Speech Preprocessing, Speech Production, Cepstrum

Jan Černocký, Valentina Hubeika

{cernocky|ihubeika}@fit.vutbr.cz

FIT BUT Brno

Agenda

- Speech Parameterization
  - Preprocessing
  - Basic Parameters: short-time energy, zero crossing rate.
- Speech production and its model.
- Spectrogram.
- Separation of excitation and modification – cepstrum.
- Approximation of cepstra according to the human auditory system's response– MFCC.

PARAMETERIZATION

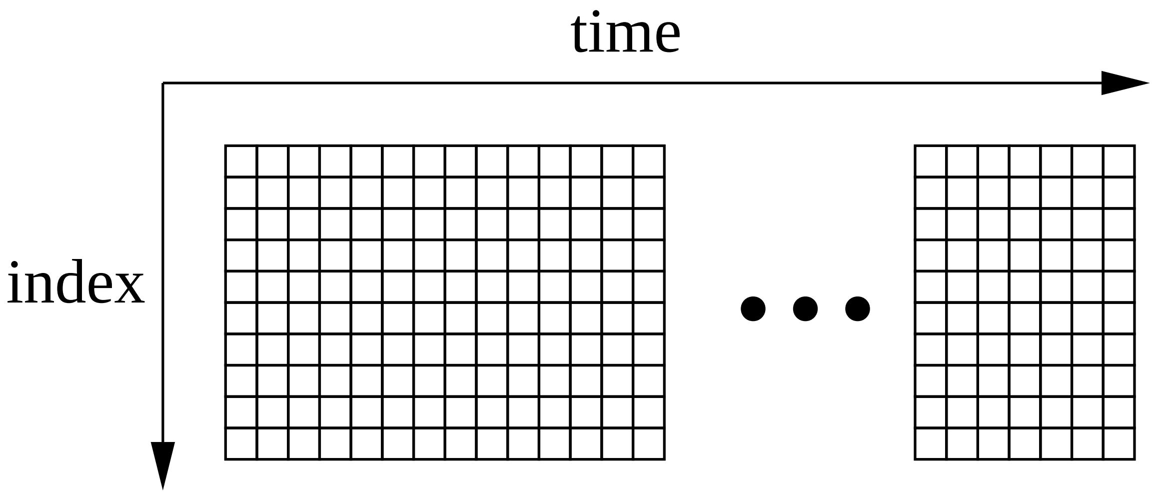
- Goal: express a signal on a limited number of values – a) "parameterization", b) "feature extraction".
- a) representation based on findings in signals processing (filter banks, Fourier transform, etc.) $\Rightarrow$ *non-parametric representation*.
- b) representation based on findings about speech production $\Rightarrow$ *parametric representation*.

BUT:

- b) also makes use of the techniques of non-parametric representation, thus difficult (sometimes unfavorable) to distinguish between the two groups.
- The calculated values are anyway usually called *parameters*.

Parameters

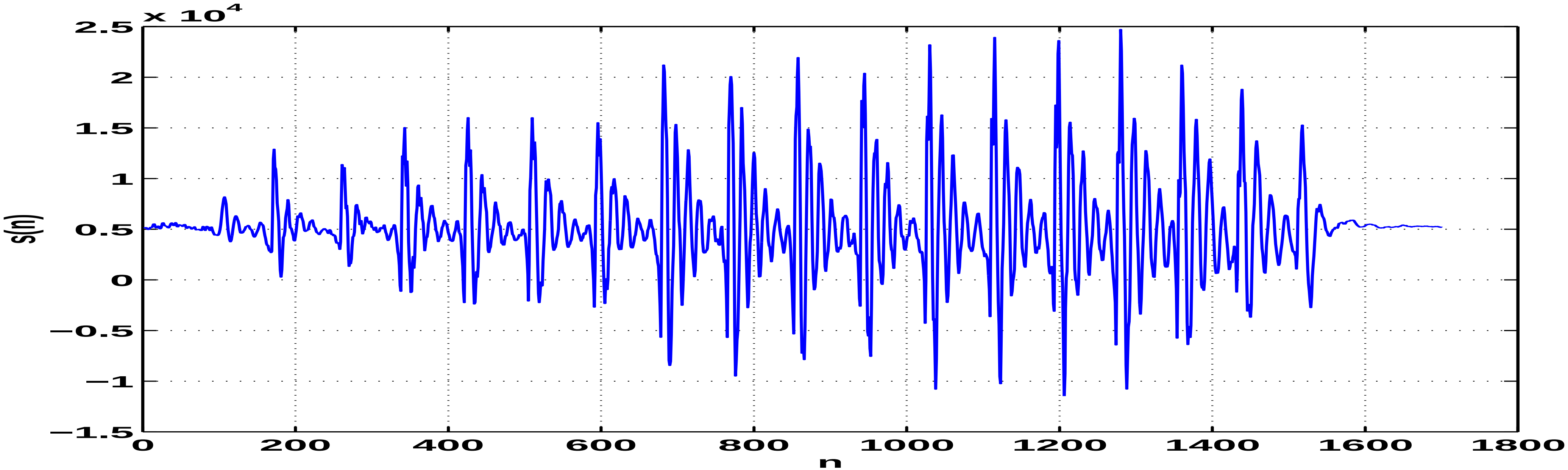
- **scalar per frame** – one number calculated from a speech frame (short-time energy or zero crossing rate).
- **vector per frame** – a set of numbers (vector) calculated from a speech frame. When having a sequence of frames, parameters are usually stored in *matrices*.



PRE-PROCESSING

Mean Normalization

Direct current offset (dc-offset) carries no useful information. Moreover, can carry disturbing information (when calculating energy).



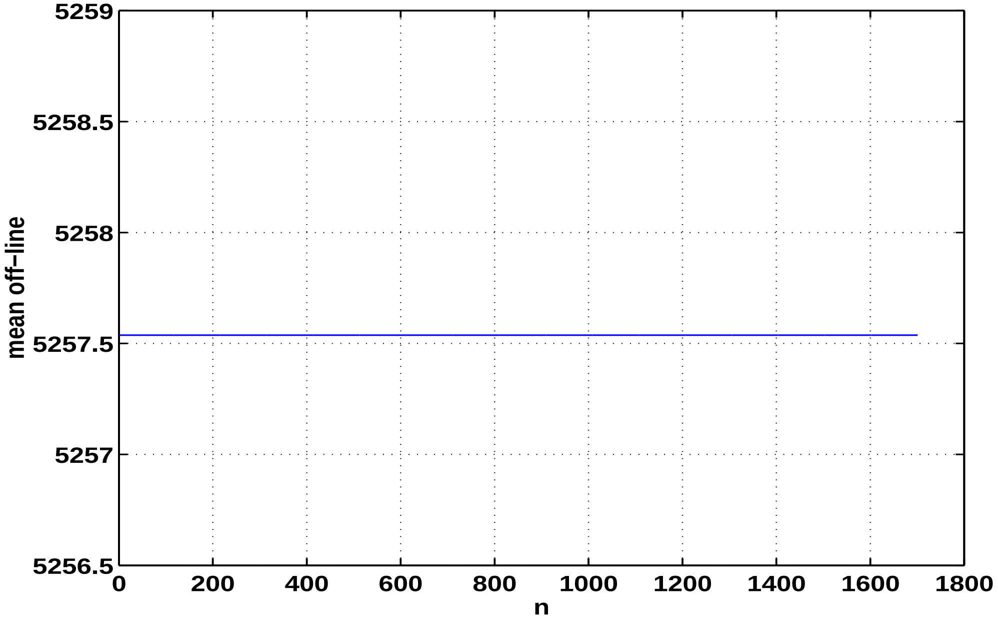
dc-offset removal!

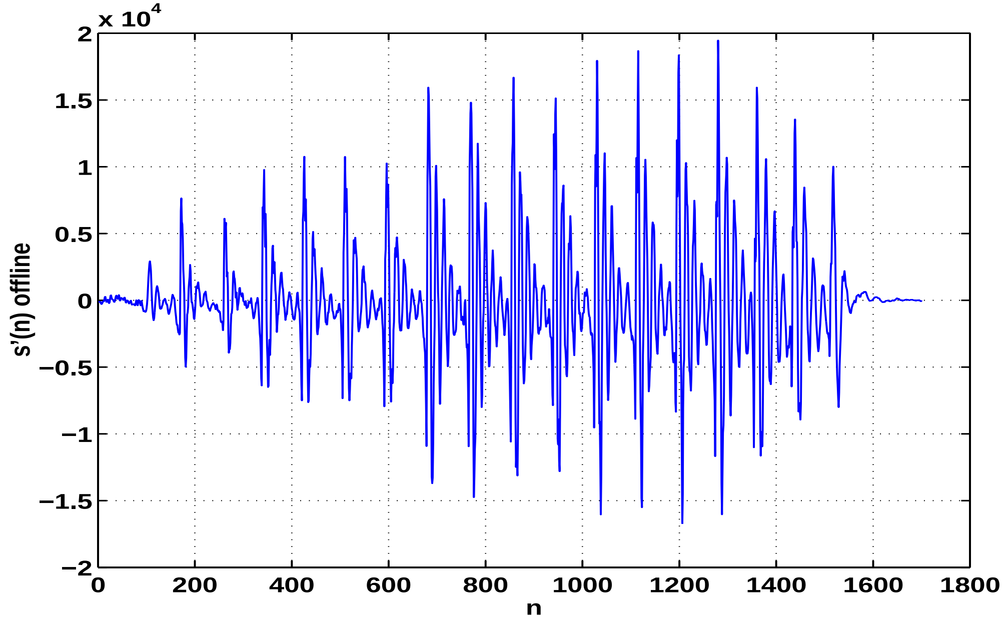
$$s'[n] = s[n] - \mu_s, \qquad \mu_s \text{ must be estimated.} \tag{1}$$

Mean Value Off-Line

Here, equivalent to the average value:

$$\bar{s} = \frac{1}{N} \sum_{n=1}^{N} s[n] \quad (2)$$





Mean Value On-Line

The whole signal is not (yet) available: either is too long or there is a flow of new values..

$$\bar{s}[n] = \gamma \bar{s}[n-1] + (1-\gamma) s[n], \quad (3)$$

where $\gamma \longrightarrow 1$. This is equivalent to passing a signal through a filter with the impulse response:

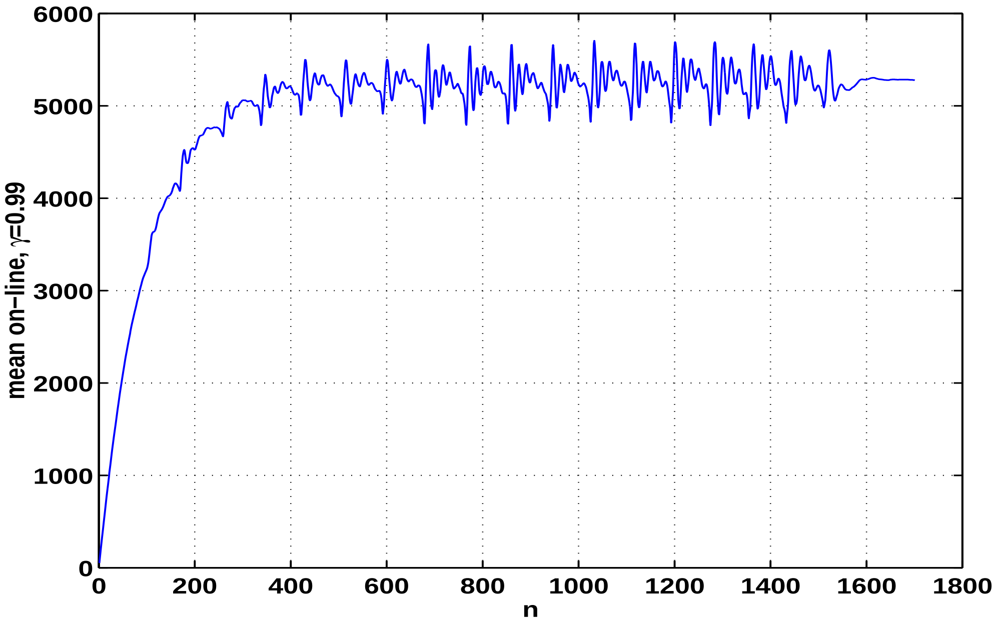
$$h = [(1-\gamma) \;\; (1-\gamma)\gamma \;\; (1-\gamma)\gamma^2 \;\; \ldots]. \quad (4)$$

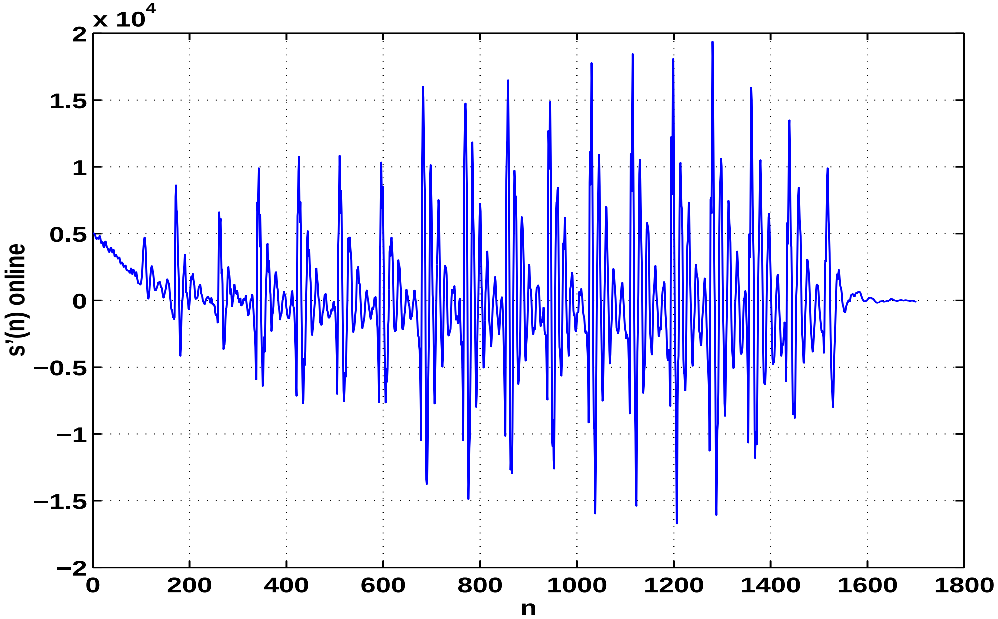
Defined as the geometric progression: the initial element is $a_0 = 1 - \gamma$ and the quotient is $q = \gamma$. The sum is thus:

$$\sum_{n=0}^{\infty} h[n] = \frac{a_0}{1-q} = \frac{1-\gamma}{1-\gamma} = 1, \quad (5)$$

(this is what we originally expected ☺).

