Předzpracování řeči, tvorba řeči, cepstrum

Jan Černocký cernocky@fit.vutbr.cz

FIT VUT Brno

Plán

- Parametrisace řeči
  - Předzpracování
  - základní parametry: krátkodobá energie, průchody nulou.
- Tvorba řeči a její signálový model.
- Spektrogram
- Oddělení buzení a modifikace – cepstrum
- Přiblížení cepstra lidskému slyšení – MFCC.

PARAMETRIZACE

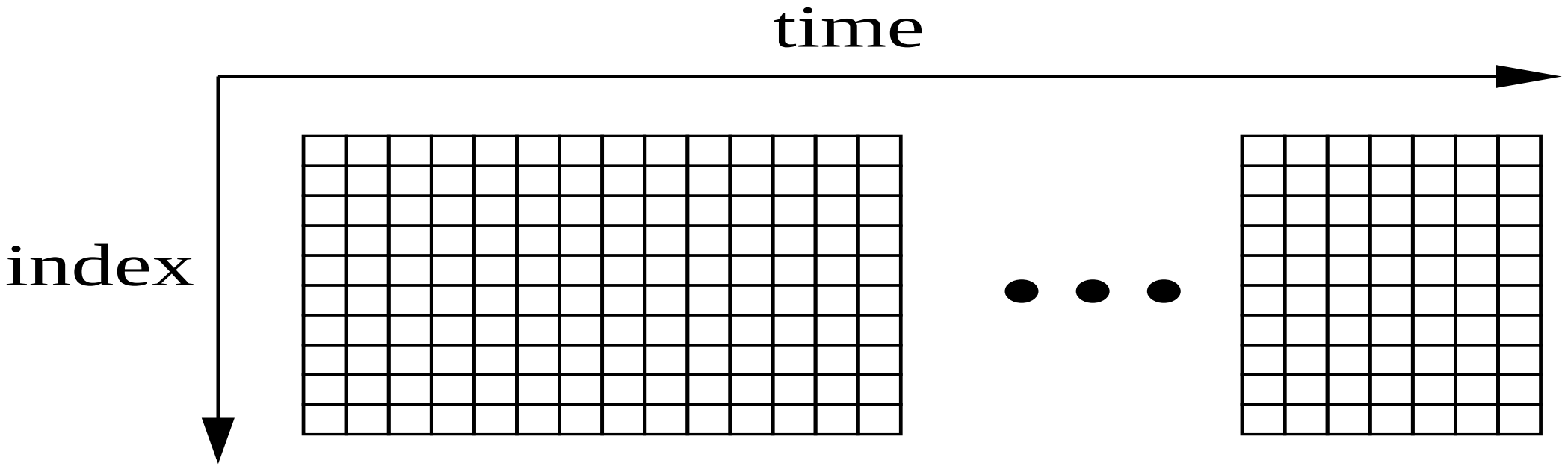
- Úkol: vyjádřit řečový signál omezeným množstvím hodnot – "parametrisace", "feature extraction".
- a) Popis založen pouze na poznatcích o zpracování signálu (banky filtrů, Fourierova transformace, atd.) $\Rightarrow$ *neparametrický popis*.
- b) Popis založen na poznatcích o tvorbě řeči $\Rightarrow$ *parametrický popis*.

ALE:

- b) používá mnoho technik neparametrického popisu, takže tyto dvě skupiny není snadné (a někdy ani žádoucí) oddělit.
- Vypočtené hodnoty stejně vždy nazýváme *parametry*.

Parametry

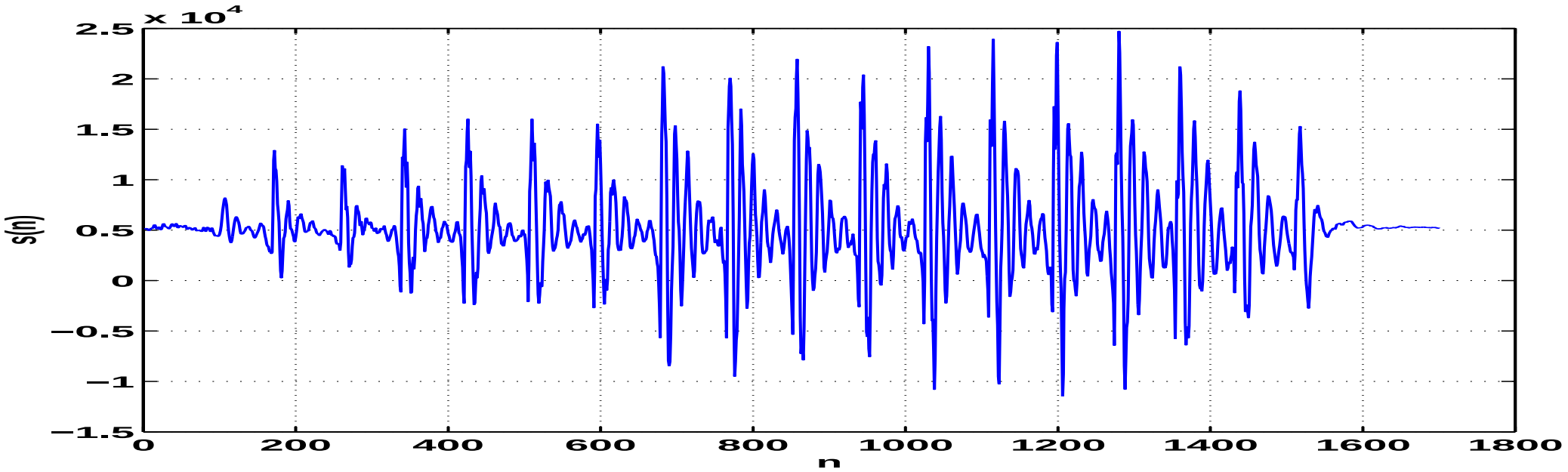
- **skalární** – jedno číslo na řečový úsek (krátkodobá energie nebo počet průchodů nulou).
- **vektorové** – sada čísel (vektor) na řečový úsek. Pokud více řečových úseků, řadíme vektorové hodnoty do *matic*.



PŘEDZPRACOVÁNÍ (pre-processing)

Ustřednění

Stejnosměrná složka (dc-offset) – nenese žádnou užitečnou informaci, naopak může být pro další zpracování rušivá (např. výpočet energie).



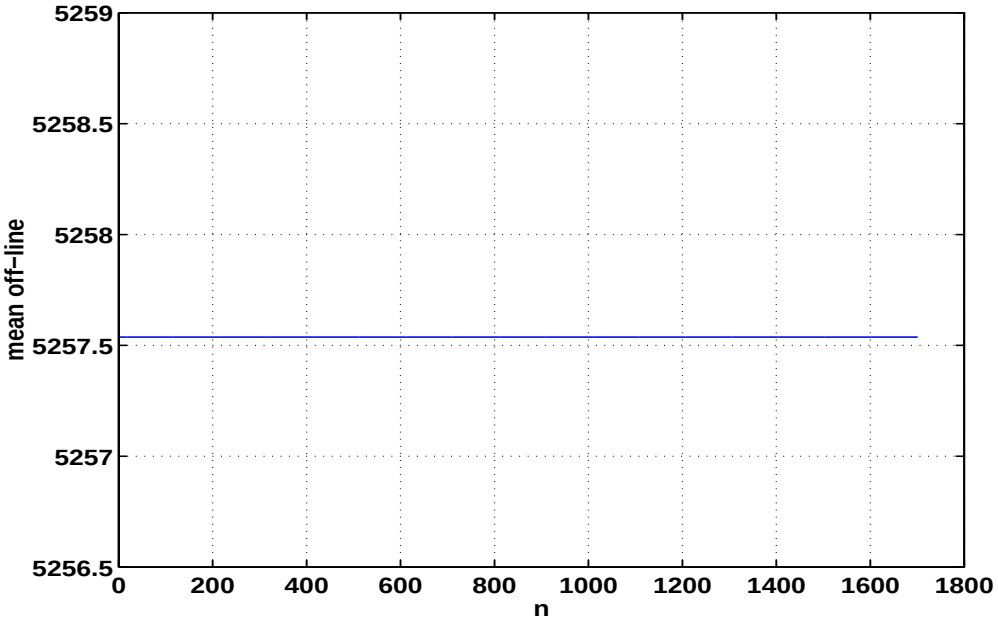
PRYČ S NÍ ! (dc-offset removal)

