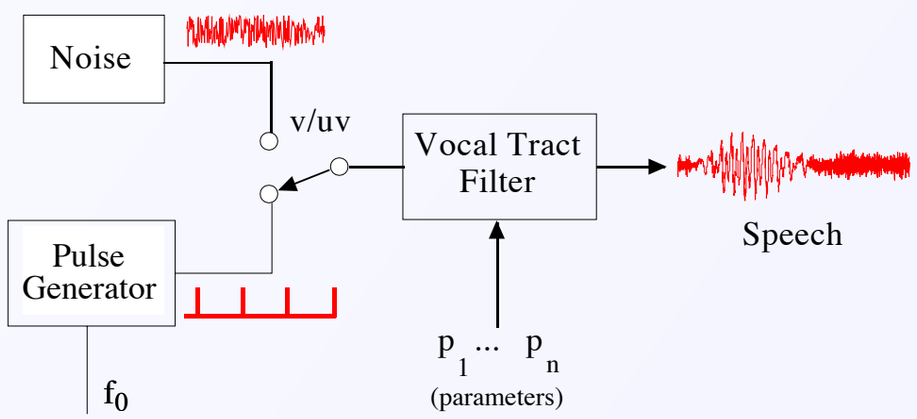
Speech signals are special

- Speech
  - …. is for communication between people
  - … has evolved to be robust
  - …. is time ordered information transfer
  - …. carries more than just the words

Speech - Sources



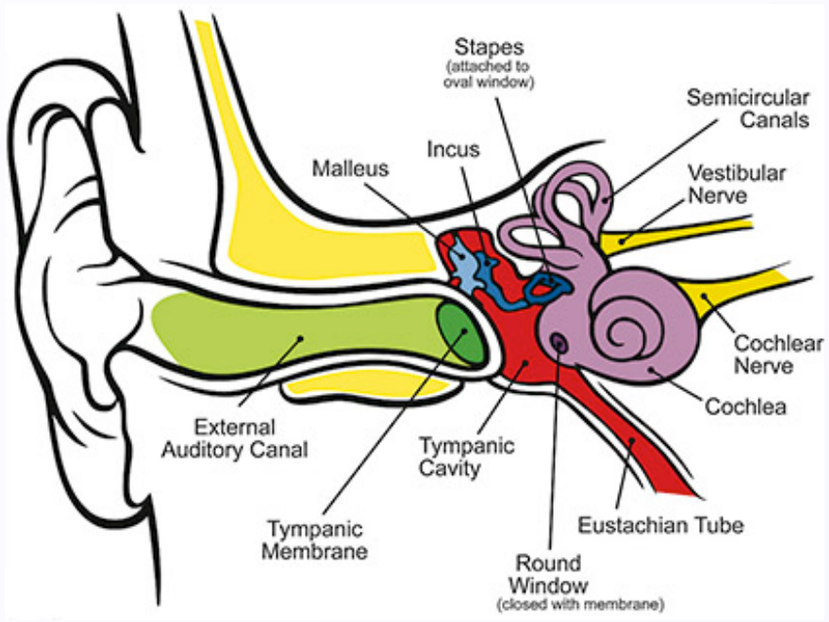
- Production
  - Concept
  - Words
  - Phonemes
  - Vocal tract



Speech - Receivers

If we want to recognise we need to know what

- basilar membrane
- cochlea



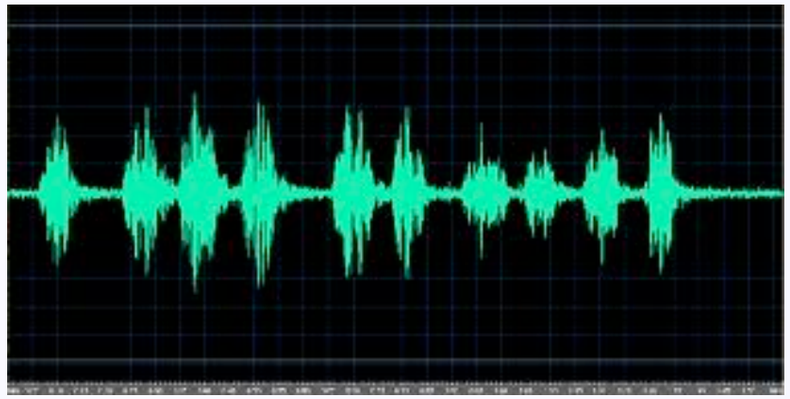
- The human ear is a very sophisticated organ
  - Acts as a non-linear frequency filter
  - Filters out distorting signals - Masking
  - Insensitive to local phase distribution