Example, $\gamma = 0.99$ (see the first computer lab):





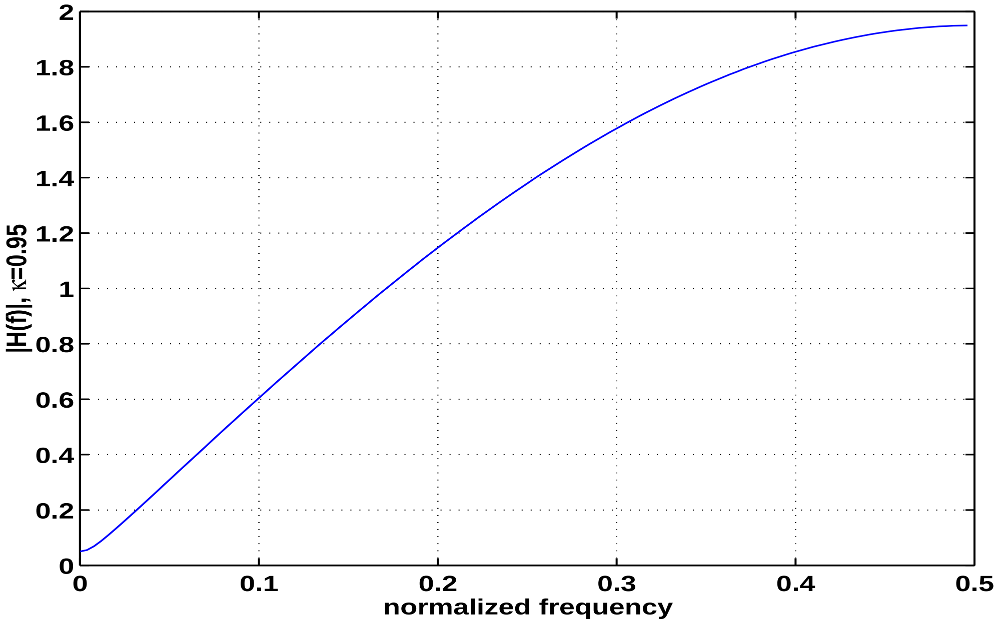
Preemphasis

Increases the magnitude of the higher frequencies with respect to the magnitude of the lower frequencies. Equalization of the speech frequency characteristics (the magnitude decreases towards higher frequencies). Rather a historical operation.

A simple first-order FIR filter:

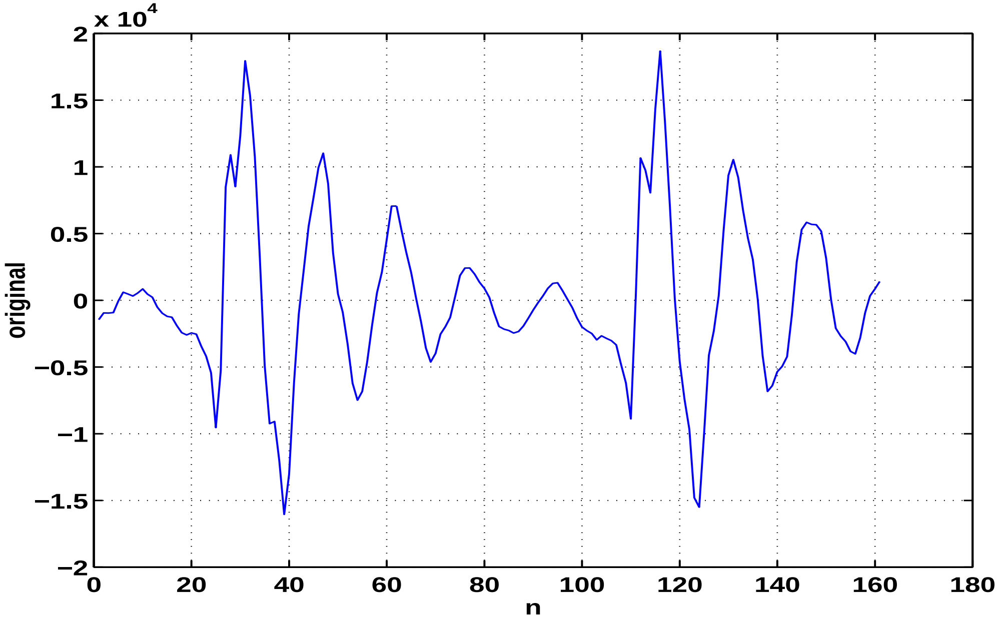
$$H(z) = 1 - \kappa z^{-1}, \tag{6}$$

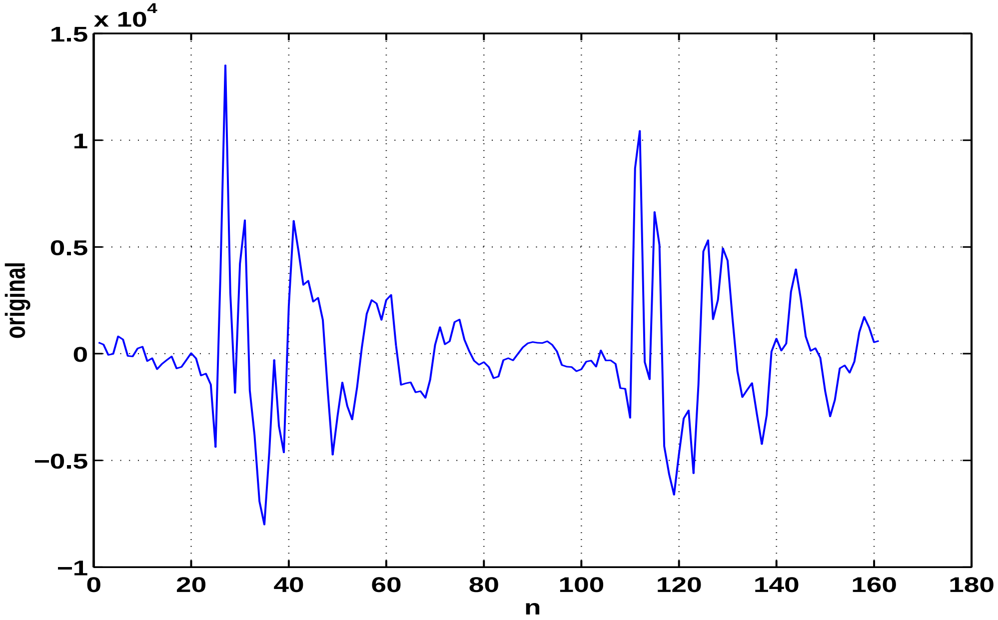
where $\kappa \in [0.9,\ 1]$. Calculated difference between the two neighboring samples. The magnitude frequency characteristics for κ=0.95:



Passing through the defined filter:

$$s'[n] = s[n] - \kappa s[n-1] \tag{7}$$





⇒ The processed signal (after applying preemphasis) contains of more higher frequencies.

FRAMES

- Why?
- Speech signal is considered as *random*, parameter estimation methods require *stationary* signals.
- Thus dividing the signal into shorter segments (segments, micro-segments, frames) within which the signal behaves (we hope) as stationary.
- Frame parameters: length l_{ram}, overlap p_{ram}, frame shift $s_{ram} = l_{ram} - p_{ram}$.

Frame Length

1. *short* enough to assume the signal (within the given length) is stationary.
2. BUT: *long* enough to provide accurate estimation of the desired parameters (features).

$\Rightarrow$ trade off (momentum of the articulation tract), typical length 20–25 ms (160–200 samples for F_s =8000 Hz).

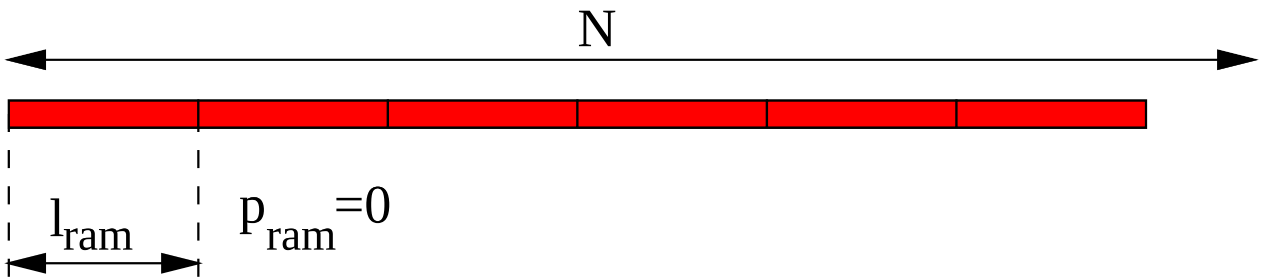
Overlap

- **small or none:** ☺ fast time shift in the signal, low memory/processor demands, ☹ the difference of the parameter values of the neighboring frames can be significant.
- **large:** ☺ slow time shift, smooth change in the parameter values, ☹ high memory/processor demands, alike parameter values (violates the independency assumption!).

$\Rightarrow$ tradeoff, typical length 10 ms, thus 100 frames per second, centi-second vectors.

How many frames per segment of the length N ?

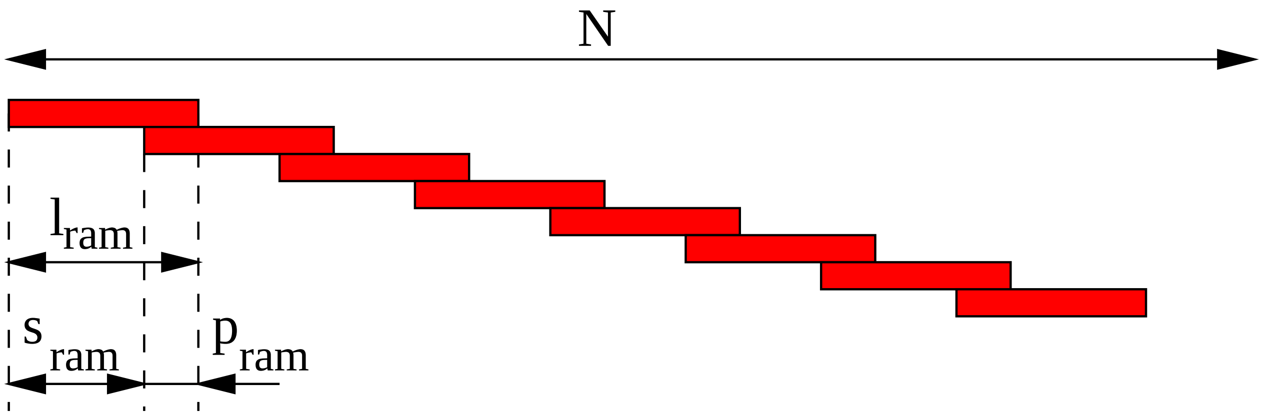
No overlap, $p_{ram} = 0$



$$N_{ram} = \left\lfloor \frac{N}{l_{ram}} \right\rfloor, \tag{8}$$

... $\lfloor \cdot \rfloor$ denotes the operation 'floor'.