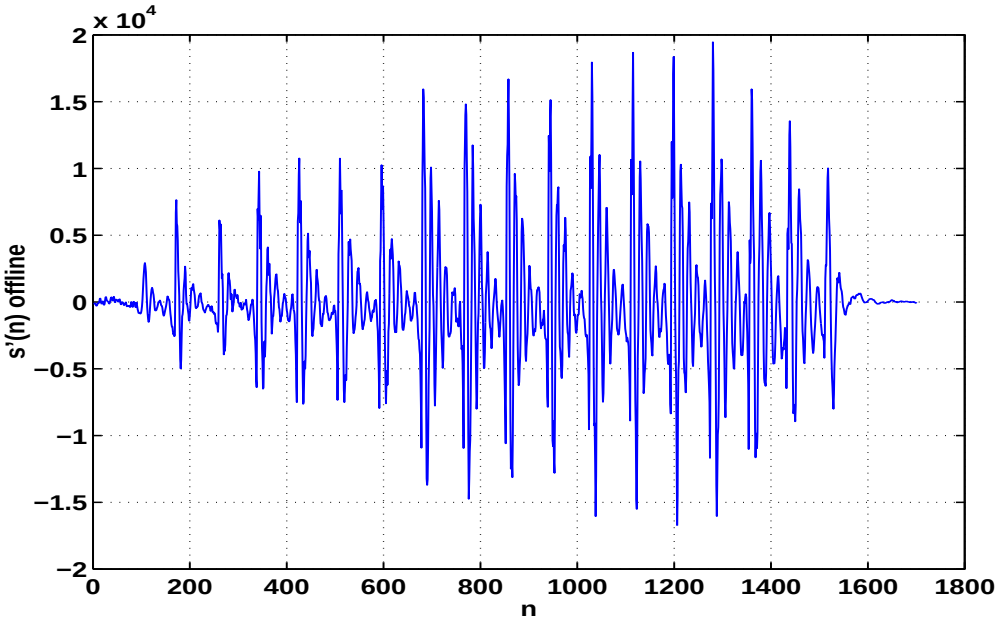
$$s'[n] = s[n] - \mu_s, \qquad \mu_s \text{ musíme odhadnout.} \tag{1}$$

Střední hodnota off-line

v pohodě — každý umí spočítat průměr:

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





Střední hodnota on-line

Nemáme k disposici celý signál: je příliš dlouhý, nebo neustále "přibývá".

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

kde $\gamma \longrightarrow 1$. To je ekvivalentní filtraci signálu filtrem s impulsní odezvou:

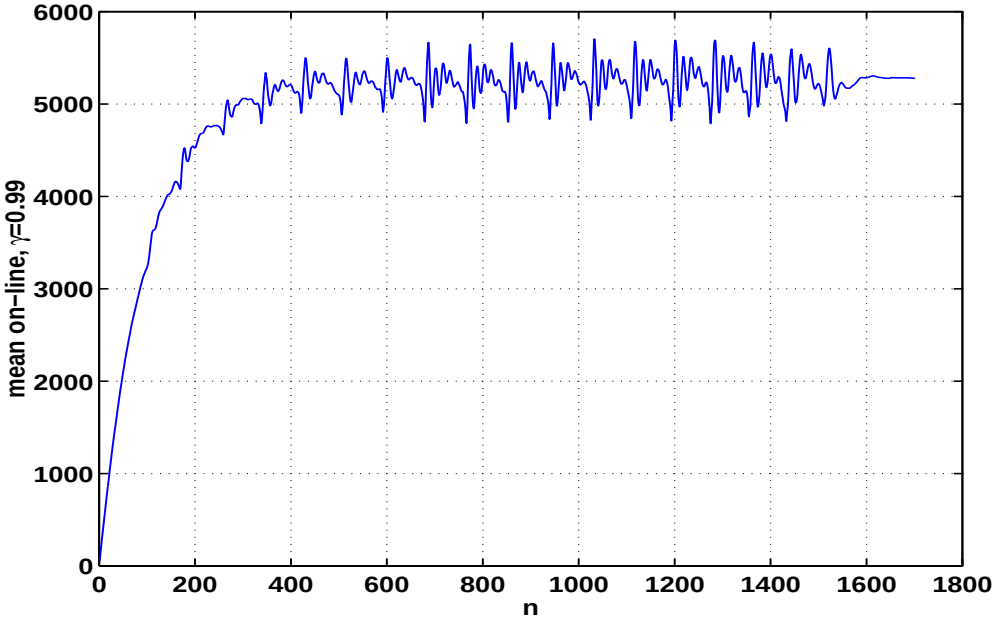
$$h = [(1-\gamma) \quad (1-\gamma)\gamma \quad (1-\gamma)\gamma^2 \quad \ldots]. \tag{4}$$

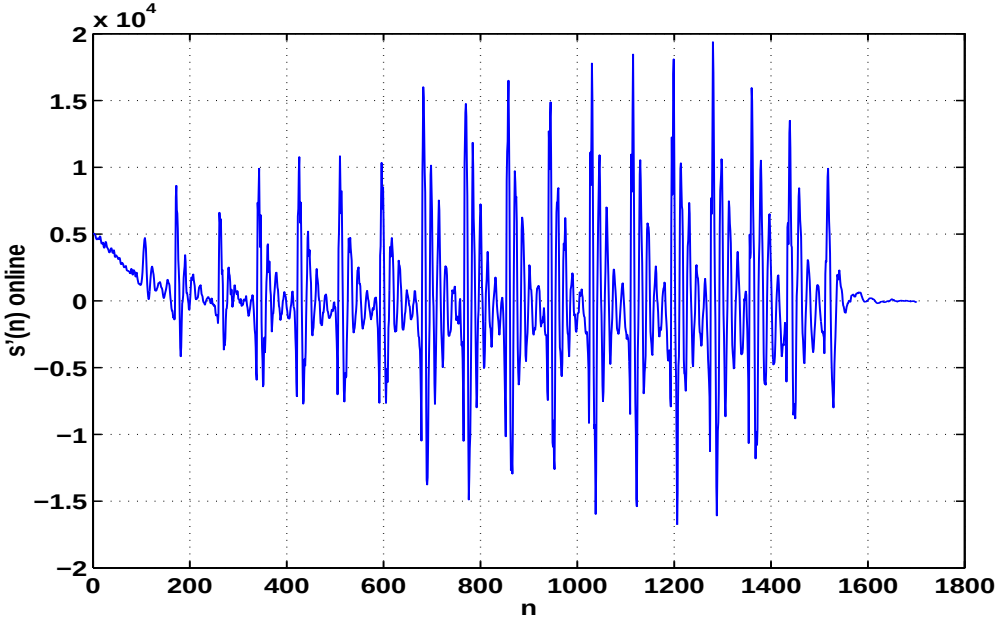
Je to geometrická posloupnost: počáteční člen $a_0 = 1-\gamma$ a kvocient $q = \gamma$. Její součet je tedy:

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

(to jsme u výrazu počítajícího střední hodnotu očekávali ☺).