- The machine listening device

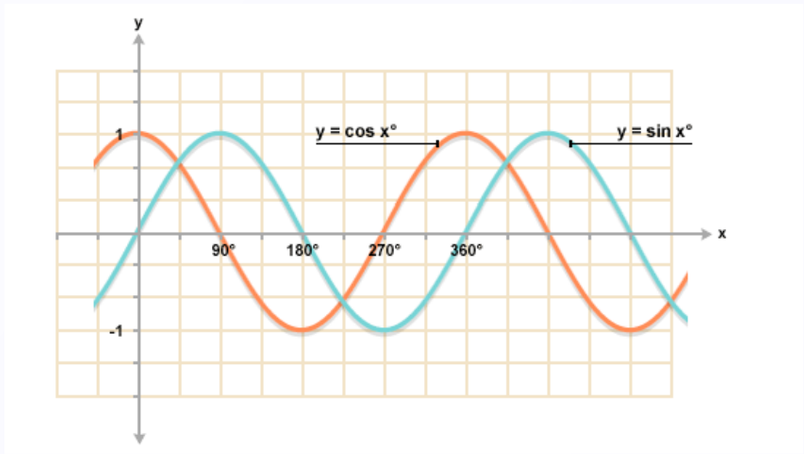
microphone **A/D**

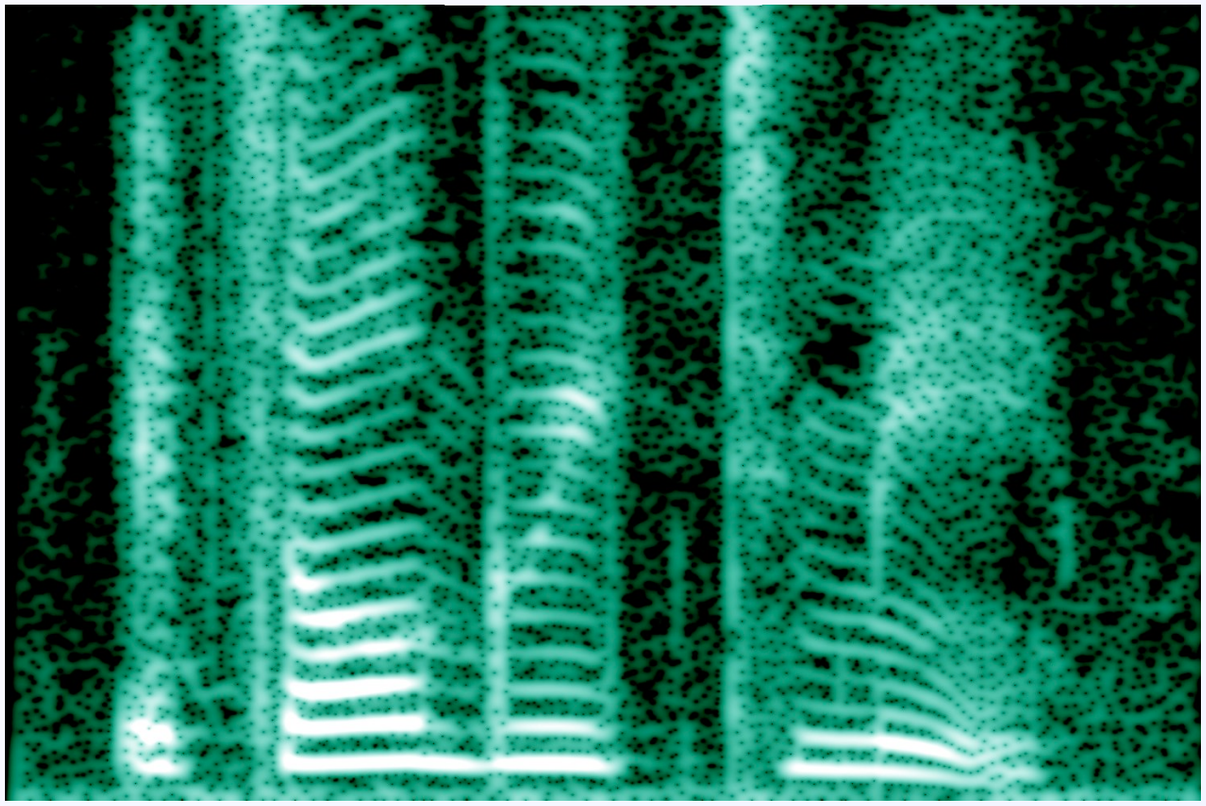
Speech - The signals

- Spectrograms
  - Separate the waveform into distinguishable elements
  -



- Frequencies
  - Strength and ‘phase’





What information is in a speech signal ?

- Content
  - The words …
- Prosody
  - The way we say it
- Identity
  - Who said it
- Physiological traits
  - State and capacity
  - Control

- Other information
  - Where did we say it
  - Others

Content

- Phonetic and word content
- Sequences of sounds
  - reproducible fragments
  - combination of fragments builds words
- We move our articulators to form content
  - Vocal tract
  - Voicing

People can read spectrograms !

Prosody information

Pitch

- perceptual quantity
- closely related to fundamental frequency F0
- The frequency range for F0 is 50-200Hz for adult males, for females 150-350Hz, for children 200-500Hz.

Rhythm

- temporal structure of human speech, many influencing actors ‹ variation can be roughly described on phone level
- stationary sounds (e.g. vowels) vs. transient sounds (e.g. stop consonants)
- position in word
- lengthening of stressed syllables

Intensity

- stress pattern
- mainly determined by phone identity and phone sequence, hence often taken from dictionary (lexical stress)
- considerably less influential than pitch/rhythm

Intonation - Signal parameters

Prosody is mostly controlled by the physical quantities:

1. **Duration** of sub-phones, phones, syllables, whole words. Can be stored on a "unit" basis. Related to rhythm (speech rate) and stress.
2. **Fundamental frequency contour** (F0): it is the longest period of a periodic waveform (only for voiced sounds). It represents the speed at which the vocal chords vibrate.
3. **Energy contour** (speech volume): scaling proportionally to the fundamental frequency contour is commonly sufficient. Local raw energy is defined as the sum over a N-sample window of the squared signal values:

Energy

Energy = Sum of squares of samples in one frame

$$E = \sum_{i=0}^{N-1} s_i^2$$

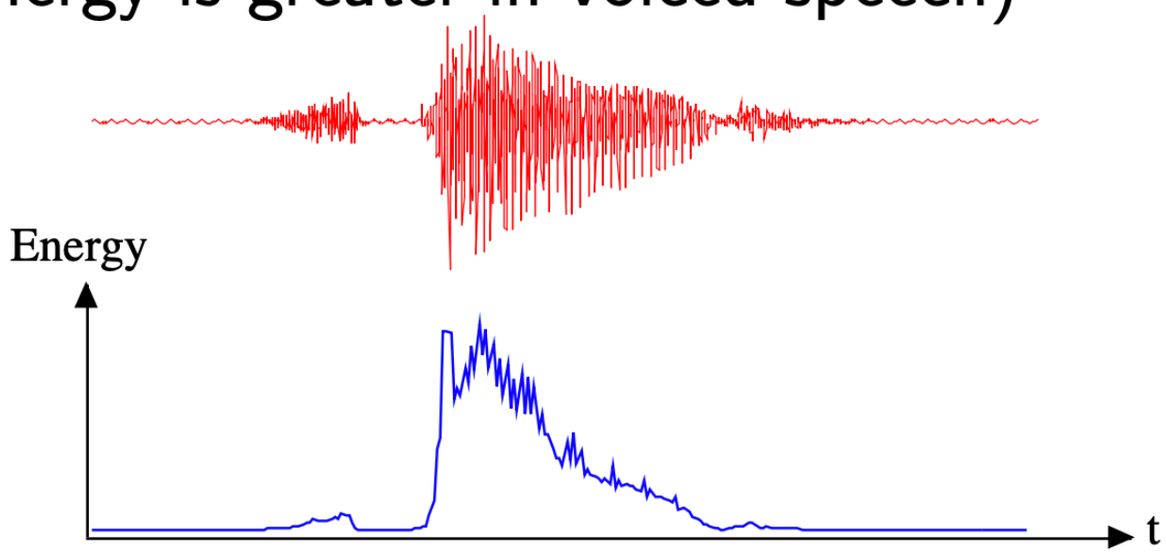
Source Code

```
for i in range(N):
        E = E + s[i] * s[i]

OR

        E = np.sum( np.square(s) )
```

Example: "skills" (energy is greater in voiced speech)



Linear prediction to model source

LP assumes each sample can be predicted from a weighted sum of the p preceding samples. Typically represents 100-200 speech sample frame by 10-15 coefficients (exploits redundancy)