Frames overlap, $p_{ram} \neq 0$



$$N_{ram} = 1 + \left\lfloor \frac{N - l_{ram}}{s_{ram}} \right\rfloor \tag{9}$$

. . . the signal must be at least one frame long.

Signal Segmentation - Windowing Function

Select a frame of a signal using a window - window(ing) function:

Rectangular – no change of the signal, selection only:

$$w[n] = \begin{cases} 1 & \text{pro} \quad 0 \leq n \leq l_{ram} - 1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

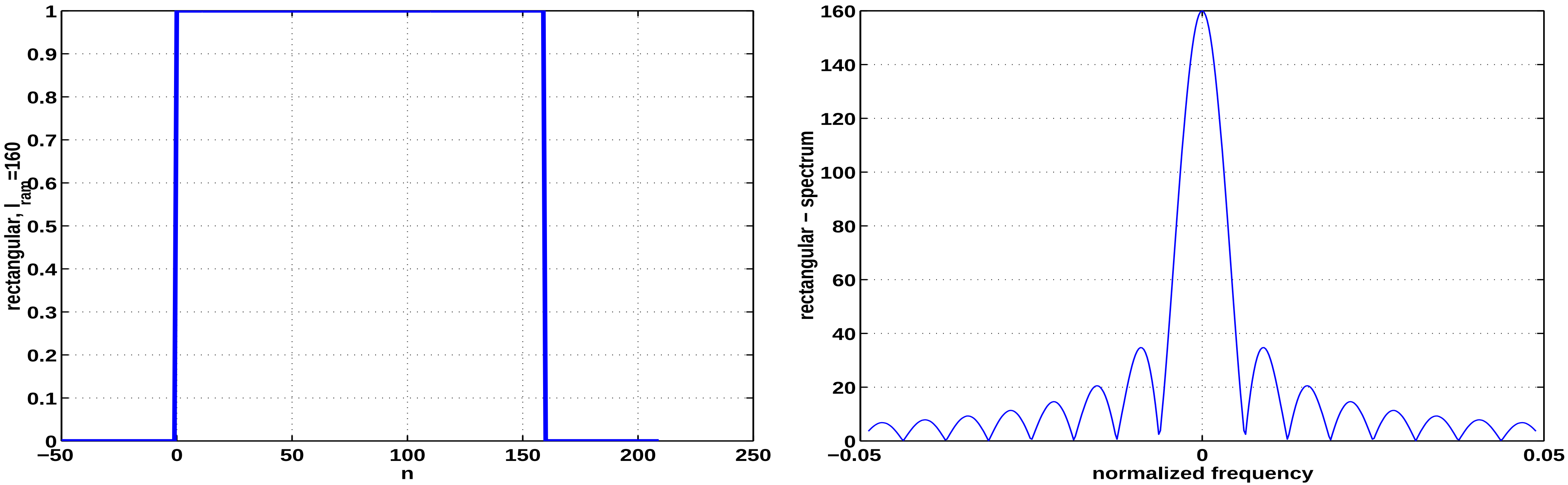
Hamming – suppresses the signal at the sides of the window, selection and weighting:

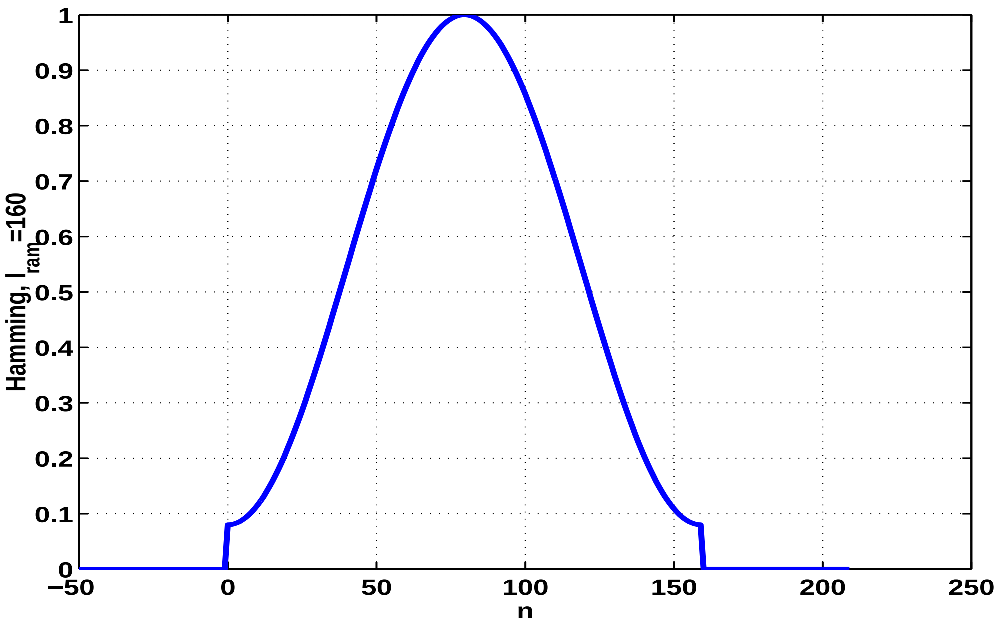
$$w[n] = \begin{cases} 0.54 - 0.46 \cos \frac{2\pi n}{l_{ram} - 1} & \text{pro} \quad 0 \leq n \leq l_{ram} - 1 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

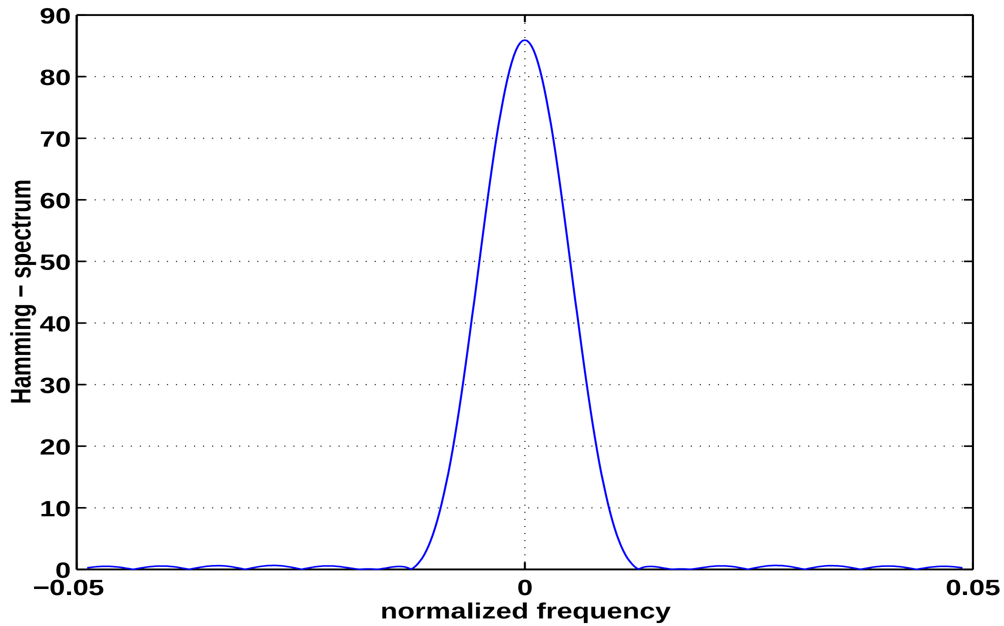
How does windowing change the spectrum of the selected segment? A product in time domain corresponds to *convolution* of the speech spectrum with the window spectrum.

$$X(f) = S(f) \star W(f) \tag{12}$$

Comparison of the rectangular and the Hamming window:




Hamming, l_ram=160
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−50
0
50
100
150
200
250
n


Hamming – spectrum
90
80
70
60
50
40
30
20
10
0
−0.05
0
0.05
normalized frequency

BASIC PARAMETERS OF SPEECH SIGNAL

all the parameters will be derived for single frames. For each frame:

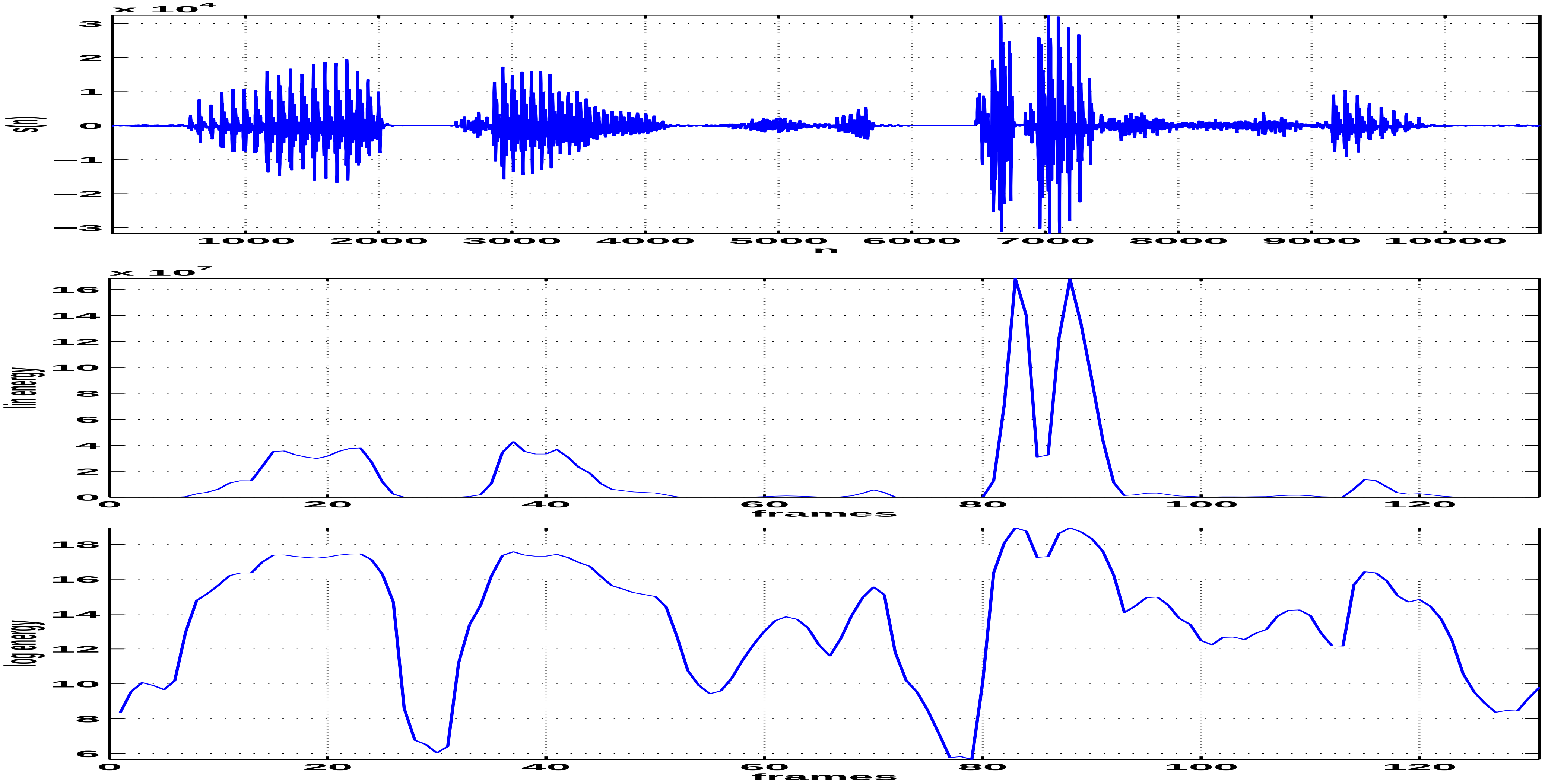
- scalar ⇒ a row vector.
- vector ⇒ matrix, columns contain dimensions of the parameter vector, rows contain a sequence of values in a particular dimension over time (time in frames).

Average Short-Time Energy

$$E = \frac{1}{l_{ram}} \sum_{n=0}^{l_{ram}-1} x^2[n] \tag{13}$$

- speech activity detector.
- separation of phonemes to voiced (high energy) and unvoiced (low energy).
- often we use log-energy.
- careful with noise and low-energy phonemes.