Příklad pro $\gamma = 0.99$ (viz také první počítačová labina):





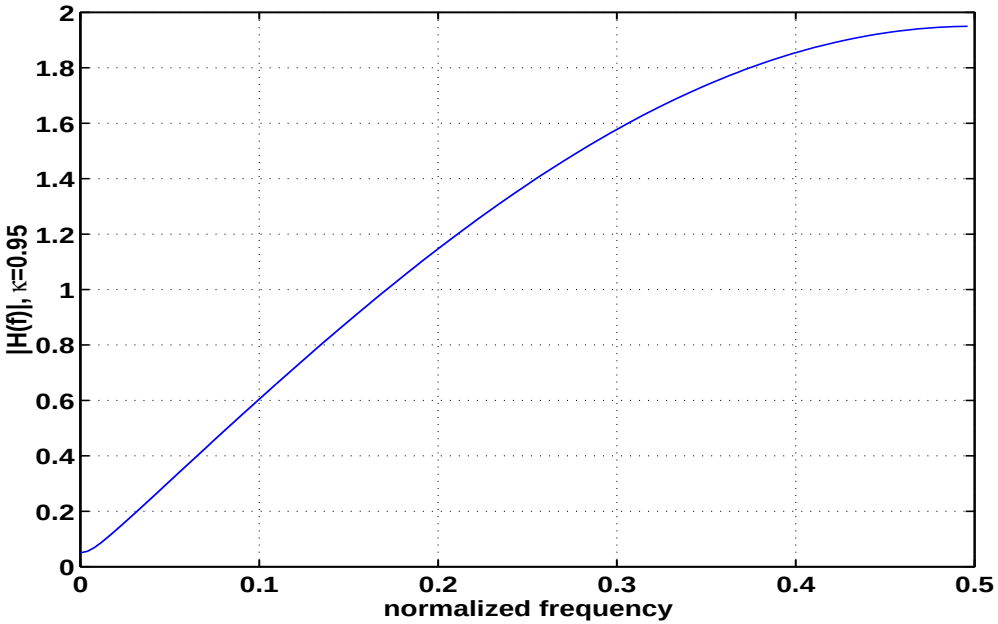
Preemfáze

Vyrovnání kmitočtové charakteristiky řeči (energie klesá směrem k vyšším frekvencím). Spíše historická operace.

Jednoduchý FIR filtr prvního řádu:

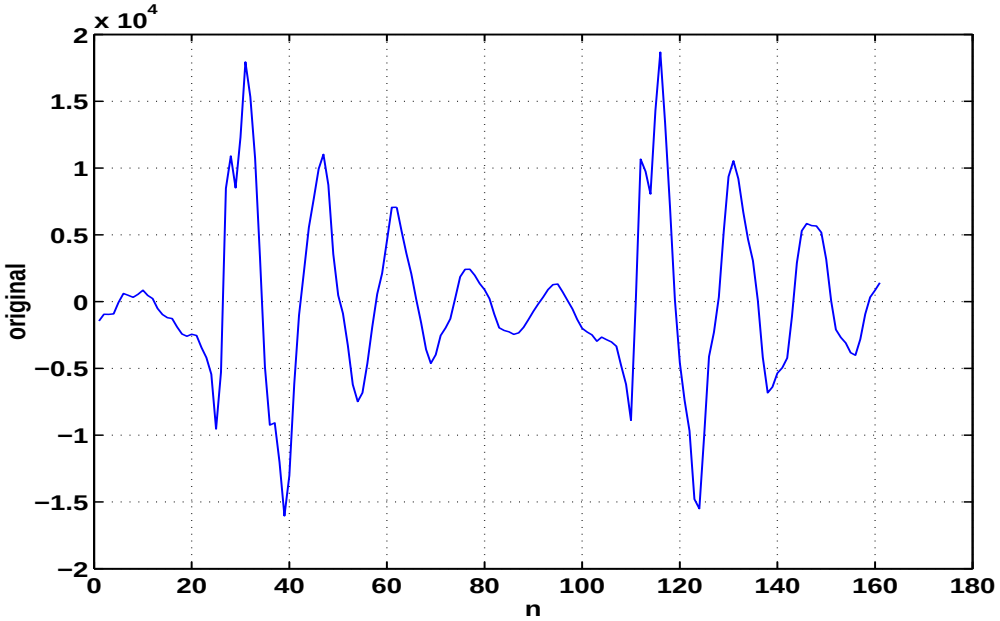
$$H(z) = 1 - \kappa z^{-1}, \tag{6}$$

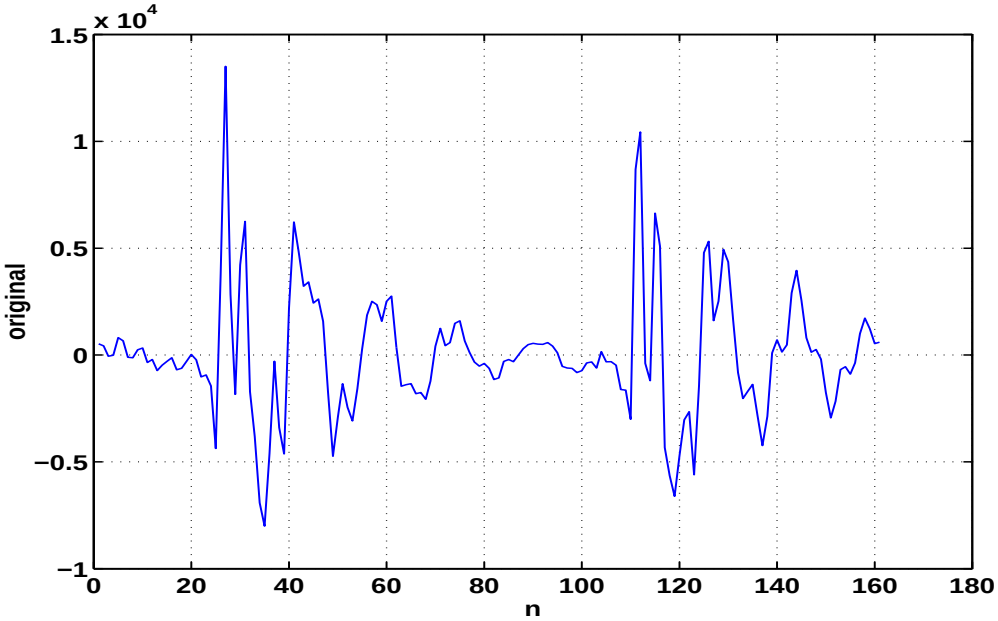
kde $\kappa \in [0.9,\ 1]$. Tedy něco jako rozdíl dvou sousedních vzorků. Modulová frekvenční charakteristika pro κ=0.95:



Filtrace tímto filtrem:

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





$\Rightarrow$ Preemfázovaný průběh je “kostrbatější” a má více “ostrých hran” – více vyšších frekvencí.

RÁMCE

- Proč ?
- Řečový signál považujeme za *náhodný*, pro metody odhadu parametrů by měl být *stacionární*.
- Jenže není $\Rightarrow$ dělení na kratší úseky (segmenty, mikrosegmenty, frames). Tam stacionární bude nebo budeme doufat, že bude. . . .
- Parametry rámců: délka (length) l_{ram}, překrytí (overlap) p_{ram}, posun rámce (frame shift) $s_{ram} = l_{ram} - p_{ram}$.