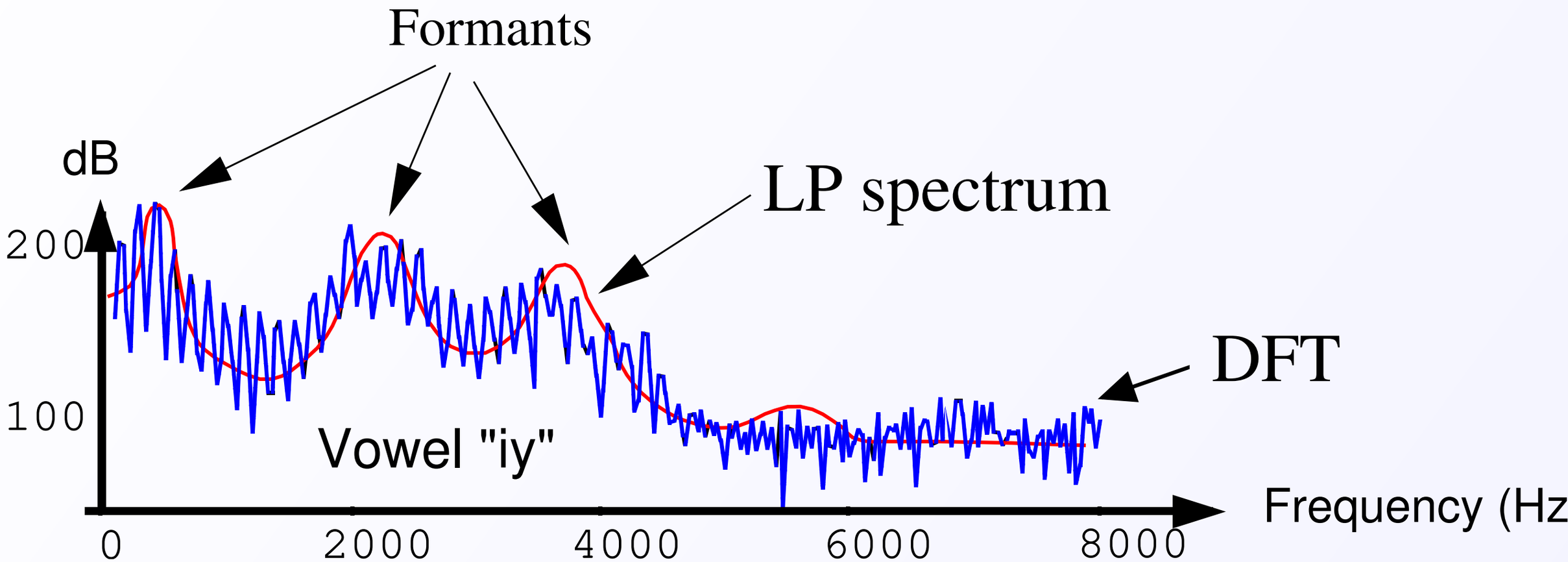
$$\hat{s}[n] \quad = \quad a_1 s[n-1] + a_2 s[n-2] + \ldots + a_p s[n-p] = \sum_{i=1}^{p} a_i s[n-i]$$

Coefficients represent the vocal tract shape and are computed by minimising the prediction error energy. In the z domain we write

$$H(z) \quad = \quad \frac{G}{A(z)}$$

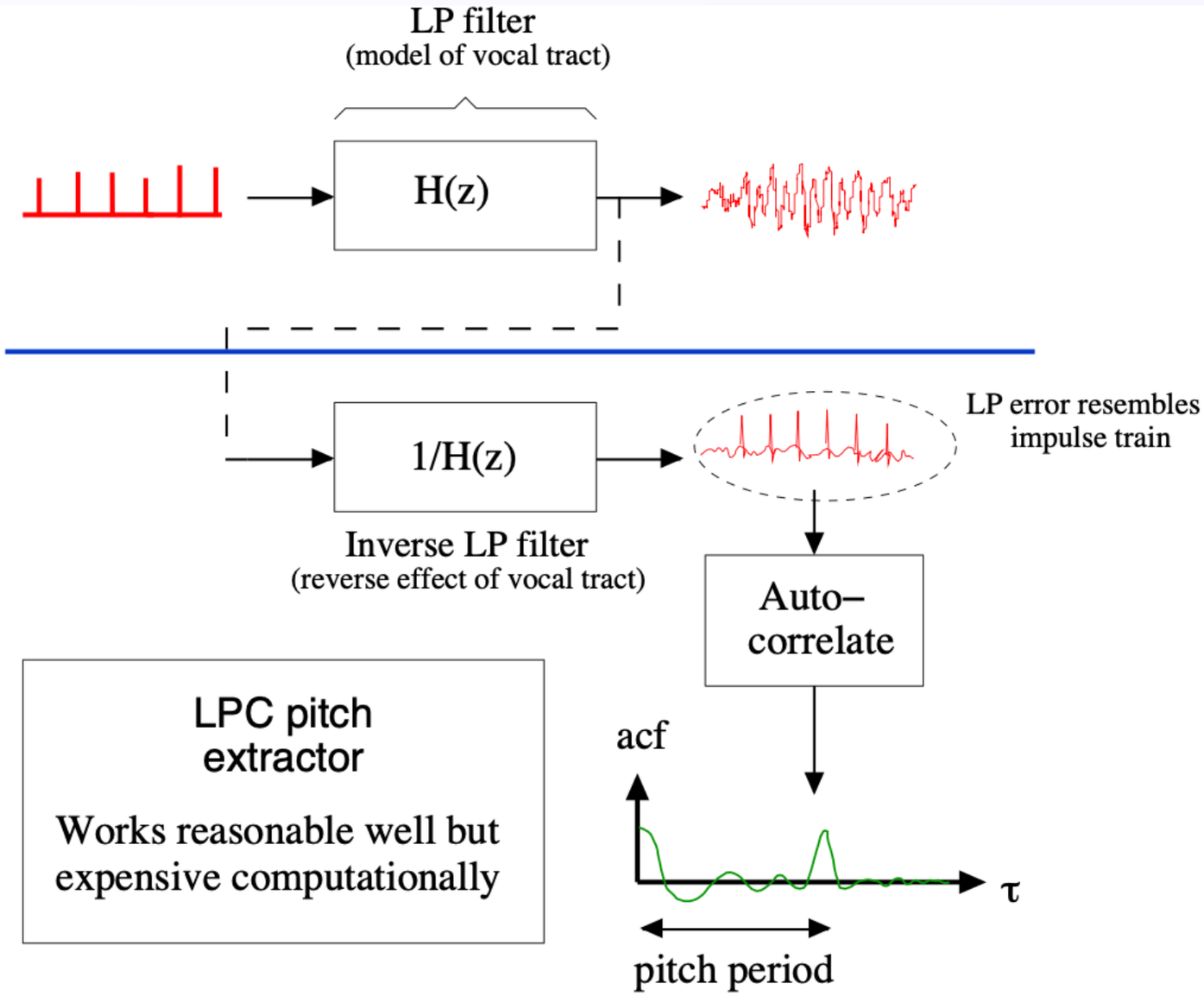
where H(z) is the synthesis filter and A(z) is the analysis filter.