Example: "létající prase"



Zero-Crossing Rate

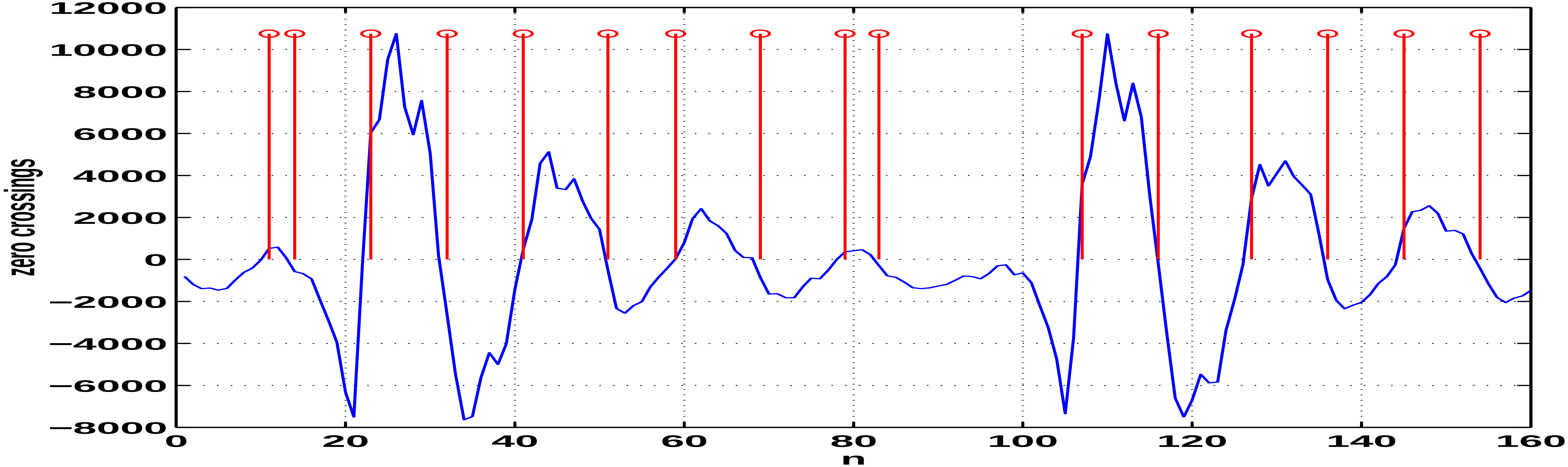
. . . rate of sign changes of the signal within a given frame.

$$Z = \frac{1}{2} \sum_{n=1}^{l_{ram}-1} |\,\text{sign}\, x[n] - \text{sign}\, x[n-1]|, \qquad (14)$$

where sign (x) is the sign function defined as:

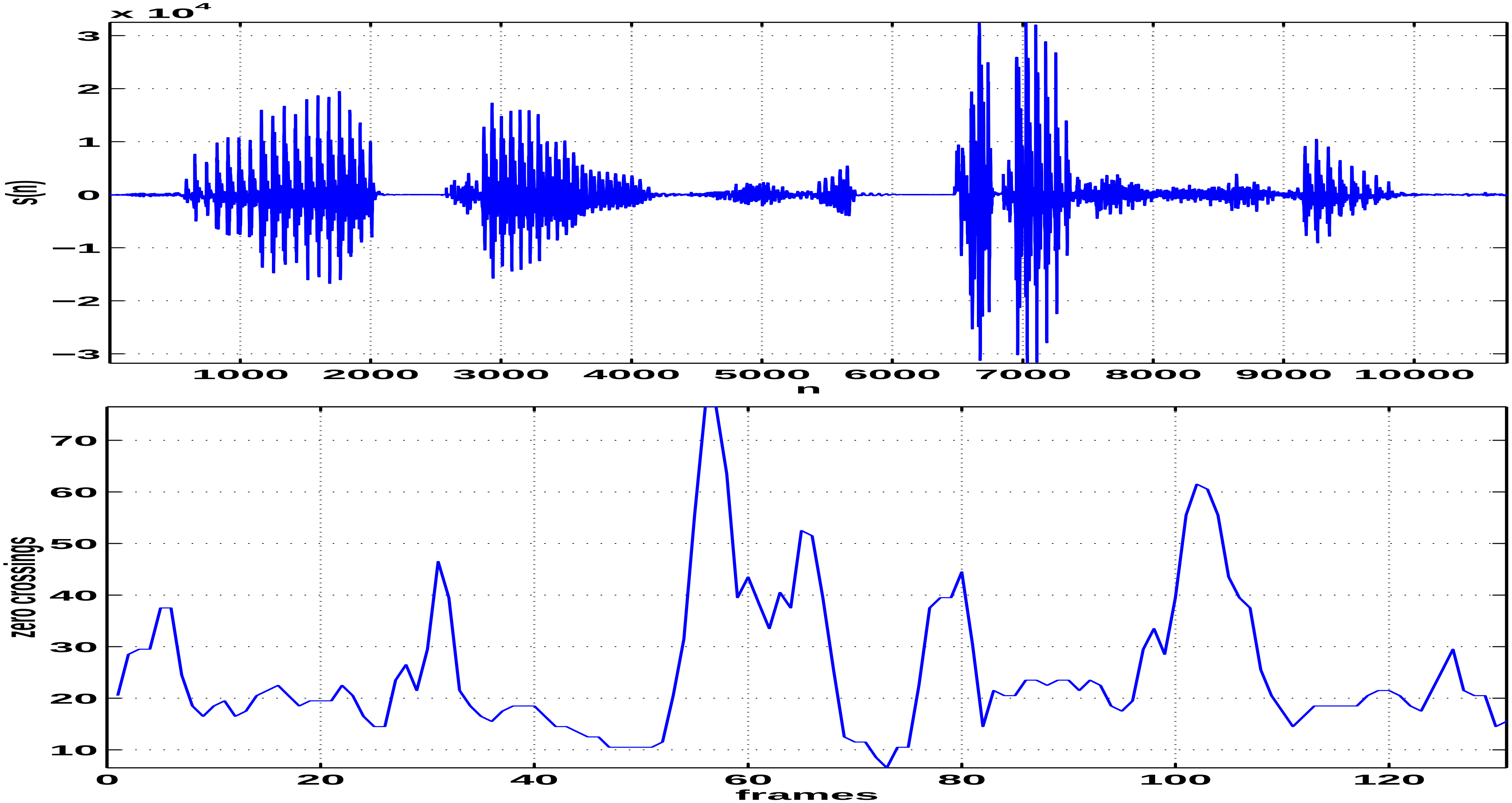
$$\text{sign}\, x[n] = \begin{cases} +1 & \text{pro} \quad x[n] \geq 0 \\ -1 & \text{pro} \quad x[n] < 0 \end{cases} \qquad (15)$$

How does it work? The function $|\,\text{sign}\, x[n] - \text{sign}\, x[n-1]|$ results in 2 when there is a change in the sign between the samples $x[n-1]$ and $x[n]$:



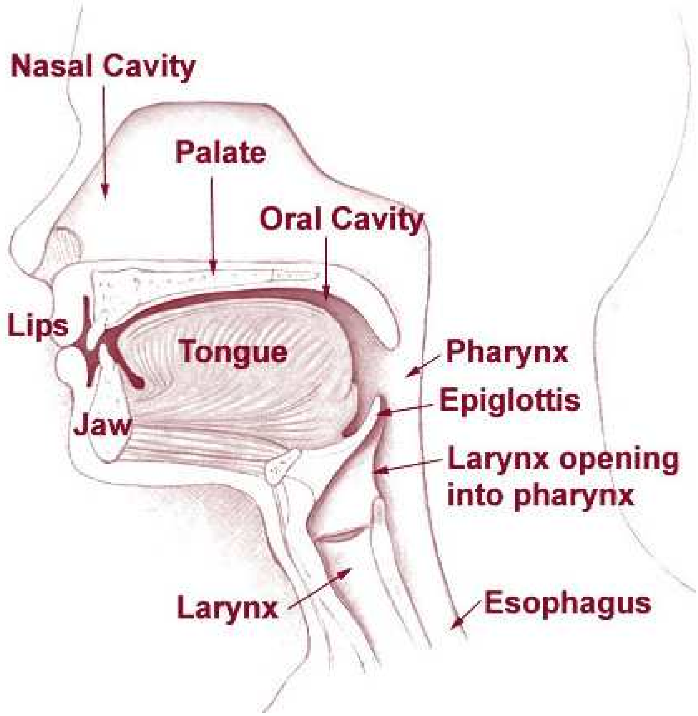
- distinguishing between the voiced (low zero-crossing rate) and unvoiced (high rate, rather like in noise).
- very sensitive to noise. . .

Example: "létající prase"



HUMAN VOCAL APPARATUS AND ITS MODEL

(Adopted from Wikipedia)



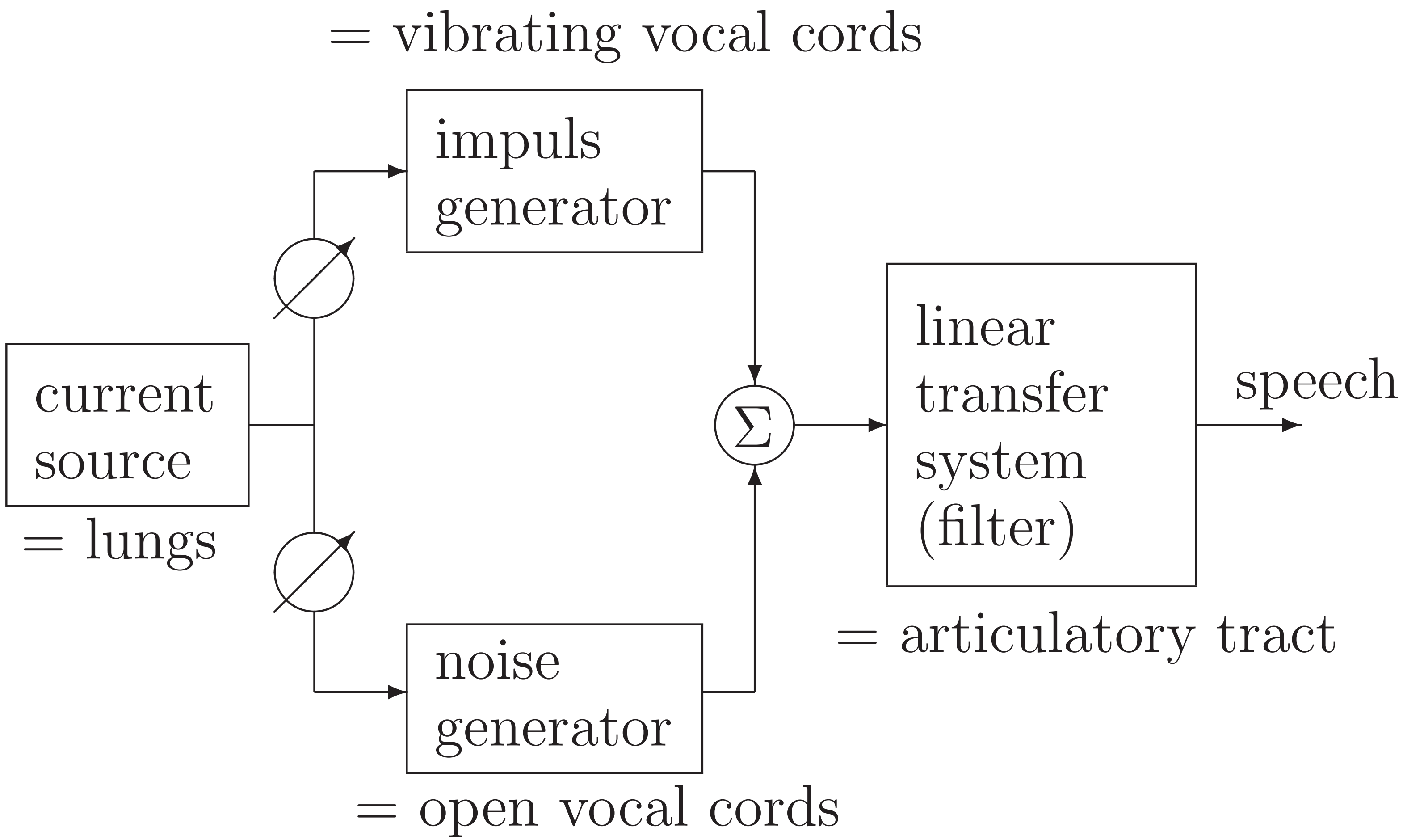
Organs and their Models in Digital Processing