Délka

1. dostatečně *malá*, aby bylo možné pokládat signál na daném úseku za stacionární.
2. ALE: dostatečně *velká*, aby bylo možné dostatečně přesně odhadnout požadované parametry.

$\Rightarrow$ kompromis (setrvačnost hlasového ústrojí), typická délka 20–25 ms (160–200 vzorků pro F_s =8000 Hz).

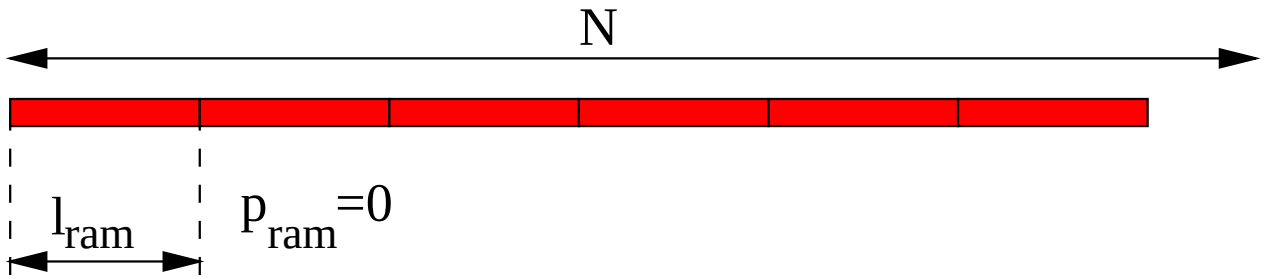
Překrytí

- **malé nebo žádné:** ☺ rychlý časový posun v signálu, malé nároky na paměť/procesor, ☹ hodnoty parametrů se od jednoho rámce ke druhému mohou hodně měnit.
- **velké:** ☺ pomalý časový posuv, "vyhlazené" průběhy parametrů, ☹ velké nároky na paměť/procesor, velmi podobné parametry (porušují požadavek nezávislosti!).

$\Rightarrow$ kompromis, typická délka 10 ms, tedy 100 rámců za vteřinu, centi-second vectors.

Kolik rámců pro řečový signál o délce N ?

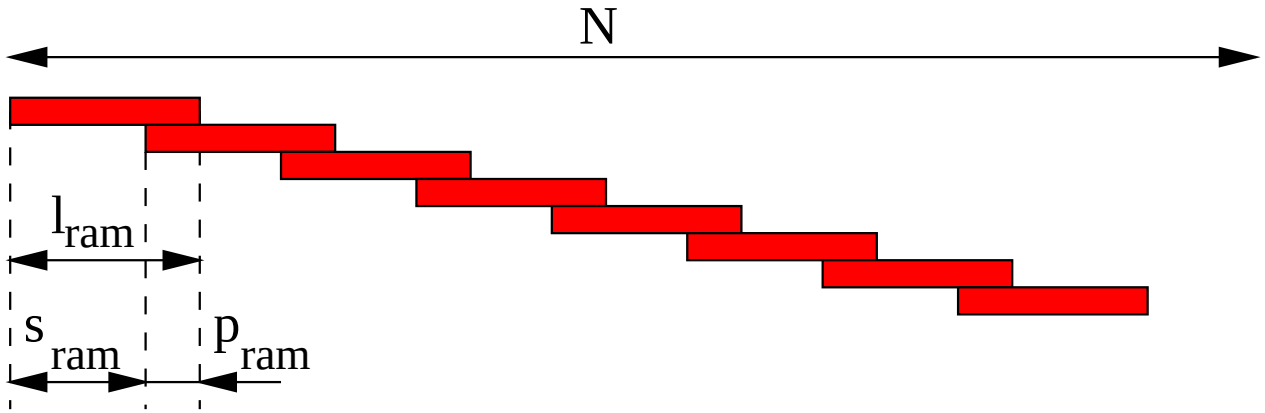
Rámce bez překrývání, $p_{ram} = 0$



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

. . . operace $\lfloor \cdot \rfloor$ značí zaokrouhlování dolů (floor).

Rámce s překrýváním, $p_{ram} \neq 0$



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

. . . pokud je signál alespoň jeden rámec dlouhý.

Výběr signálu do rámců - okénkové funkce

Při "vykrojení" rámce je použito "okno" - window(ing) function:

Pravoúhlé (rectangular) – se signálem neudělá nic:

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

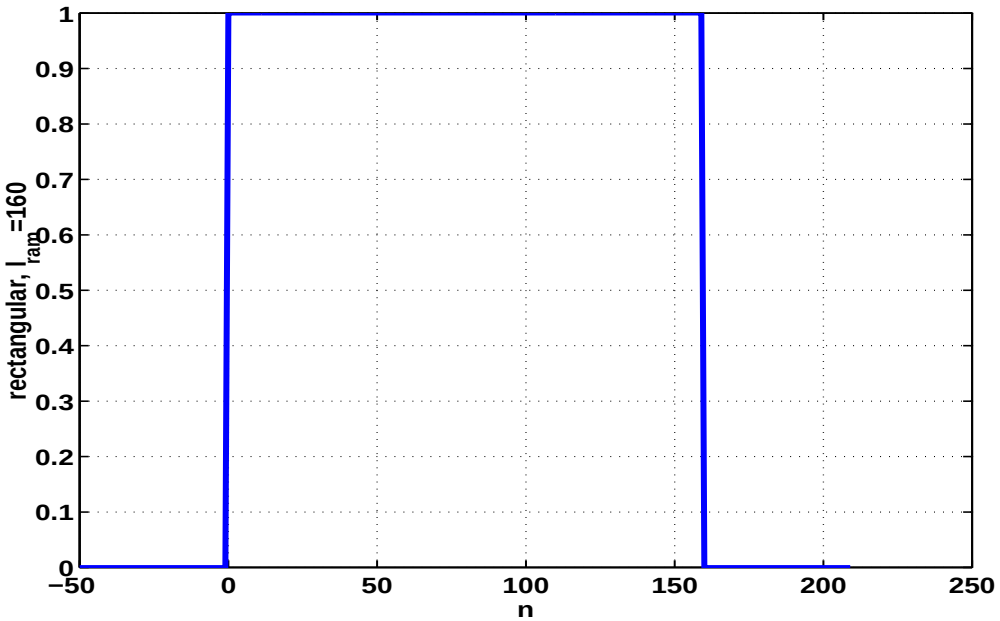
Hammingovo – utlumení signálu na okrajích:

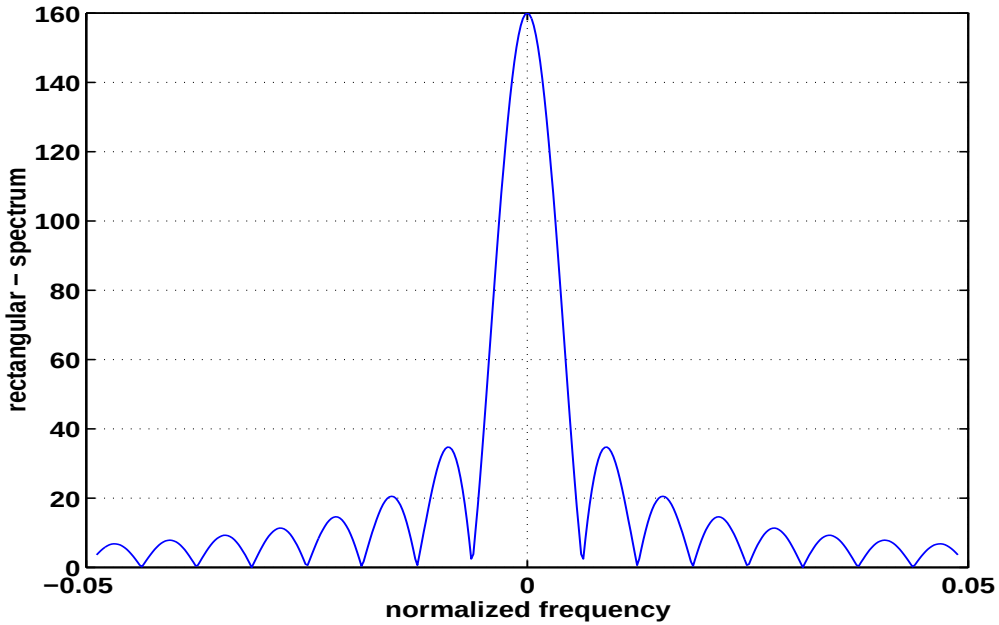
$$w[n] = \begin{cases} 0.54 - 0.46 \cos \dfrac{2\pi n}{l_{ram} - 1} & \text{pro} \quad 0 \leq n \leq l_{ram} - 1 \\ 0 & \text{jinde} \end{cases} \tag{11}$$

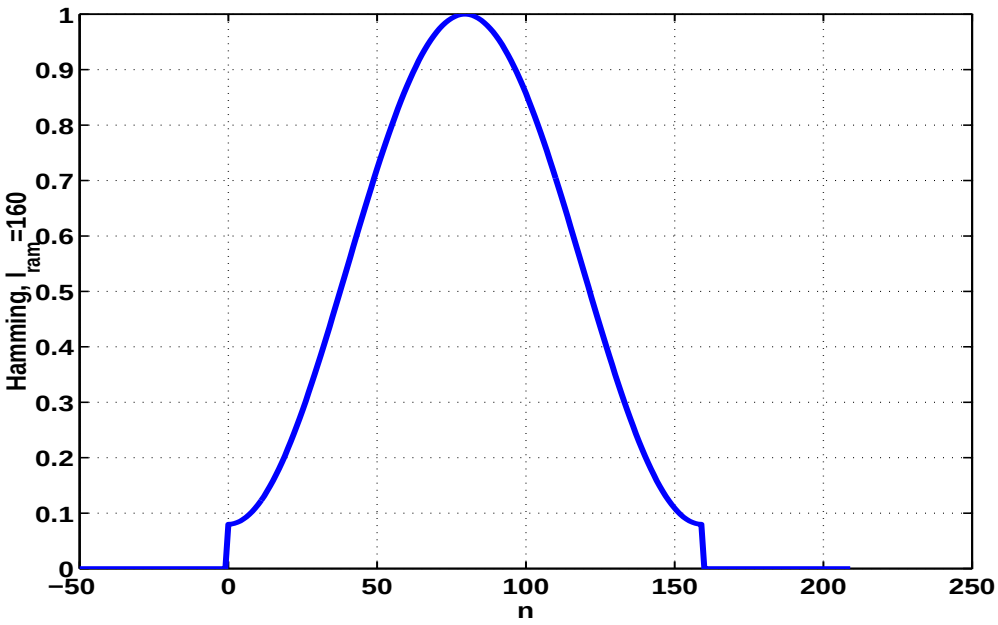
Jak změní okno spektrum vykusovaného signálu ? Násobení signálu oknem v časové oblasti odpovídá *konvoluce* spektra řeči se spektrem okna:

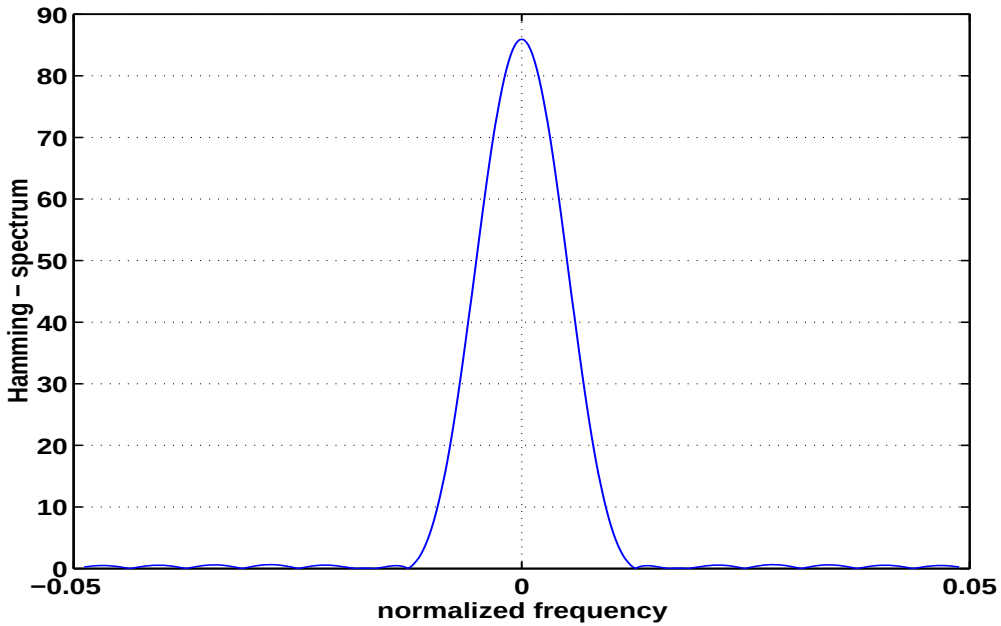
$$X(f) = S(f) \star W(f) \tag{12}$$

Srovnání pravoúhlého a Hammingova:






Hamming, l_ram=160
n
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


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

ZÁKLADNÍ PARAMETRY ŘEČOVÉHO SIGNÁLU

všechny budeme odvozovat na jednotlivých rámcích. Pro všechny rámce:

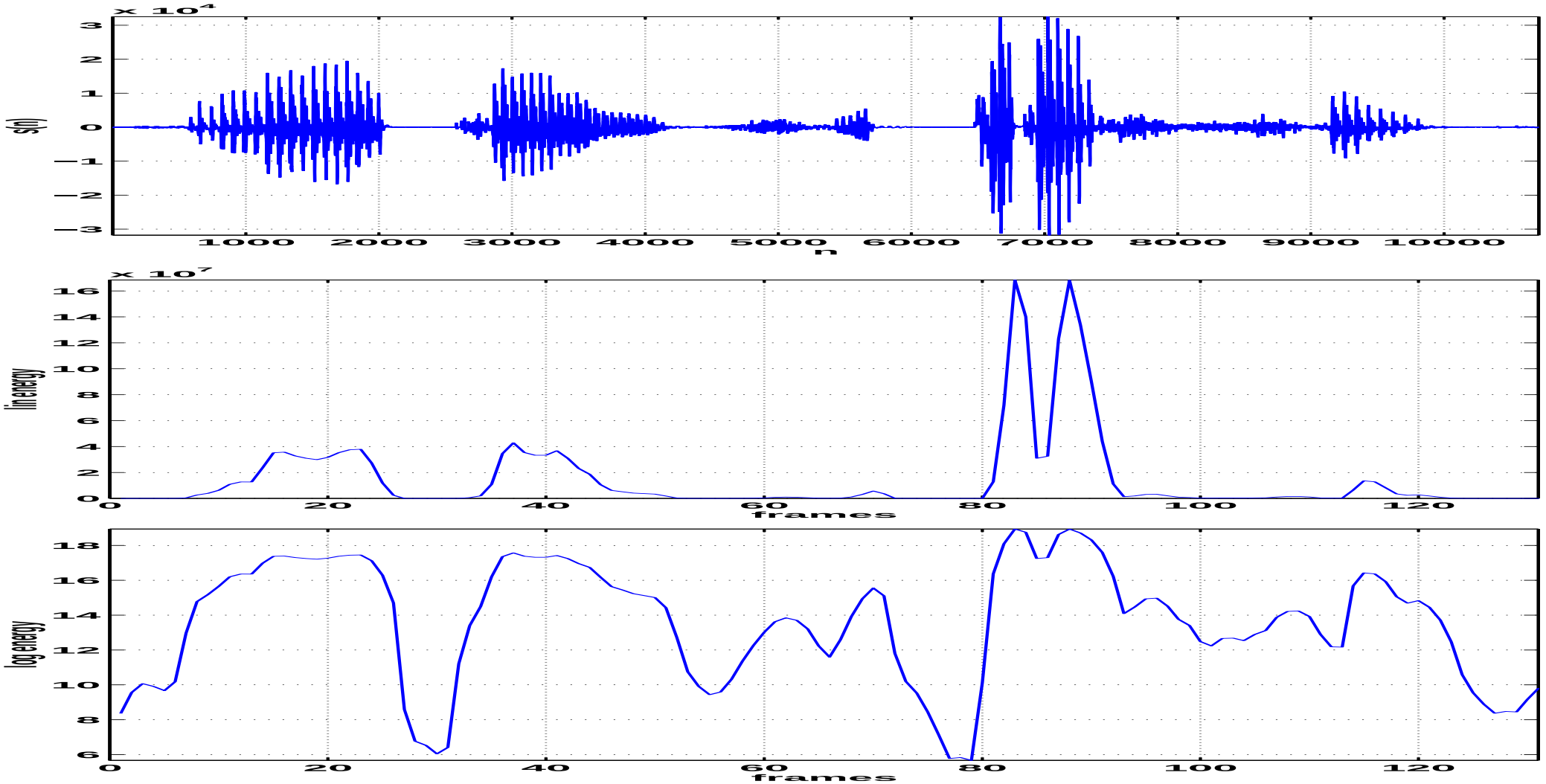
- skalární $\Rightarrow$ jeden řádkový vektor.
- vektorové $\Rightarrow$ matice, svisle index, vodorovně (hrubý) čas.

Střední krátkodobá energie

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

- detektor řečové aktivity.
- rozlišení hlásek na znělé (vysoká energie) a neznělé (nízká).
- často log-energie.
- pozor na šum a na nízkoenergetické hlásky.

Příklad: “létající prase”



Počet průchodů nulou (zero-crossing rate)

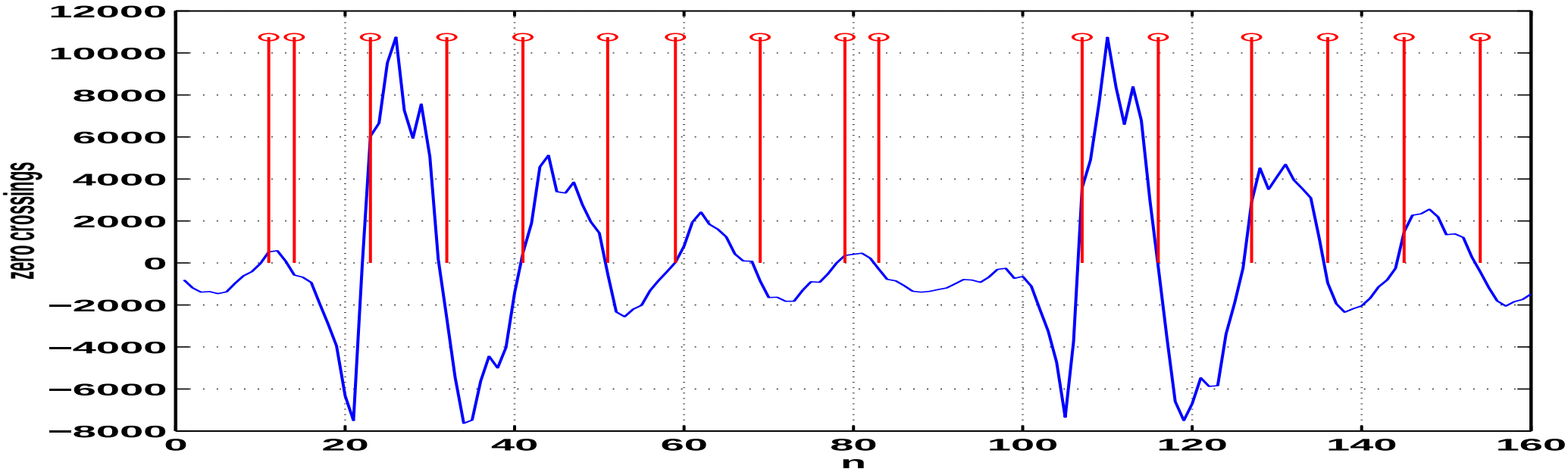
. . . kolikrát za rámec proleze signál nulou.

$$Z = \frac{1}{2} \sum_{n=1}^{l_{ram}-1} |\,\mathrm{sign}\, x[n] - \,\mathrm{sign}\, x[n-1]|, \tag{14}$$

kde $\mathrm{sign}\,(x)$ je zjednodušená znaménková funkce:

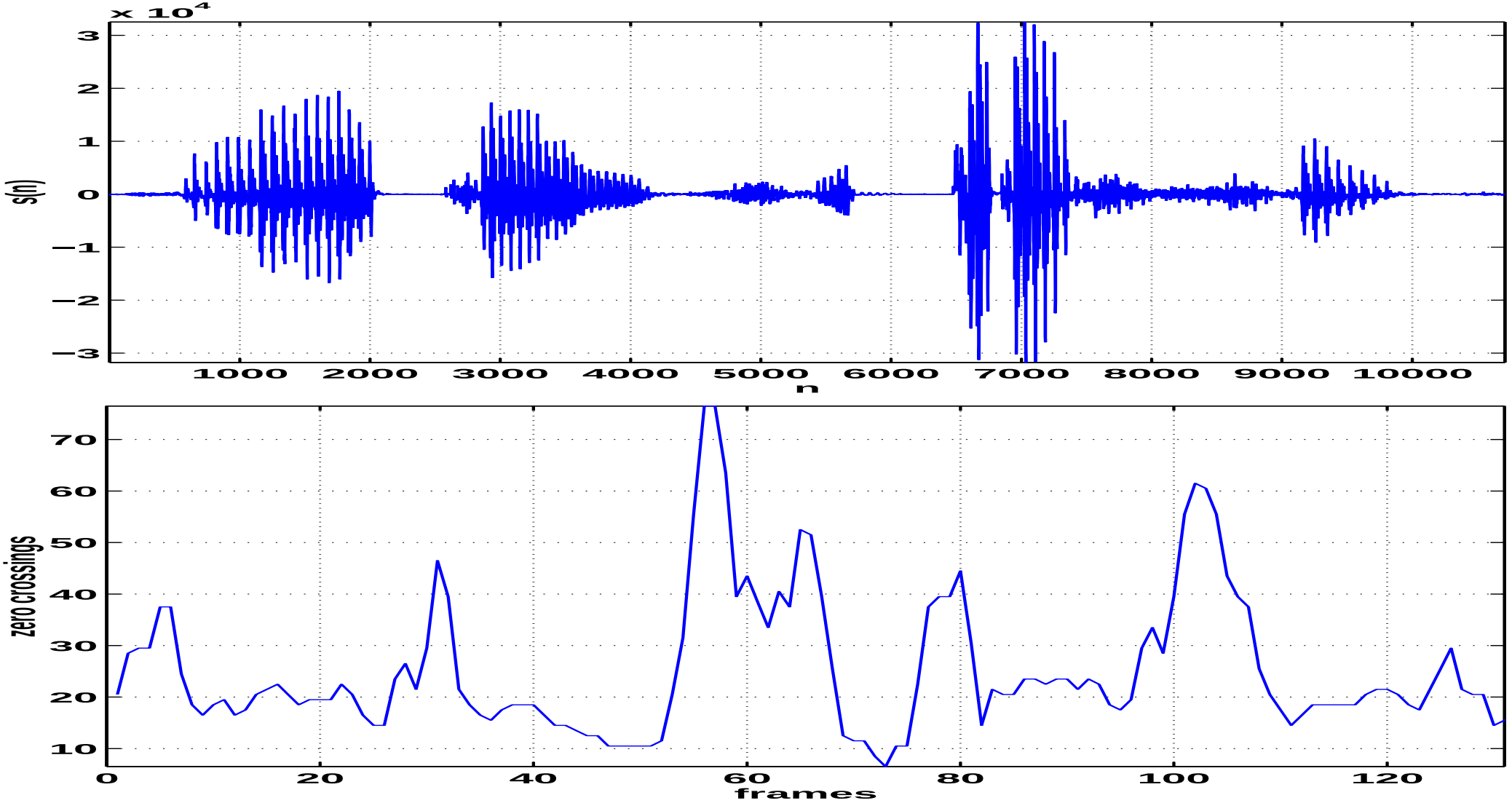
$$\mathrm{sign}\, x[n] = \begin{cases} +1 & \text{pro} \quad x[n] \geq 0 \\ -1 & \text{pro} \quad x[n] < 0 \end{cases} \tag{15}$$

Jak to funguje ? Funkce $|\,\mathrm{sign}\, x[n] - \,\mathrm{sign}\, x[n-1]|$ dává hodnotu 2 pro každý případ, že se od vzorku $x[n-1]$ ke vzorku $x[n]$ změní znaménko:



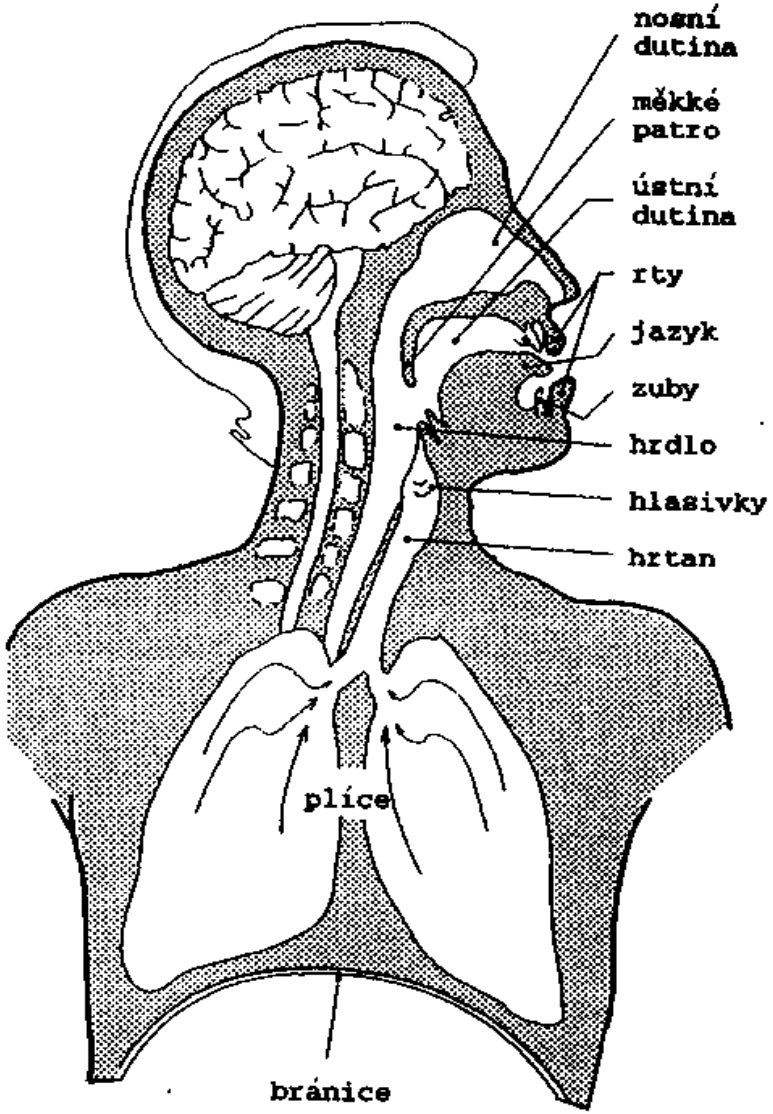
- rozlišení hlásek na znělé (málo průchodů) a neznělé (více jako šum – více průchodů).
- velmi citlivé na šum...

Příklad: “létající prase”



HLASOVÉ ÚSTROJÍ ČLOVĚKA A JEHO MODEL

S laskavým svolením autora převzato z J. Psutka: Komunikace s počítačem mluvenou řečí, Academia Praha 1995.