LP Spectrum


Formants
dB
200
100
LP spectrum
DFT
Vowel "iy"
Frequency (Hz
0
2000
4000
6000
8000

Pitch extraction

1. First obtain the LP error signal by passing the speech through the LP inverse filter
2. Find pitch frequency by autocorrelation on error signal (formant info has been removed) and search for peaks
3. Use continuity constraints for pitch tracking



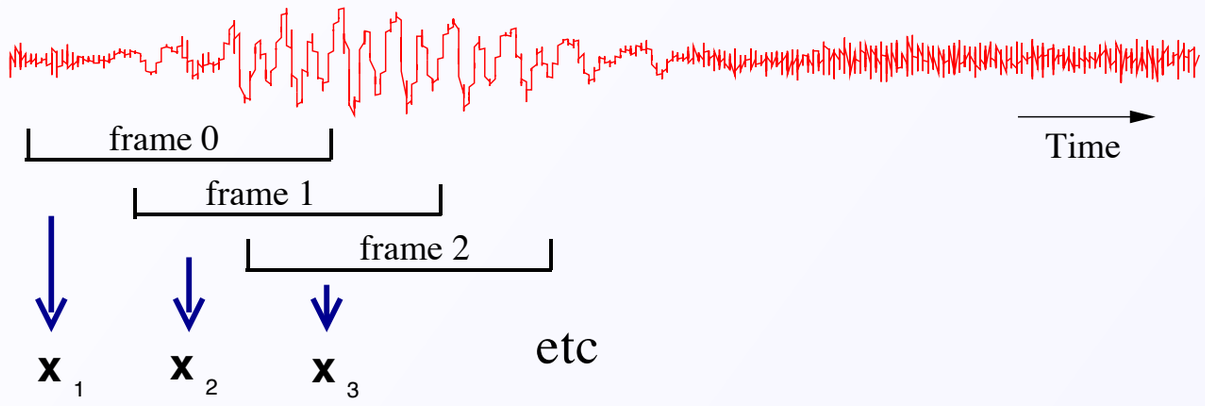
LPC pitch extractor

Works reasonable well but expensive computationally

In practise a reduced bandwidth signal and a low-order LP analysis is used (**SIFT**).

Front-end

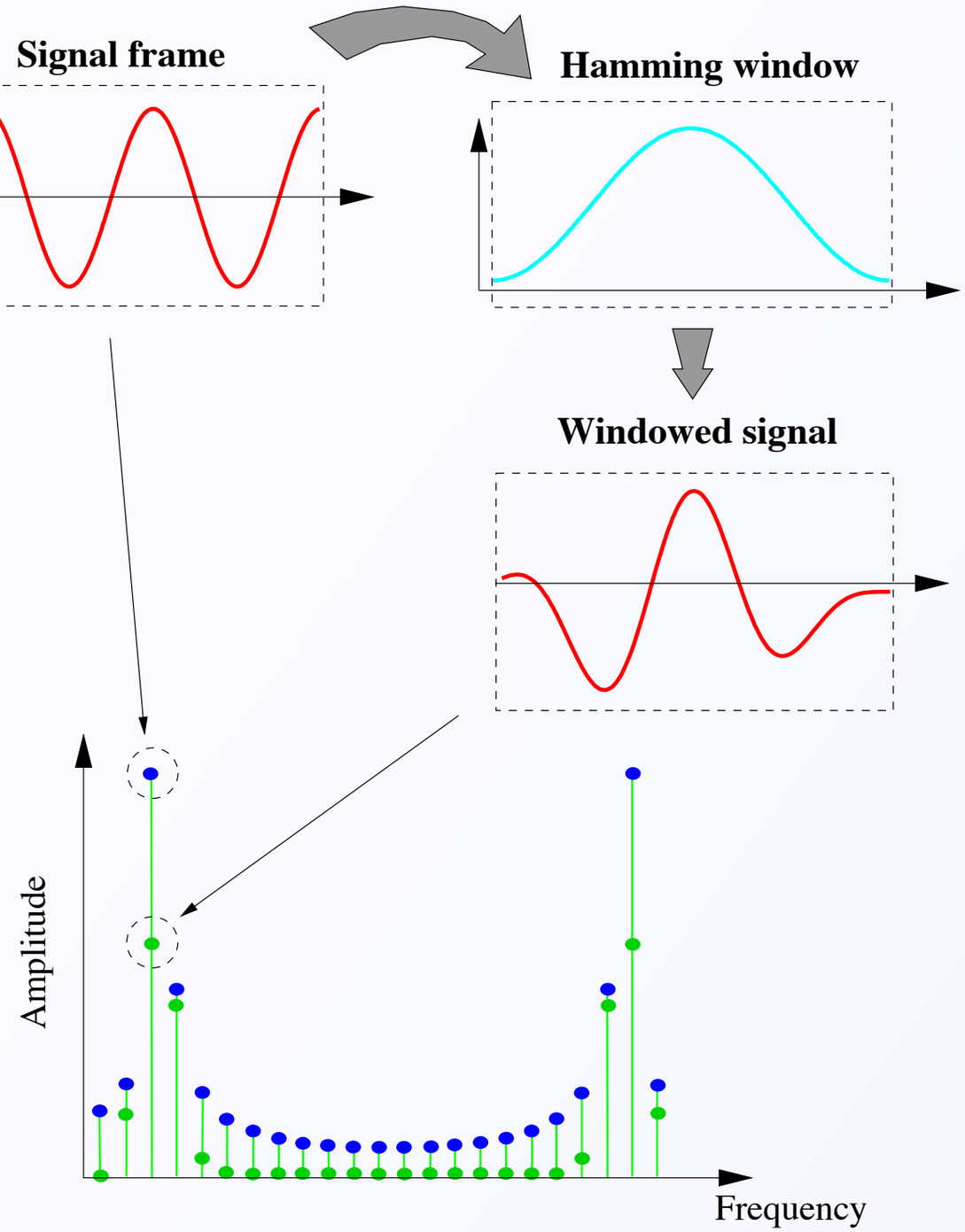
The by far most commonly used front-end is based on Mel-Frequency Cepstral Coefficients (**MFCCs**). Typically every 10ms a speech frame with a length of 25ms is taken.



After application of a **Hamming window**, **Mel-filterbank**, **log conversion** and the Discrete Fourier Transform (**DCT**) we arrive typically at a feature vector $\mathbf{x}_i$ of size 12. As normally the zero-th cepstral coefficient is excluded, a measure of raw energy needs to be added to the feature vector.

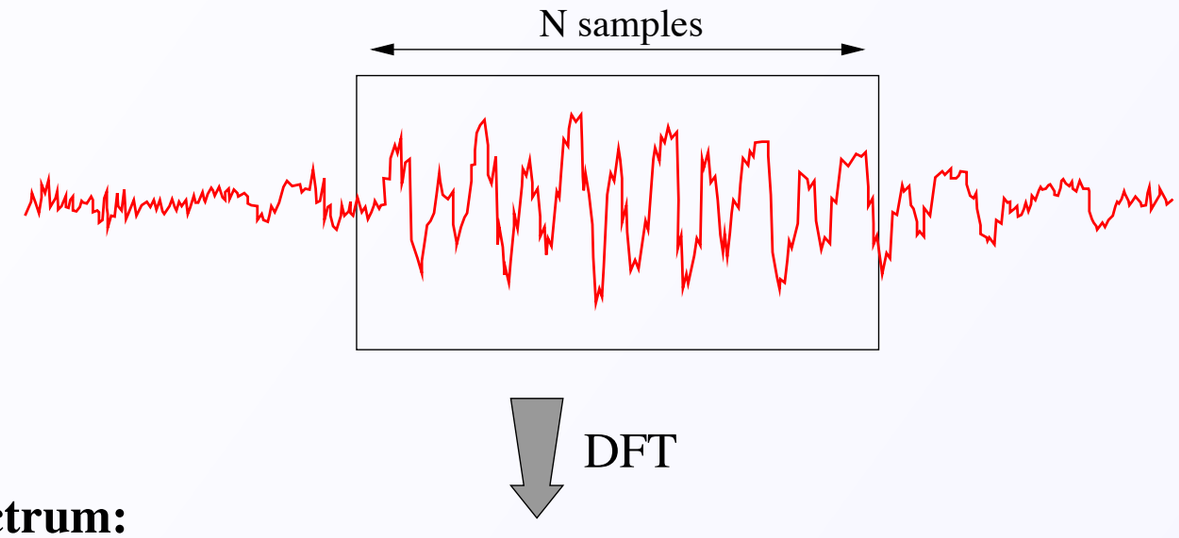
What are the reasons for all these steps ?