- **lungs** — energy source — **signals: none**
- **larynx** — energy modulation – **signals: excitation**.
  - opened vocal cords — noise.
  - vibrating vocal cords — periodic signal (tone). **fundamental frequency**:

males	90–120 Hz
females	150–300 Hz
children	350–400 Hz

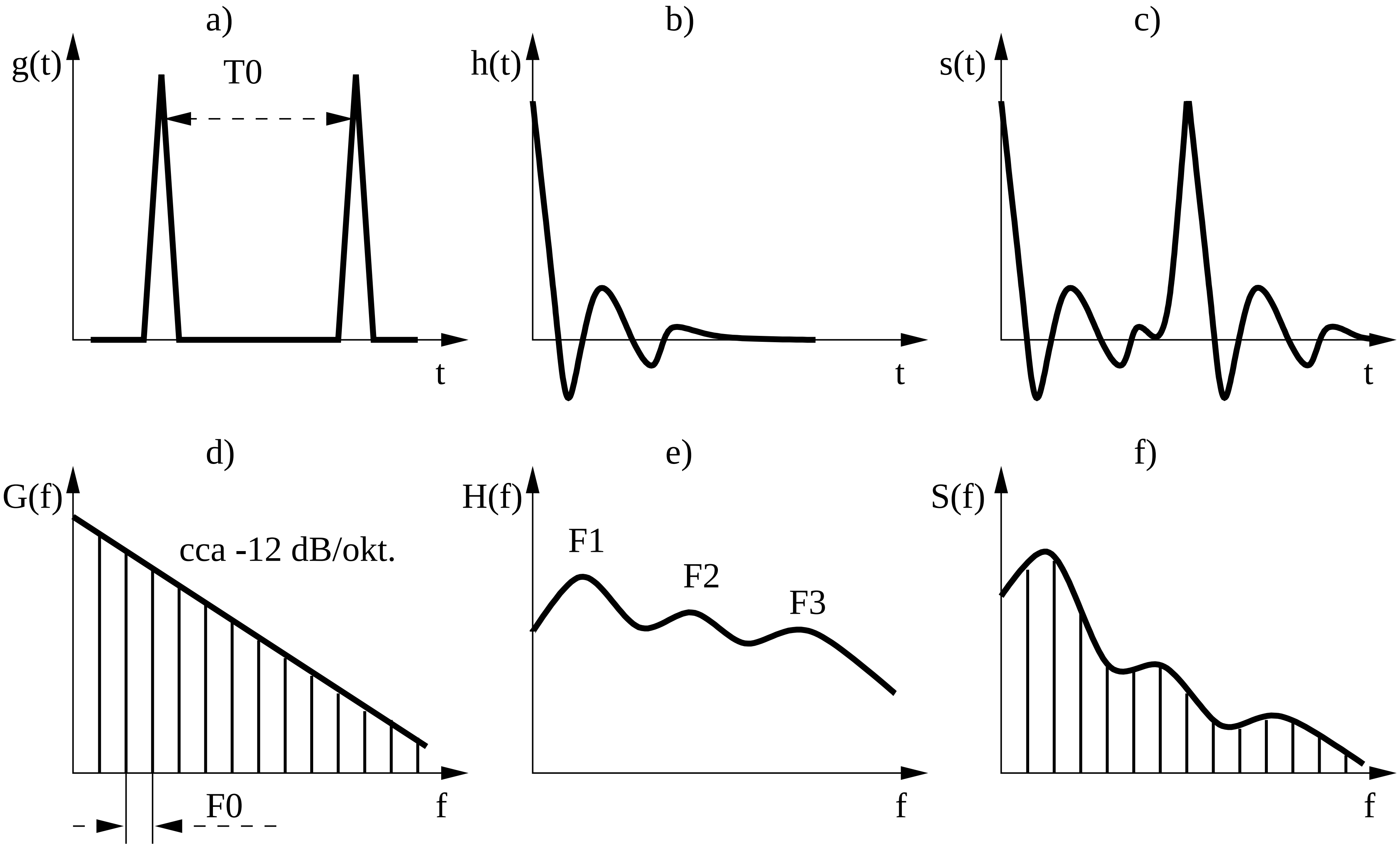
- **vocal (articulatory) tract** — modification tract — **signals: filter**.
  - pharynx.
  - velum.
  - tongue.
  - oral and nasal cavity.
  - teeth.
  - lips.

Model



Transfer system: linear filter - usually IIR.

Vocal Tract Model in Time and Frequency Domain



Top part – time behaviour, bottom part – spectrum.

- a) and d) excitation: T_0 is the period, F_0 is the fundamental frequency (pitch).
- b) and e) articulation tract: F_1 to F_3 are the formants (resonance frequency of the vocal tract), are given by the physical configuration of the vocal tract.
- c) and f) the resulting signal and its spectrum.

The resulting signal is given in the time domain by *convolution*:

$$s(t) = g(t) \star h(t) = \int_{-\infty}^{+\infty} g(\tau)h(t-\tau)d\tau. \tag{16}$$

Convolution in time domain corresponds to *product* in frequency:

$$S(f) = G(f)H(f). \tag{17}$$

A relevant task in speech processing is **de-convolution**; the goal is to separate excitation and modification.

SPECTROGRAM

One spectrum is not enough (speech is non-nonstationary) ⇒ representation of the spectrum (strictly speaking PSD) behaviour over time:

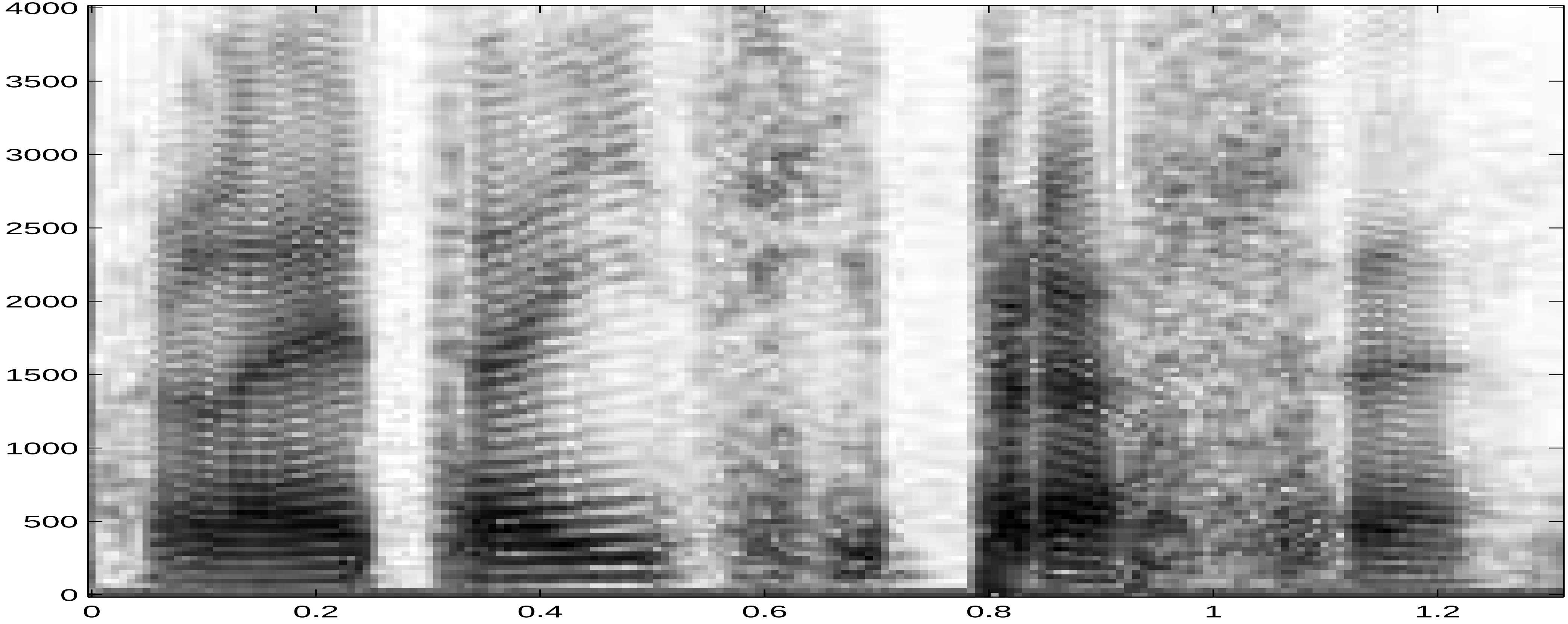
- segment speech into frames.
- estimate the PSD for each frame, usually using DFT.
- depict:
  - horizontal axes represents time ("rough" time in frames).
  - vertical axes represents frequency.
  - color represents energy.

Depending on the frame length we talk about:

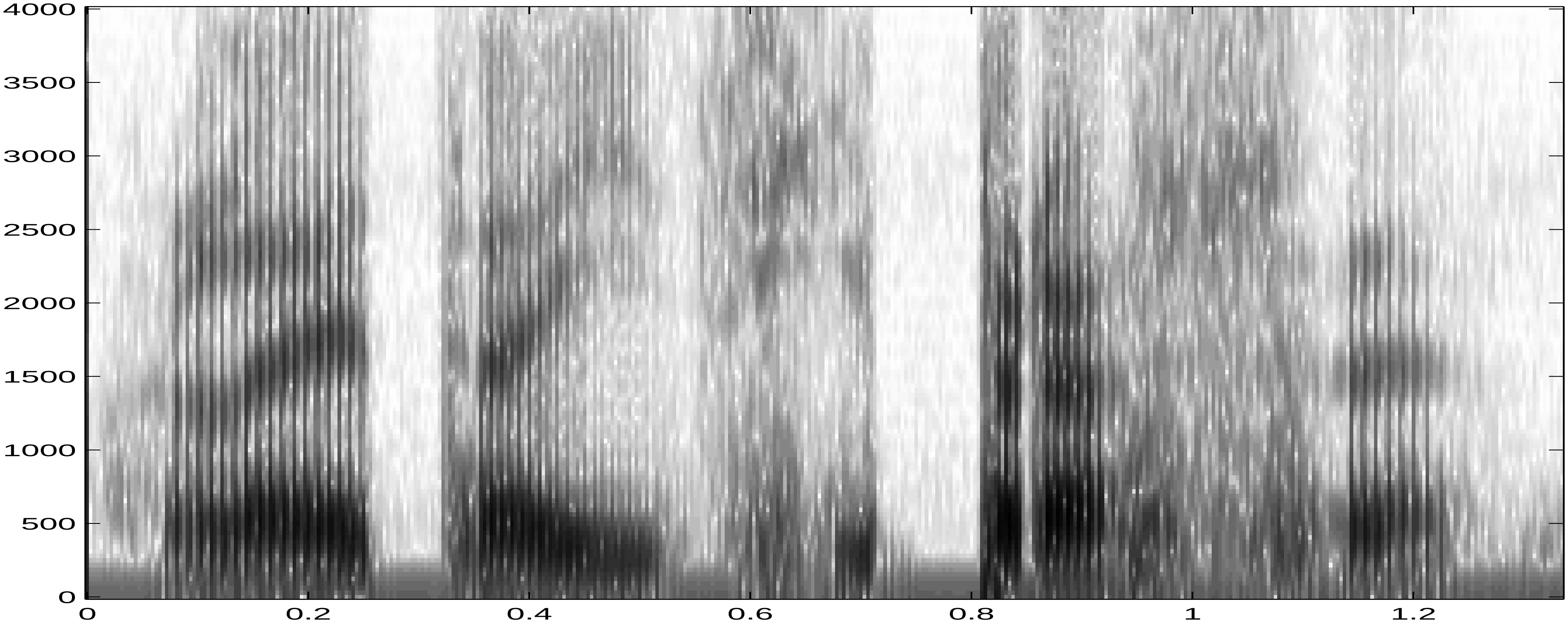
- long-term spectrogram.
- short-term spectrogram.

☹ Drawback of DFT: Fine scale in frequency and time domain cannot be satisfied simultaneously

long-term: `specgram(s,256,8000,hamming(256),200);`



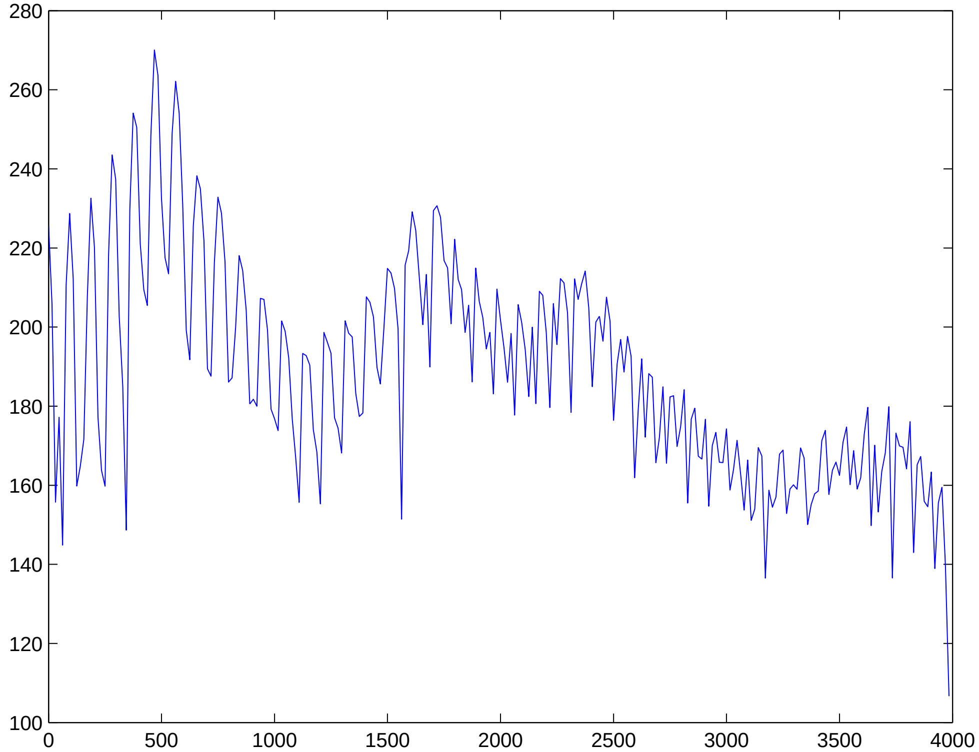
short-term: `specgram(s,256,8000,hamming(50));`



CEPSTRUM

. . . separates excitation from modification – convenient for encoding; dropping excitation frequency in speech processing (excitation carries information dependent on speaker, mood,. . .)

What can we do.. 1: filter off frequency lower than 400 Hz and get rid of the fundamental frequency. . . **BAD IDEA:**



- fundamental frequency folds are present along the whole spectrum.
- we can loose the fist formant.
- land line band starts on 300 Hz and we still can recognize the pitch.
- . . . so we need a better approach.

⇒ **Cepstrum**

Challenge

Excitation $e(t)$ is convoluted with the filter (modification) impulse response:

$$s(t) = g(t) \star h(t) = \int_{-\infty}^{+\infty} g(\tau)h(t-\tau)d\tau, \quad (18)$$

which in frequency domain corresponds to *product*:

$$S(f) = G(f)H(f). \quad (19)$$

we cannot well separate the two components in either domain. Solution: **non-linearity**, which can translate product to summation.

Definition of Cepstrum

$$\ln G(f) = \sum_{n=-\infty}^{+\infty} c(n) e^{-j2\pi fn} \tag{20}$$

The $c(n)$ values are the **cepstral coefficients**. Since $G(f)$ is an even function, $c(n)$ are real and the following holds:

$$c(n) = c(-n) \tag{21}$$

The sum in the equation is the definition of DFT, hence we can compute the $c(n)$ as:

$$c(n) = \mathcal{F}^{-1}\left[\ln G(f)\right] \tag{22}$$

DFT-cepstrum

$$c(n) = \mathcal{F}^{-1}\left\{\ln |\mathcal{F}[s(n)]|^2\right\}, \tag{23}$$

- spectrum $\longrightarrow$ cepstrum.

Can it really "break" convolution?

$$s(n) = e(n) \star h(n), \quad (24)$$

$$S(f) = E(f)H(f) \text{ a thus } |S(f)|^2 = |E(f)|^2 \; |H(f)|^2. \quad (25)$$

For the cepstrum calculation we make use of the linearity of the inverse Fourier transform:

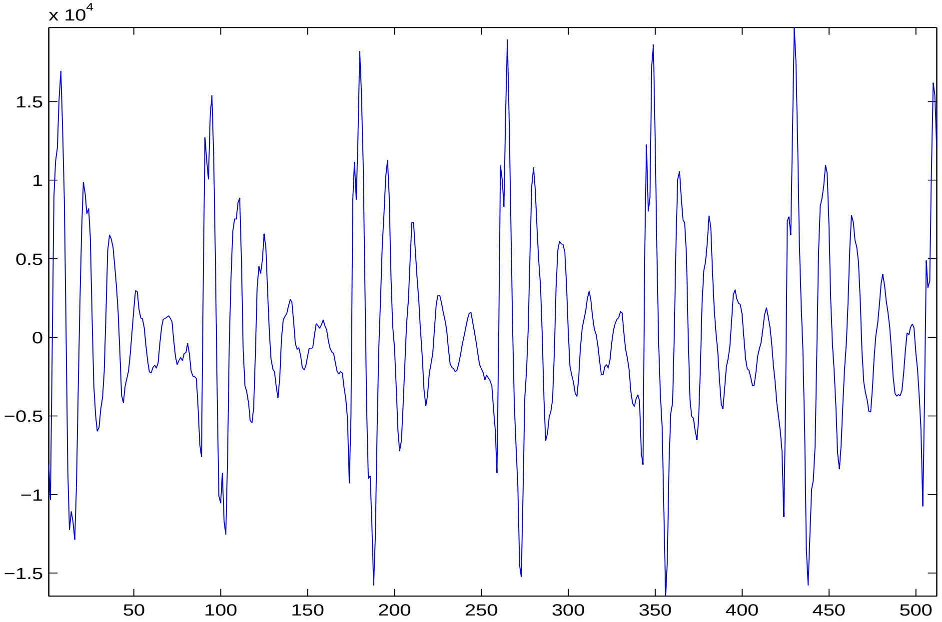
$$\mathcal{F}^{-1}(a+b) = \mathcal{F}^{-1}(a) + \mathcal{F}^{-1}(b).$$

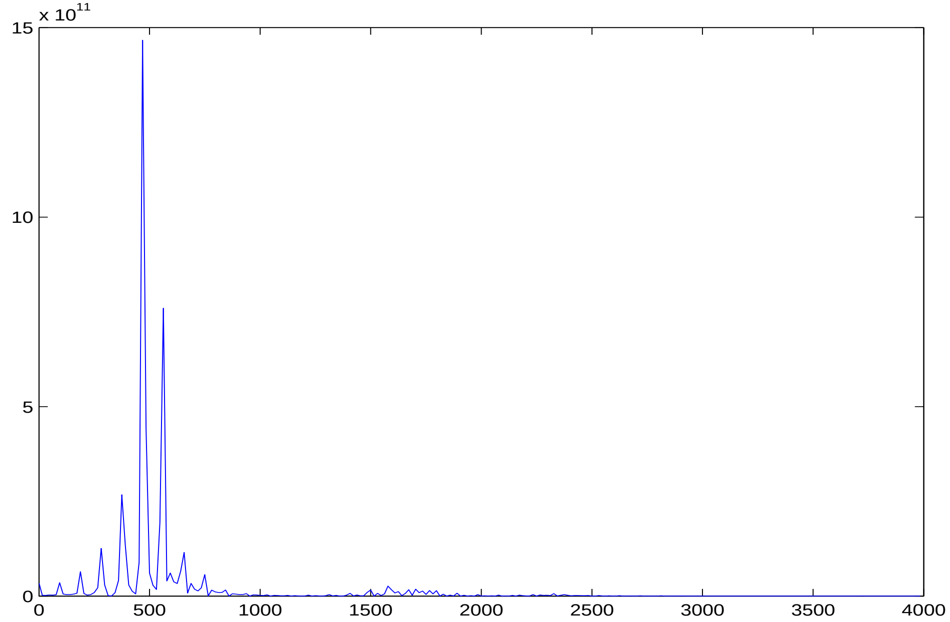
It results in:

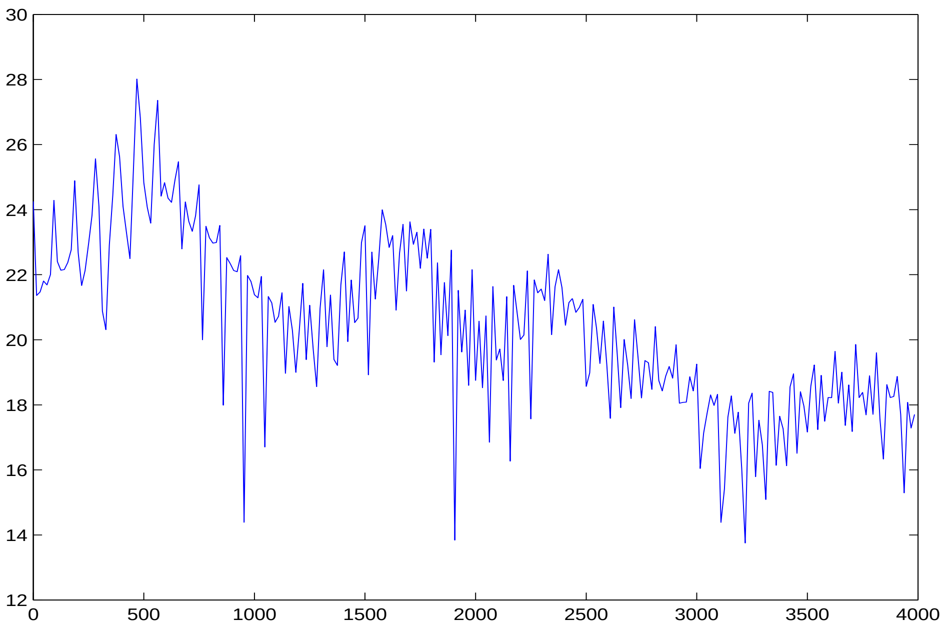
$$\begin{aligned} c(n) &= \mathcal{F}^{-1}\left\{\ln[|E(f)|^2 \; |H(f)|^2]\right\} = \mathcal{F}^{-1}\left\{\ln|E(f)|^2 + \ln|H(f)|^2\right\} = & (26) \\ &= \mathcal{F}^{-1}\left\{\ln|E(f)|^2\right\} + \mathcal{F}^{-1}\left\{\ln|H(f)|^2\right\} = c_e(n) + c_h(n) & (27) \\ & & (28) \end{aligned}$$

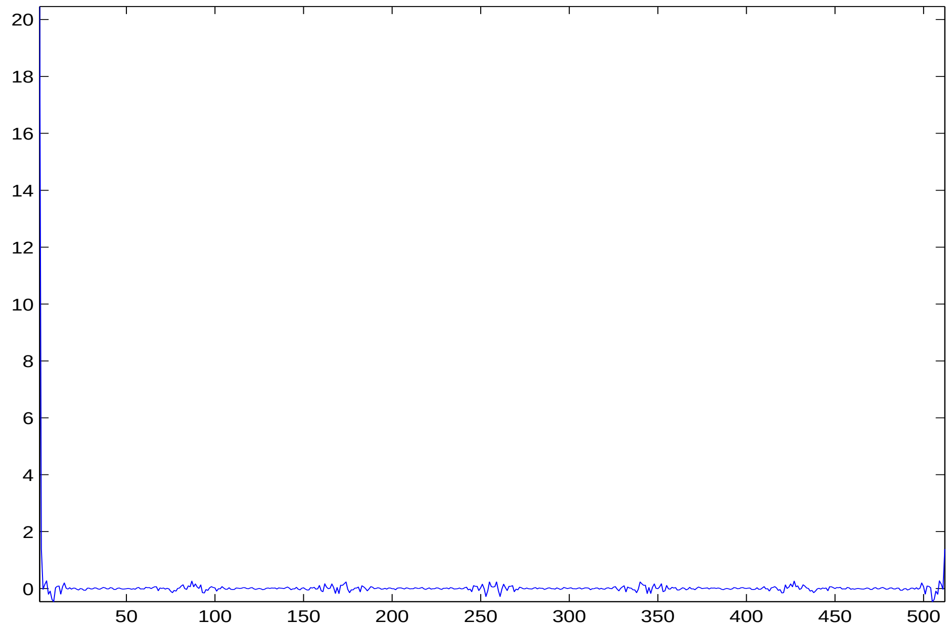
Convolution becames **summation**. The coefficients $c_e(n)$ and $c_h(n)$ are separable in frequency, which allows to separate them by windowing.