Windowing

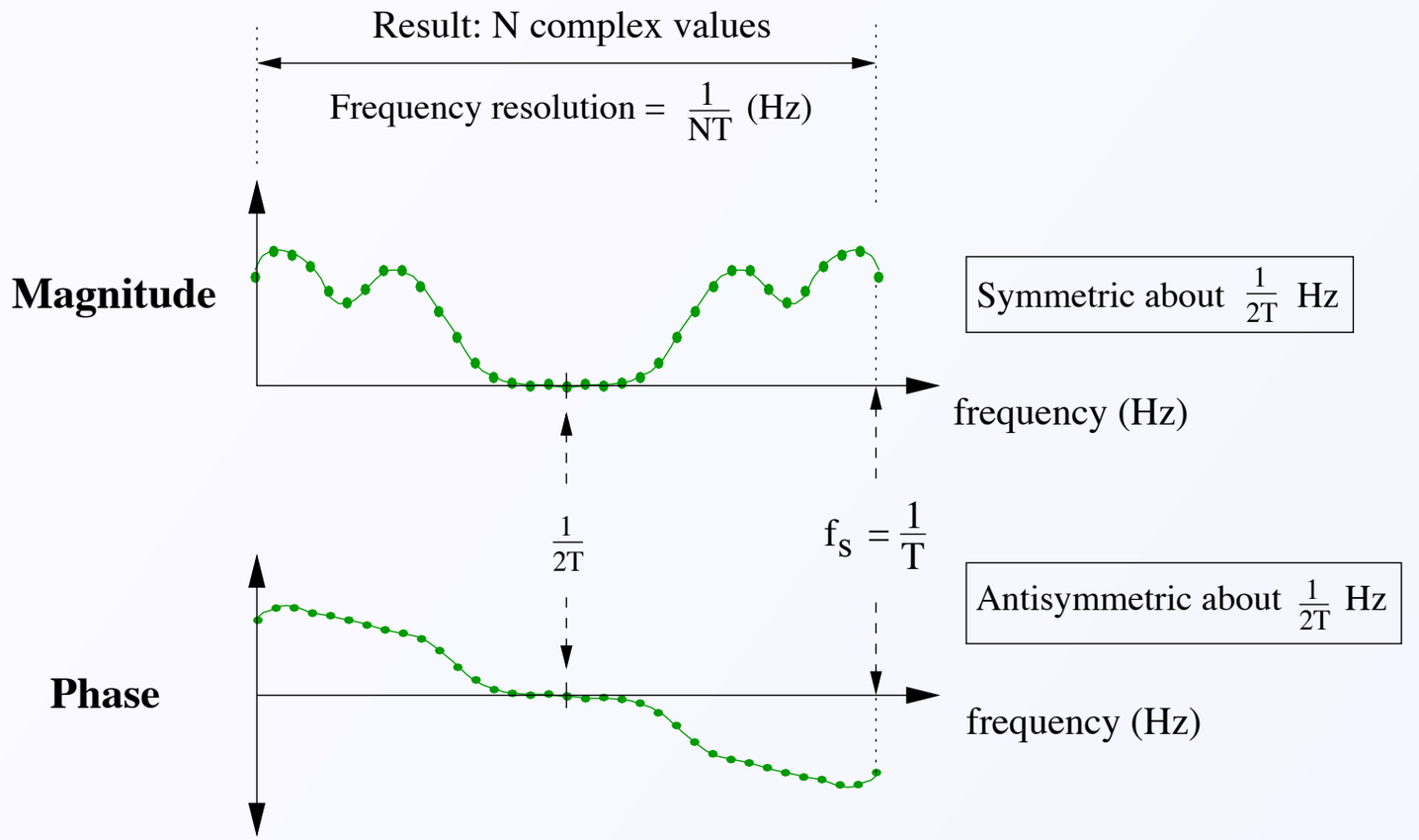

Signal frame
Hamming window
Windowed signal
Amplitude
Frequency

The discrete Fourier transform

Signal:



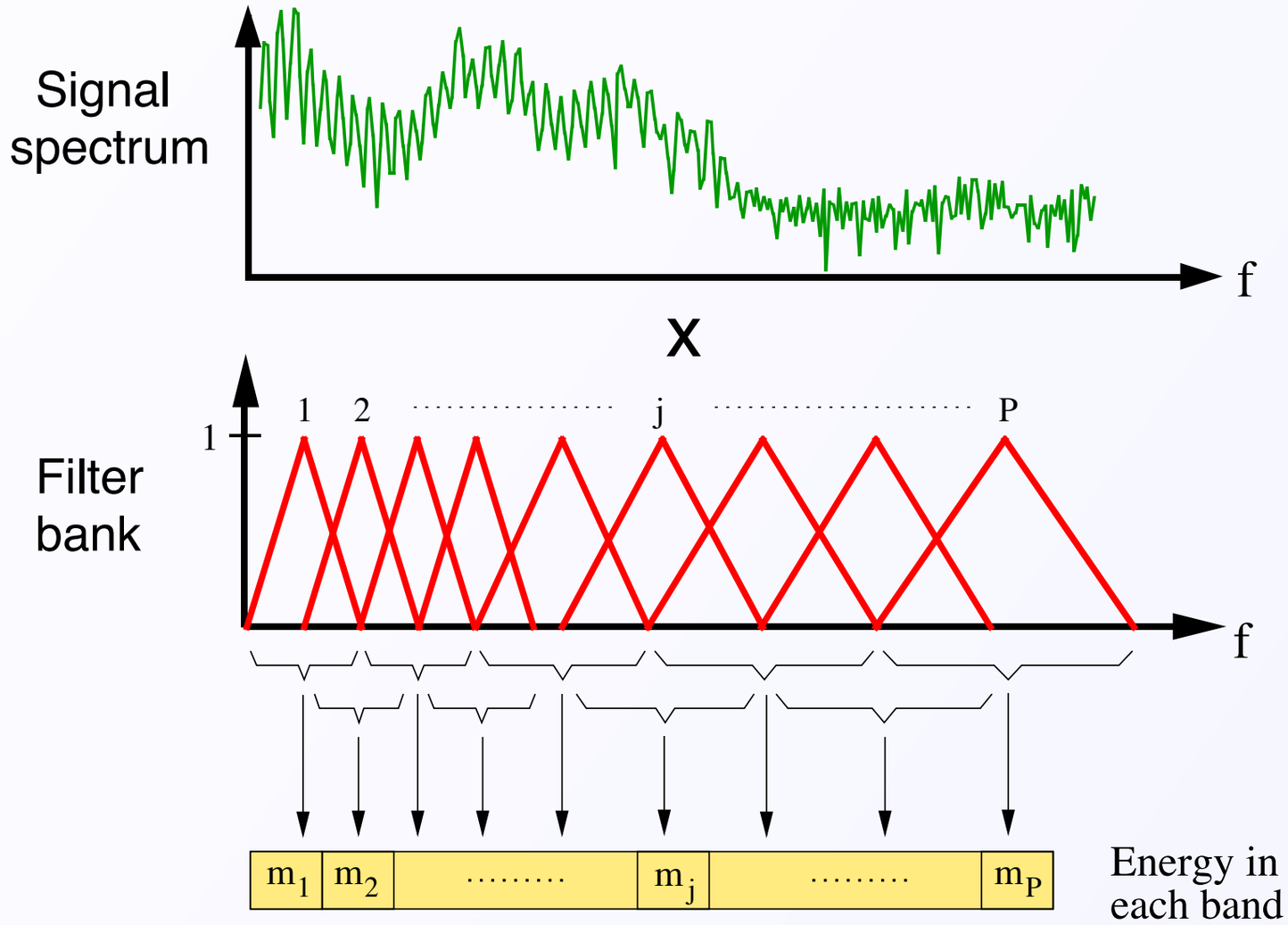
Spectrum:



Padding for use of the Fast Fourier Transform (FFT)

Mel-Scale Filterbanks

Band analysis to model masking effects by the ear together with non-linear frequency and bandwidth allocation.

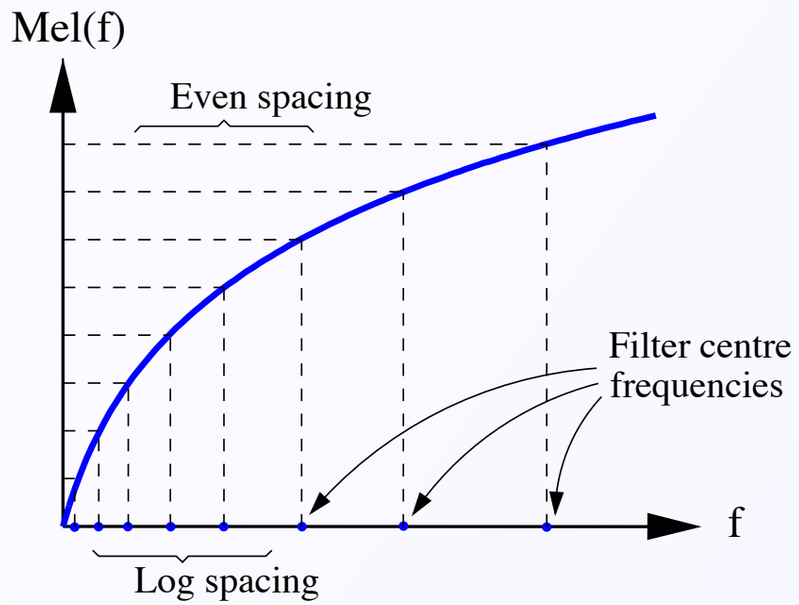


The energy in each frequency band is computed using a DFT.

The spacing of the center frequencies is based on the **Mel-scale**. The Mel-scale is defined as

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

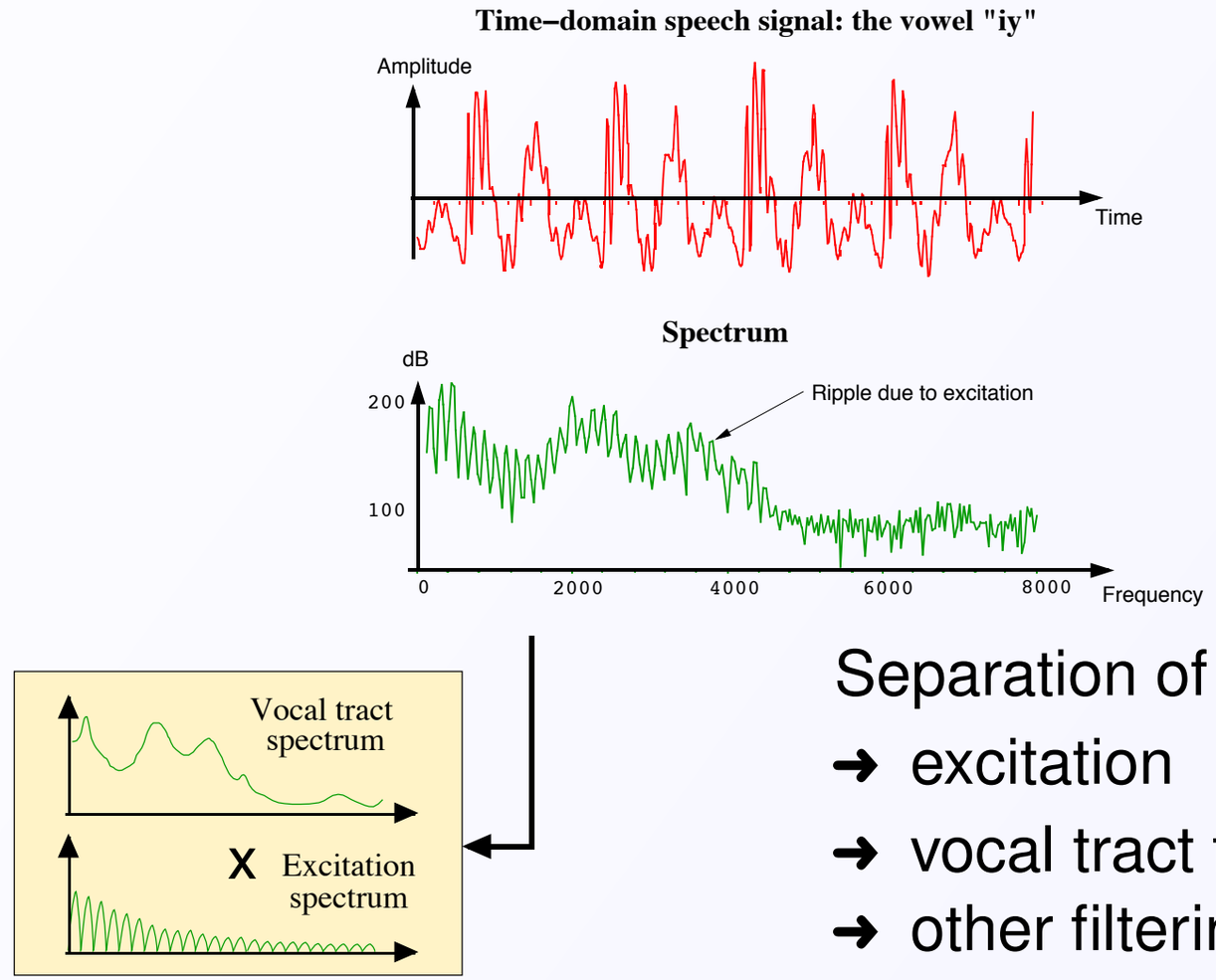
This frequency scale is shown below:



The scale is often regarded as being approximately linear up to 1kHz and logarithmic thereafter.