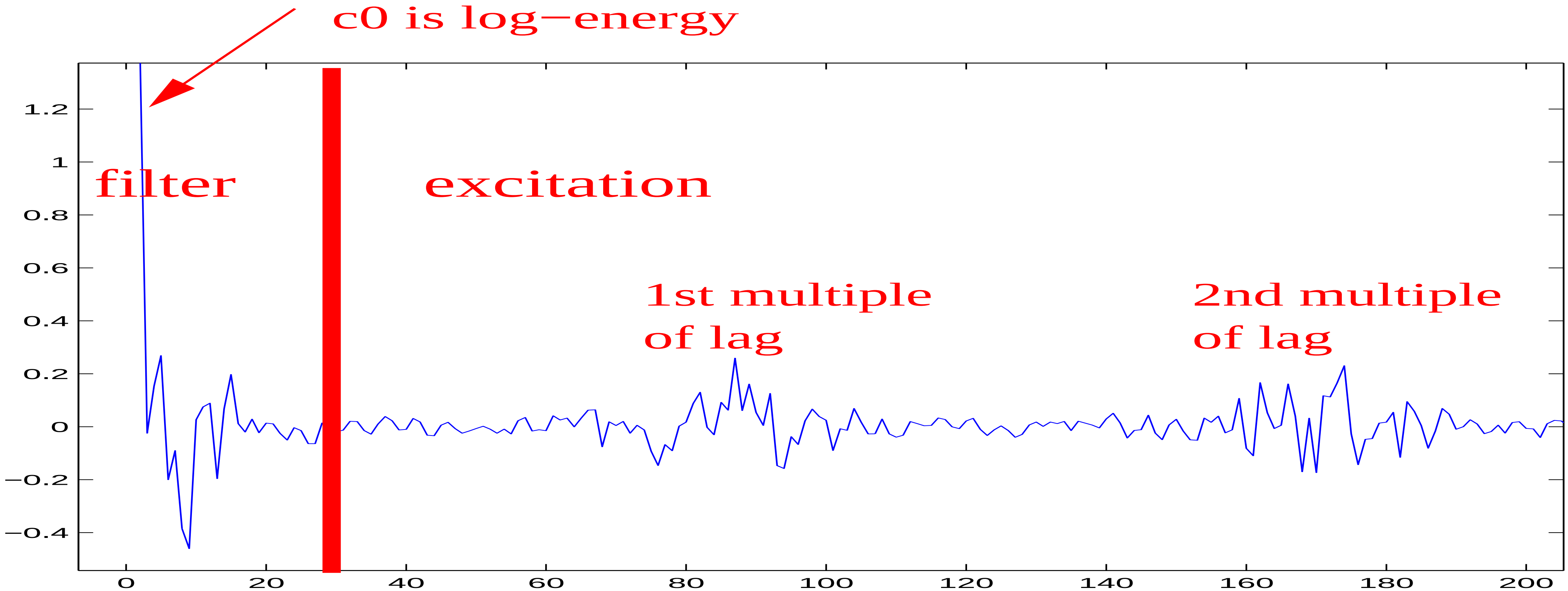
signal, $\quad |\mathcal{F}[s(n)]|^2, \quad \ln|\mathcal{F}[s(n)]|^2, \quad \mathcal{F}^{-1}\left\{\ln|\mathcal{F}[s(n)]|^2\right\}.$







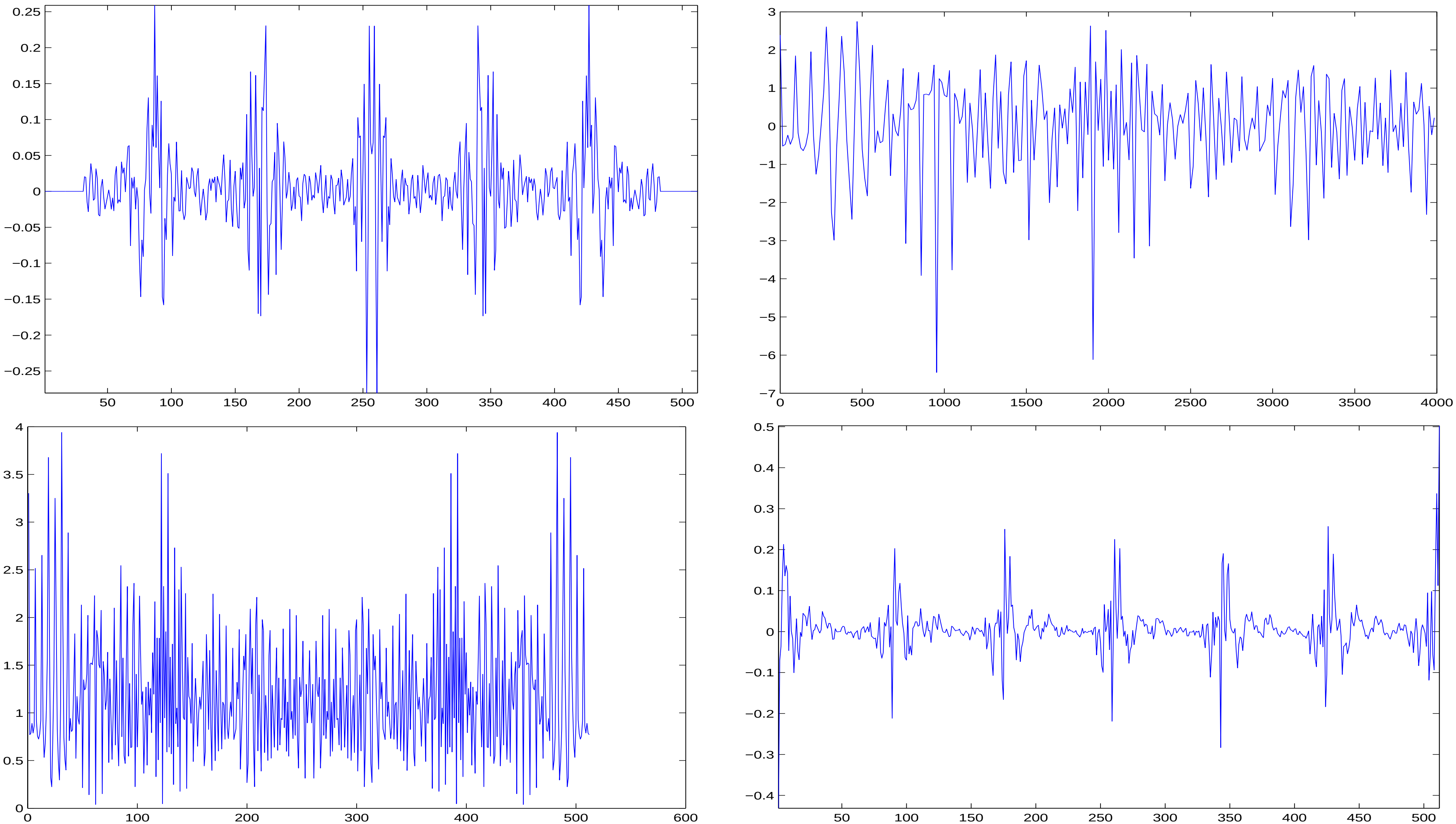




For the sampling frequency F_s =8000 Hz, we can separate excitation from modification in frequency domain using threshold of 30.

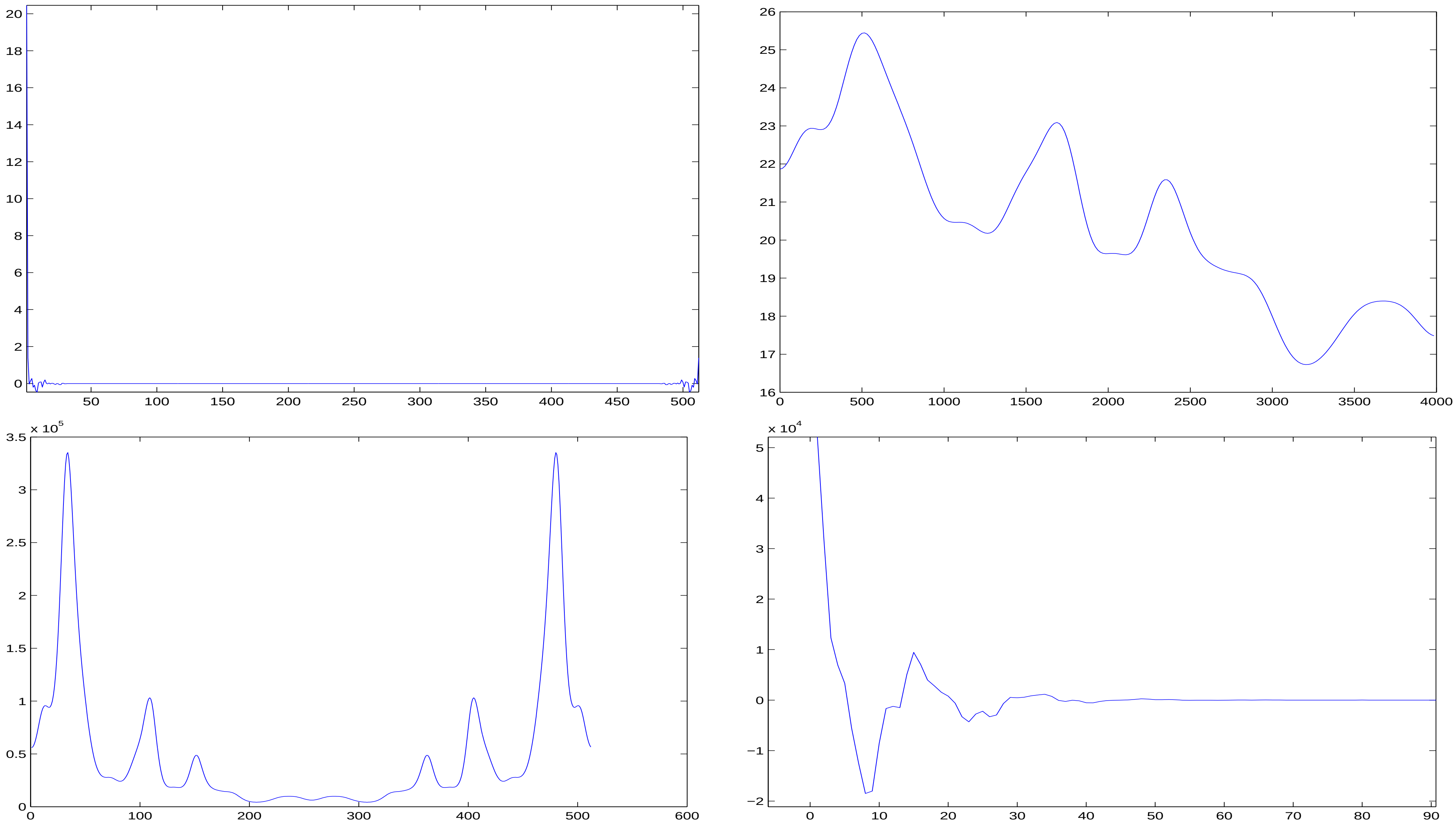
Excitation only – set the cepstra related to modification to zero:

modified cepstrum, $\ln|\mathcal{F}[s(n)]|^2$, $|\mathcal{F}[s(n)]|$, signal (after IDFT, phases of the original signal are used).



Modification only – set the cepstra related to excitation to zero:

modified cepstrum, $\ln|\mathcal{F}[s(n)]|^2$, $|\mathcal{F}[s(n)]|$, signal (after IDFT, phases are set to zero).



Mel-frequency cepstrum – MFCC

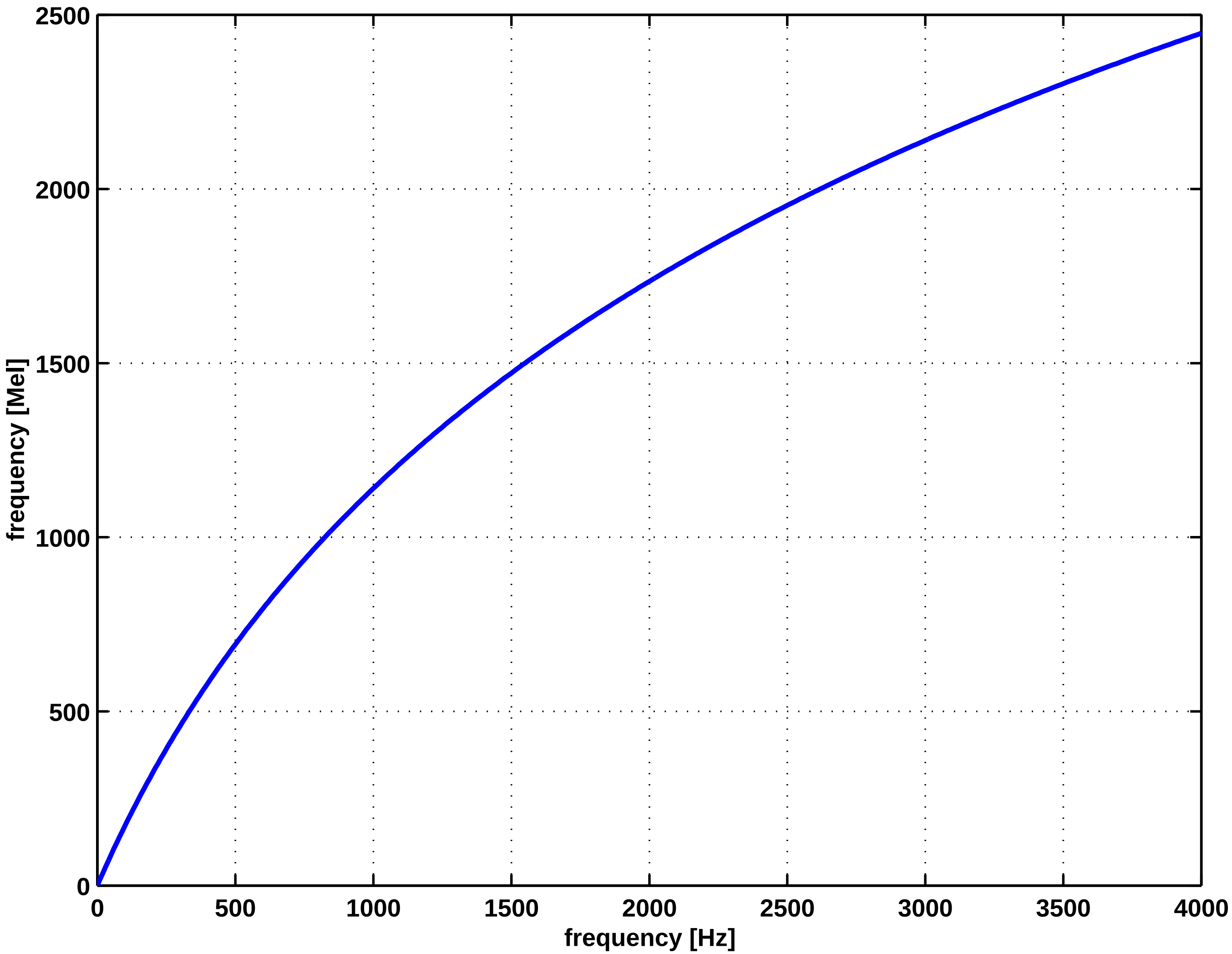
- DFT has equivalent frequency resolution along the axes.
- Human ear has higher resolution on lower frequencies than on higher frequencies.
- We want to adjust cepstrum to human hearing.

how do we do it?

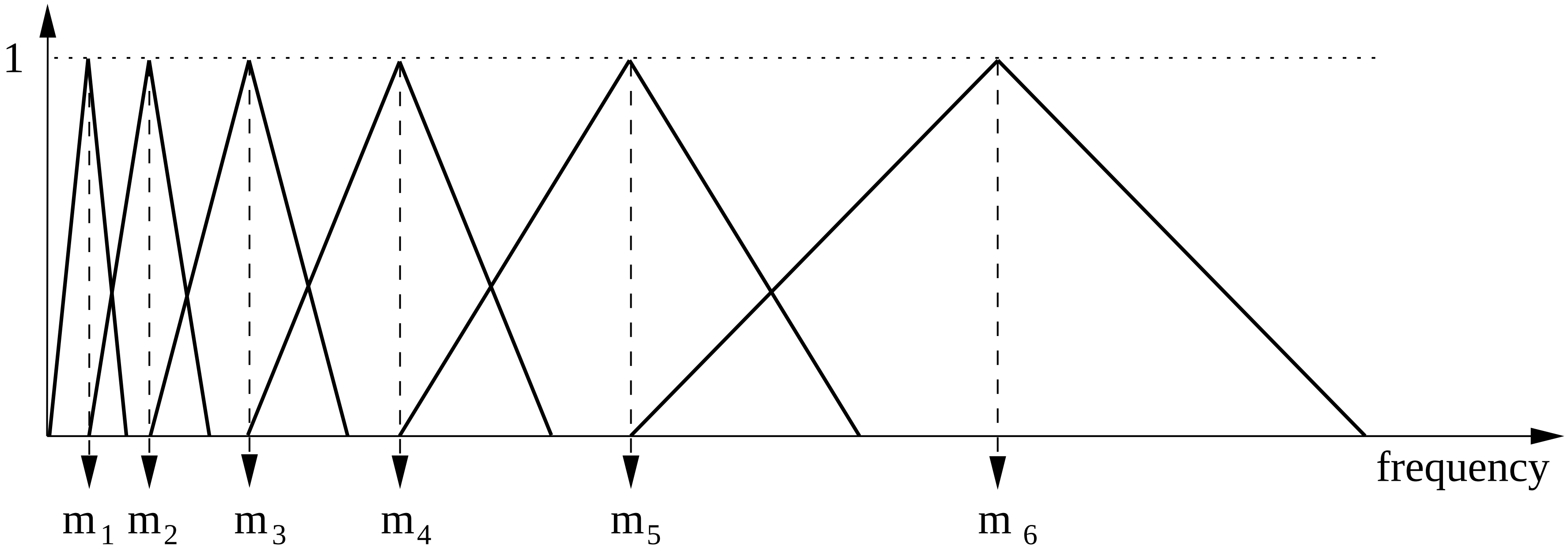
- Place filters along the frequency axes *non-linearly*, calculate the energy on the output, use the calculated values instead of DFT when calculating cepstra.
- Calculate non-linear frequency axes and place the filters on the modified axes linearly, . . .

Convert Hertz to Mel (to transform the frequency axes):

$$F_{Mel} = 2959 \log_{10}(1 + \frac{F_{Hz}}{700}) \tag{29}$$



When linearly placed on the Mel axes, the filters correspond to the non-linearly placed filters on the Hertz axes:



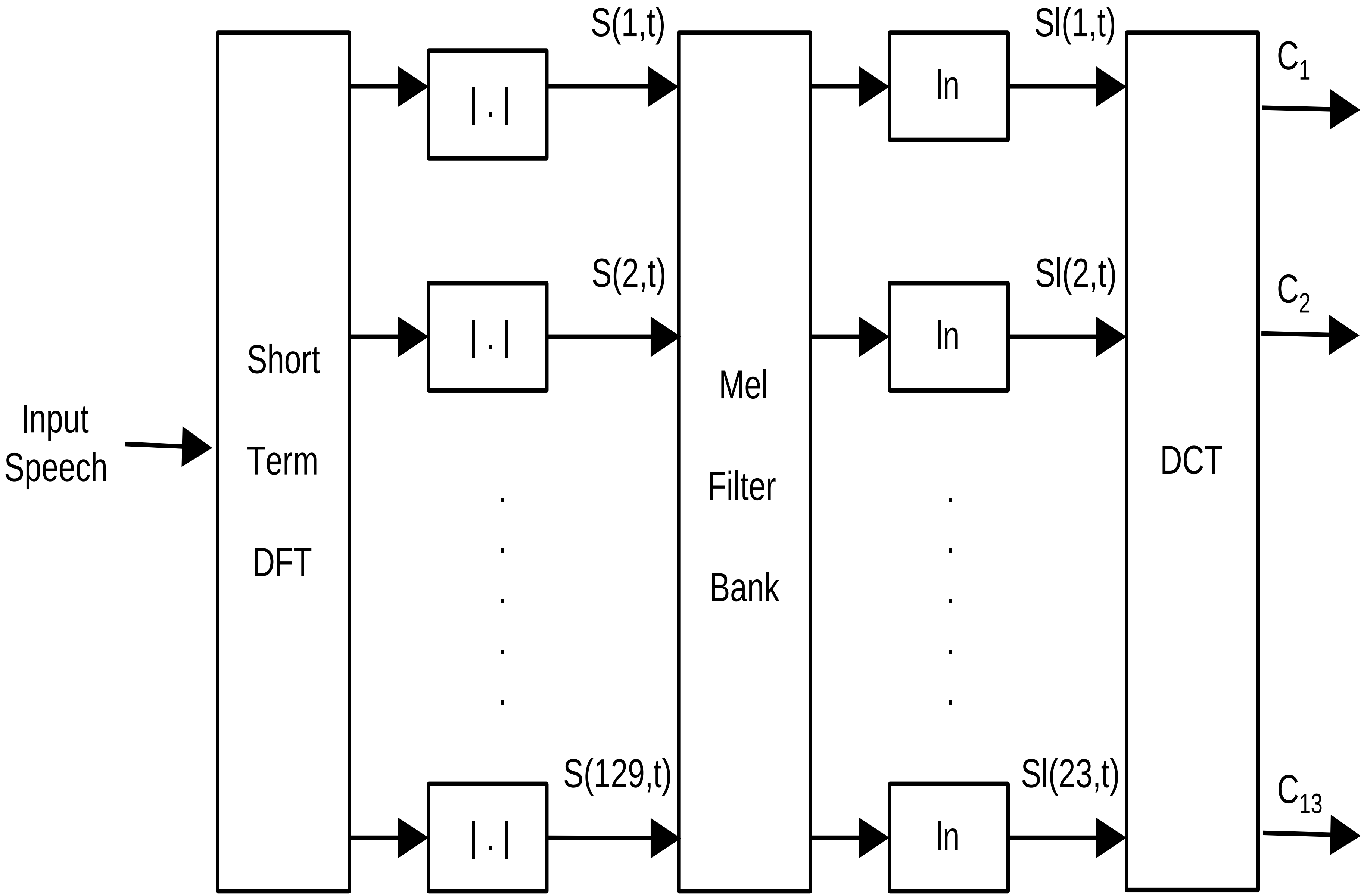
Energy estimation:

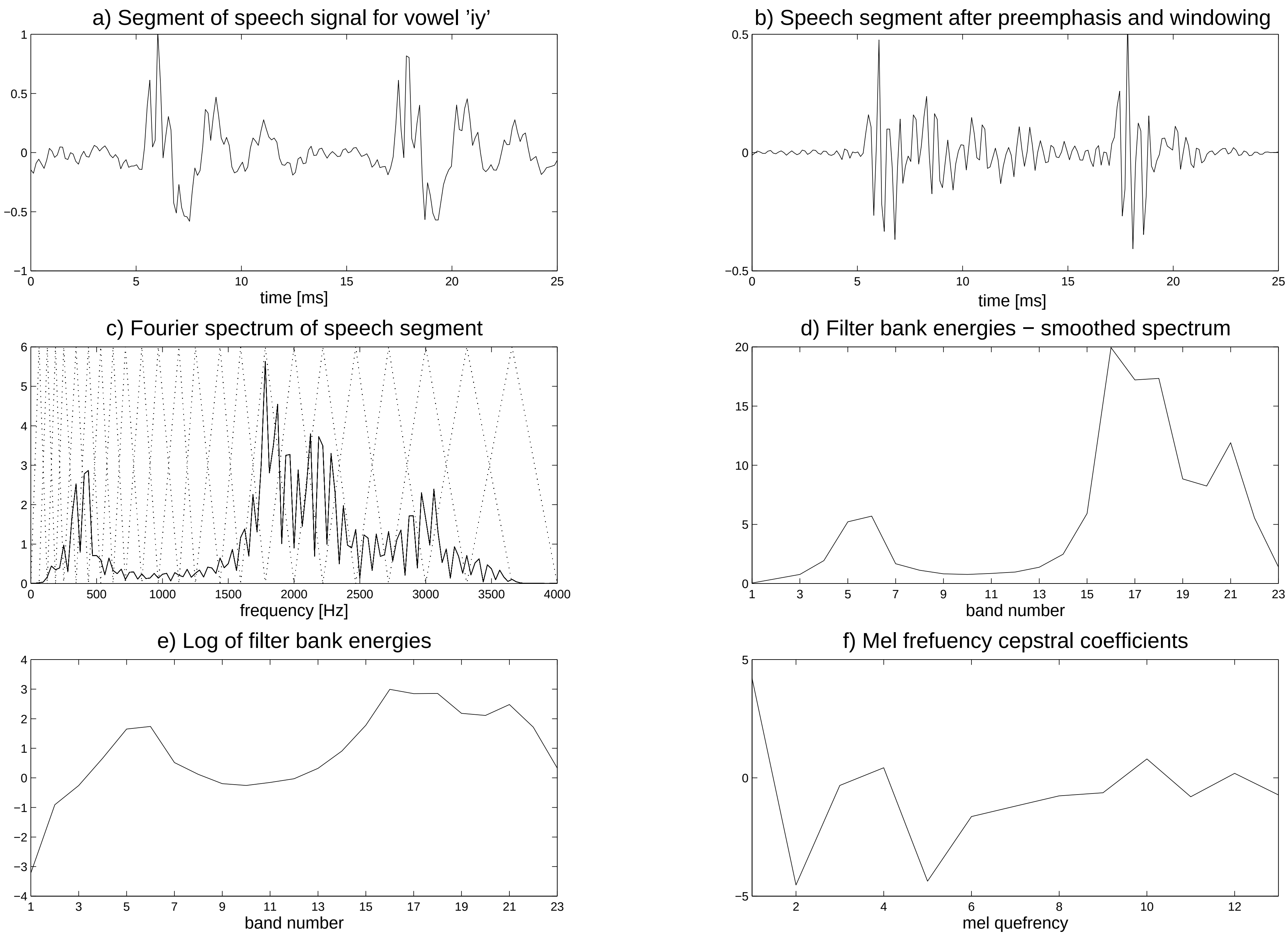
1. construct filter bank, filter the input signal in time domain and calculate energy: $\sum_n s_i^2(n)$ ... TOO DIFFICULT.
2. apply DFT, power, multiply by a triangular window and add up. (used in the HTK toolkit).

Inverse FT can be realized by the discrete cosine transform (DCT) ... (without derivation: makes use of symmetry of the spectrum and the fact that the result should be real):

$$c_{mf}(n) = \sum_{i=1}^{K} \log m_k \cos\left[n(k-0.5)\frac{\pi}{K}\right] \qquad (30)$$

$\Rightarrow$ **Mel-frequency cepstral coefficients (MFCC)**


Input
Speech
Short
Term
DFT
| . |
| . |
| . |
S(1,t)
S(2,t)
S(129,t)
Mel
Filter
Bank
ln
ln
ln
Sl(1,t)
Sl(2,t)
Sl(23,t)
DCT
C1
C2
C13


a) Segment of speech signal for vowel 'iy'
time [ms]
b) Speech segment after preemphasis and windowing
time [ms]
c) Fourier spectrum of speech segment
frequency [Hz]
d) Filter bank energies − smoothed spectrum
band number
e) Log of filter bank energies
band number
f) Mel frefuency cepstral coefficients
mel quefrency