Cepstral Analysis

The source-filter model of speech production regards the spectrum of the signal as the product of the excitation spectrum and the vocal tract frequency response. The aim is to separate these:



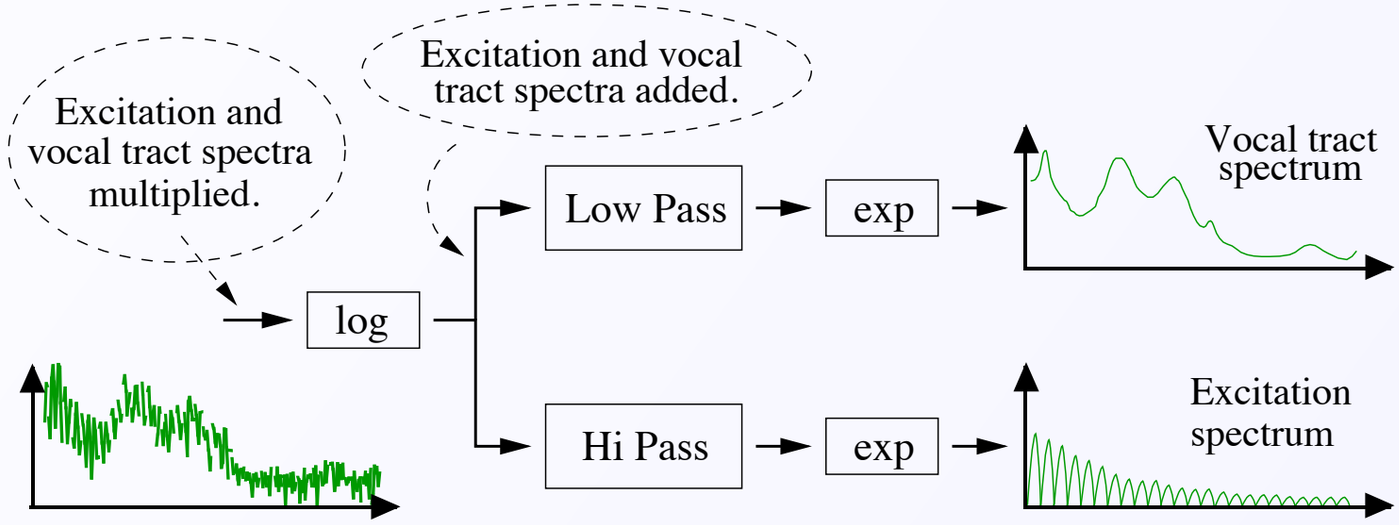
Separation of

- ➜ excitation
- ➜ vocal tract frequency response
- ➜ other filtering effects

Homomorphic Filtering

The excitation manifests itself as a quickly varying ripple in the spectrum. If these were **added** then the signals could be separated. However the signals have been **multiplied**! The solution is to take the logarithm to convert multiplication to addition:

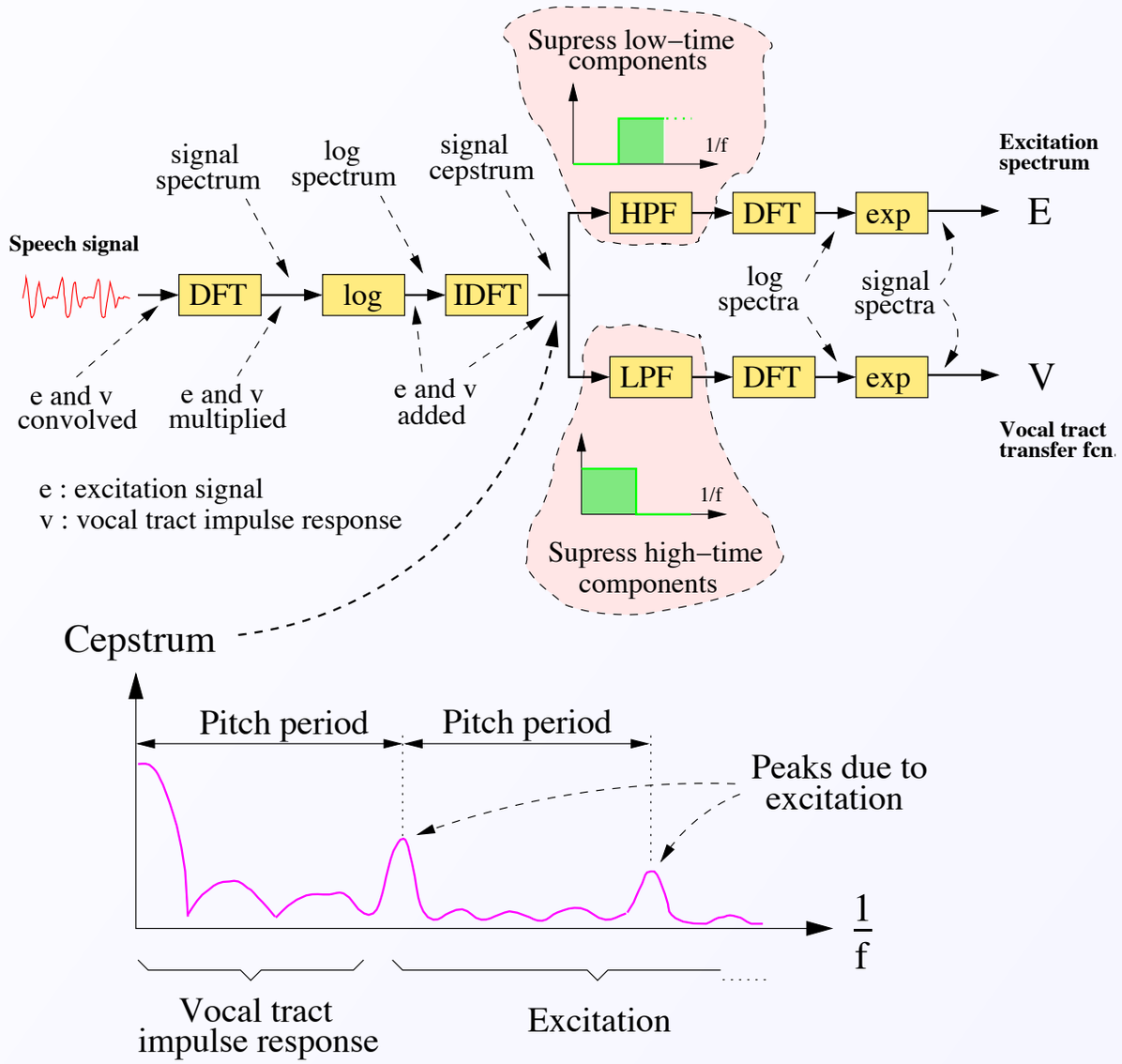
$$\begin{aligned} \log(a \cdot b) &= \log a + \log b \\ e^{\log y} &= y \end{aligned}$$



This approach is called **homomorphic filtering**. We are filtering the log spectrum as we would normally filter in the time domain.

The Cepstrum

Homomorphic filtering usually employs the DFT. Note that for the real cepstrum the log is applied to the magnitude spectrum.



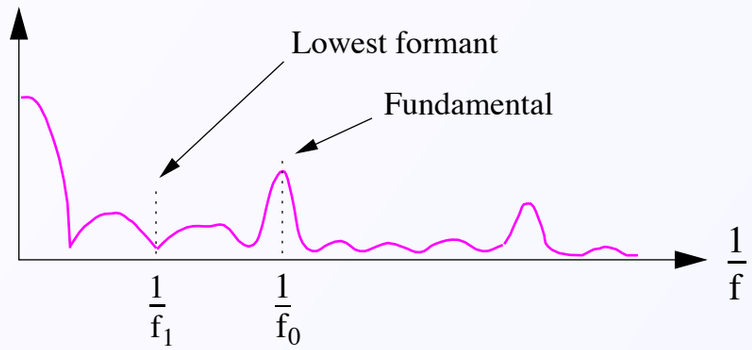
In cepstrum vocal tract impulse response decays rapidly and can be separated (by windowing) from the excitation.

The cepstrum is computed in **quefrency** domain and filtering in this domain is called **liftering**. Note that taking the IDFT of the cepstrum doesn't return to the time domain because of the non-linear log operation.

Assume that have the lowest formant at frequency f_1 and the fundamental at f_0 then cepstral coefficients

➜ $c_0 \rightarrow c_h$ encodes vocal tract response if $h \geq \frac{1}{f_1 T}$

➜ $c_p \rightarrow c_{N-1}$ include the major pitch peak if $p \geq \frac{1}{f_0 T}$



➜ If $f_0 < 250Hz$ and $f_1 > 500Hz$, and $T = 62.5\mu s$ then $h = 32$, $p = 64$

Making h smaller increases the smoothing over the whole spectrum.

Discrete Cosine Transform

Cepstral coefficients can be derived from the Mel filterbank energies using a simplified version of the DFT known as the discrete cosine transform (DCT). This uses the fact that the log magnitude spectrum is real-valued, symmetric with respect to 0 and periodic in frequency.

$$c_n = \sqrt{\frac{2}{P}} \sum_{i=1}^{P} m_i \cos \left[\frac{n(i - \frac{1}{2})\pi}{P} \right]$$

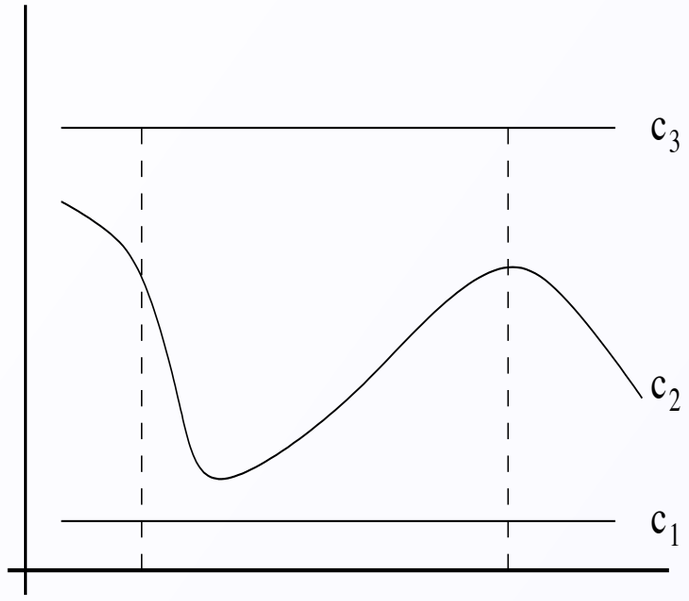
where P is the number of filterbank channels.

The representation found in this way is known as **Mel-frequency cepstral coefficients** (or **MFCCs**).

- ➜ The DCT decorrelates the spectral coefficients and allows them to be modelled with diagonal Gaussian distributions
- ➜ The number of parameters needed to represent a frame of speech is reduced. This in turn reduces memory and computation requirements.
- ➜ Note that c_0 is a measure of the signal energy

Dealing with variation - Differentials

MFCCs describe the instantaneous speech signal spectrum, but take no account of signal dynamics. Moreover, the use of **HMMs** requires to pack dynamic information into local vectors because of the **frame independence assumption**.



$$R_n(t) = \frac{\sum_{\tau=-\delta}^{\delta} \tau c_n(t+\tau)}{\sum_{\tau=-\delta}^{\delta} \tau^2}$$

This situation can be improved by including rates of change of the coefficients (differentials) in the representation. Typically regression coefficients are used to calculate the best straight line through a number of frames (here $2\delta + 1$)

For similar reasons the representation can also be augmented to include the 2nd differential of the *static* coefficients by again applying the regression formula to the 1st order differentials. The standard representation in modern HMM-based speech recognisers includes static, Δ and $\Delta\Delta$ coefficients.

Cepstral mean normalisation (CMN)

Any fixed linear filter that is applied to the speech (e.g. since the speech was recorded over a particular channel such as a specific microphone or over telephone) causes the measured speech spectrum, $Y(\omega)$ to be the product of the original speech $S(\omega)$ and the frequency response of the **linear filter** $H(\omega)$.

$$Y(\omega) = S(\omega)H(\omega)$$

Taking the logarithm

$$\log Y(\omega) = \log S(\omega) + \log H(\omega)$$

and

$$\log |Y(\omega)| = \log |S(\omega)| + \log |H(\omega)|$$

The distortion of the speech signal in the log-domain is additive !

CMN (2)

If $\log|H(\omega)|$ can be estimated it can be subtracted and the measured speech is independent of channel! The idea is that (within an additive constant) $\log|H(\omega)|$ as the average value of $\log|Y(\omega)|$ over the utterance (or a set of utterances over the same channel) since it is assumed that on average $\log|S(\omega)|$ is flat.

$\mathbf{y}_t$ represents a frame of observed speech in the spectral domain, hence $\mathbf{y}_t = \mathbf{s}_t + \mathbf{h}_t$ ($\mathbf{h}_t$ is assumed to be constant over time). Then

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{s}_t + \mathbf{h}_t) = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{s}_t + \mathbf{h}) = \frac{1}{T}T\mathbf{h} + \frac{1}{T}\sum_{t=1}^{T}(\mathbf{s}_t) \approx \mathbf{h}$$

Since the cepstrum is a linear transformation of the log spectrum, as are MFCCs (multiply the filter energies by a fixed matrix), for MFCCs the average cepstrum (e.g. over an utterance) can be subtracted from all the static coefficients. This is termed **cepstral mean normalisation** (*subtraction*, *removal*).

An ASR front-end