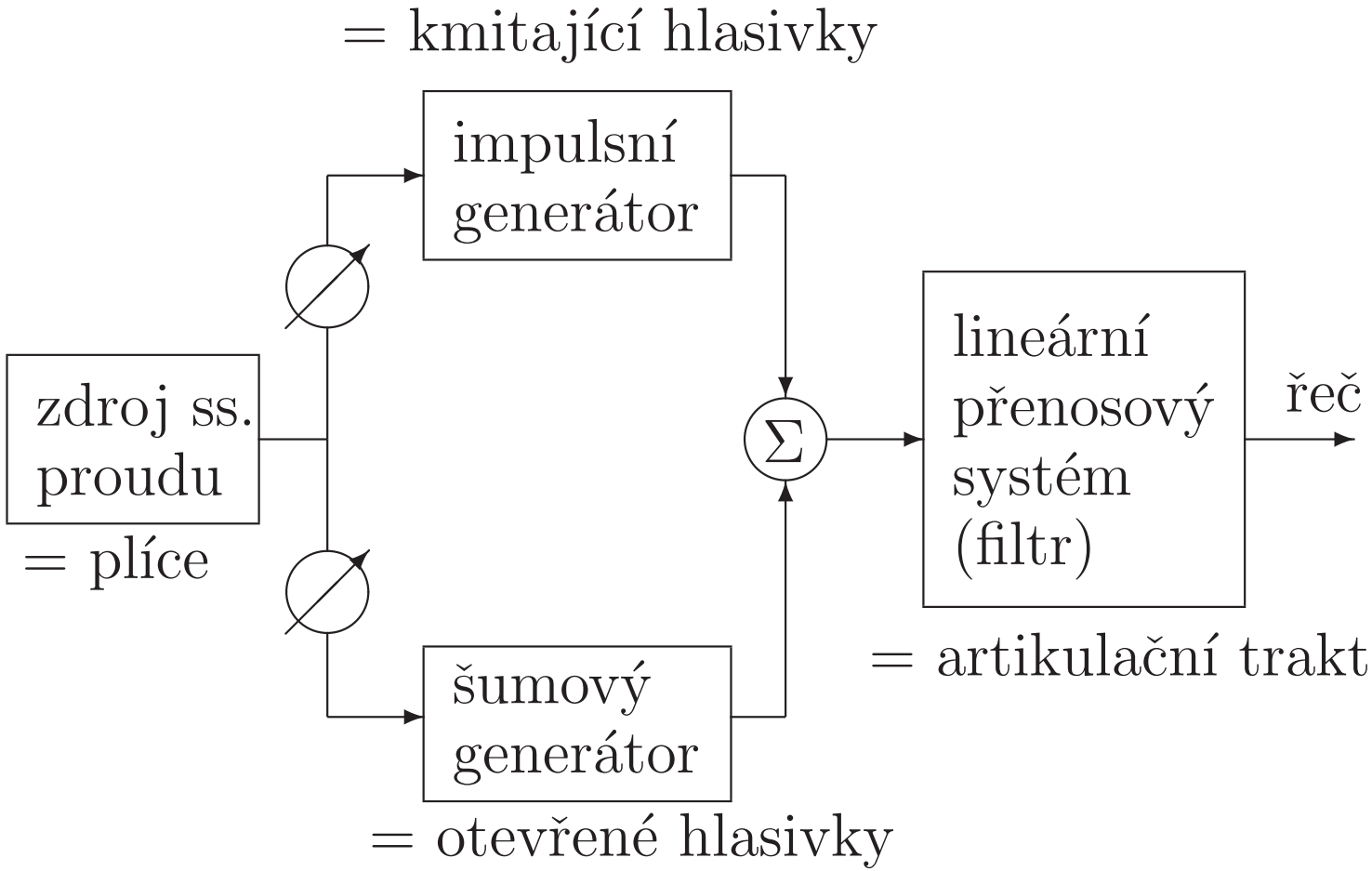
"Maso" a jeho modelování v číslicovém zpracování

- **plíce** — zdroj energie — **signály: NIC**
- **hrtan** (larynx) — modulace energie – **signály: buzení (excitation)**.
  - otevřené hlasivky — šum.
  - kmitající hlasivky — periodický signál (tón). **základní tón řeči**:

muži	90–120 Hz
ženy	150–300 Hz
děti	350–400 Hz

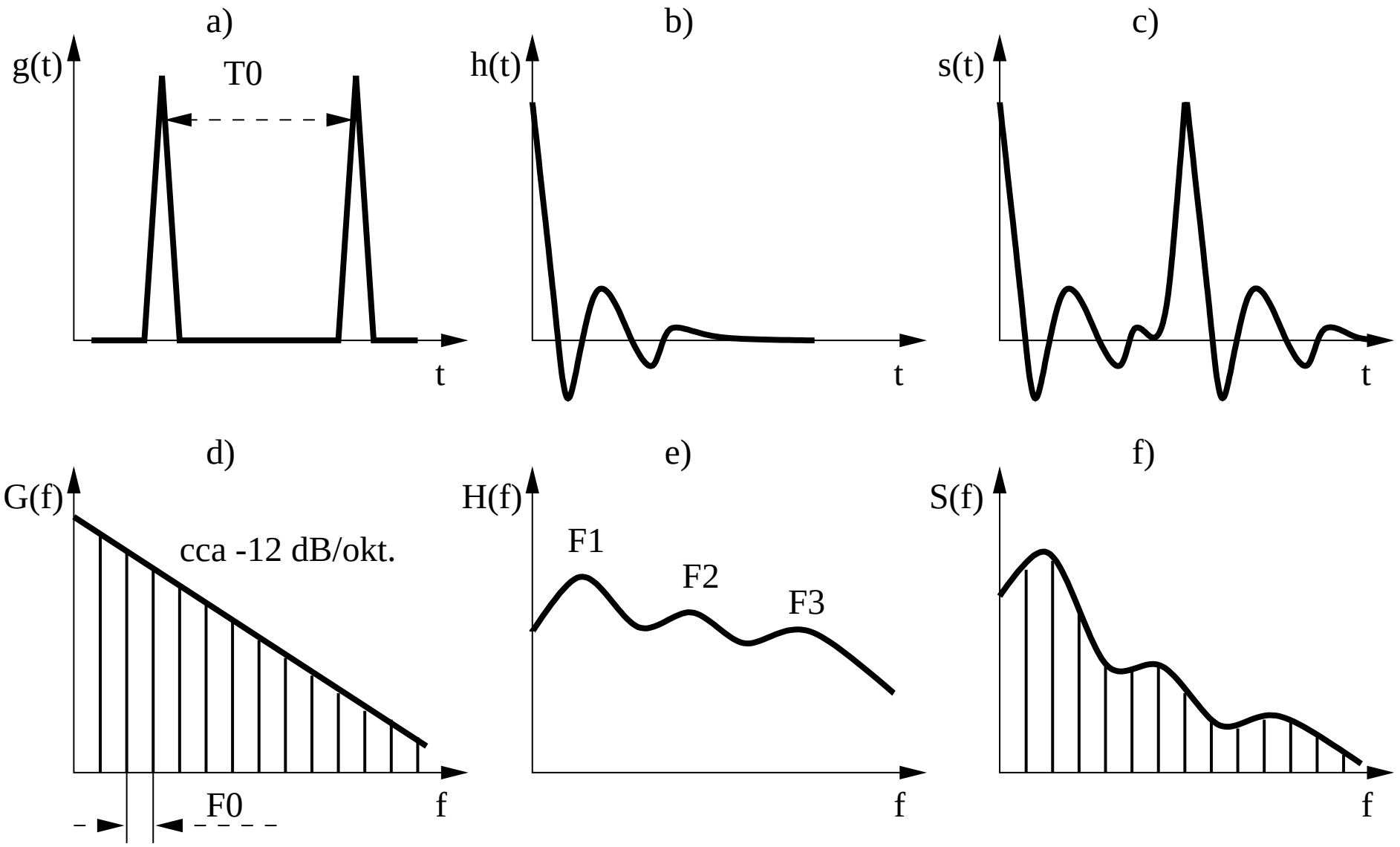
  - Buzení je většinou *smíšené* (moderní kodéry, GSM . . .)
- **hlasový (artikulační) trakt** — modifikující ústrojí — **signály: filtrace**.
  - hltan (pharynx).
  - měkké patro (vélum).
  - jazyk.
  - ústní dutina.
  - nosní dutina.
  - zuby.
  - rty.

Model



Přenosový systém: lineární filtr - nejčastěji IIR.

Model hlasového traktu v časové a frekvenční oblasti



Horní polovina – časové průběhy, spodní polovina – spektra.

- a) a d) buzení: T_0 je perioda, F_0 je frekvence základního tónu.
- b) a e) artikulační trakt: F_1 až F_3 jsou formanty (rezonanční frekvence hlasového traktu), dány jeho fyzickou konfigurací.
- c) a f) výsledný signál a jeho spektrum.

Výsledný signál je dán v časové oblasti *konvolucí*:

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

Do kmitočtové oblasti se tato konvoluce promítá jako *součin*:

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

Důležitý (a nelehký) úkol ve zpr. řeči je **dekonvoluce**, kdy se snažíme oddělit vliv buzení a artikulačního traktu.

SPEKTROGRAM

Jedno spektrum nestačí (řeč je nestacionární) $\Rightarrow$ znázornění průběhu spektra řeči (přesněji spektrální hustoty výkonu – PSD) v čase:

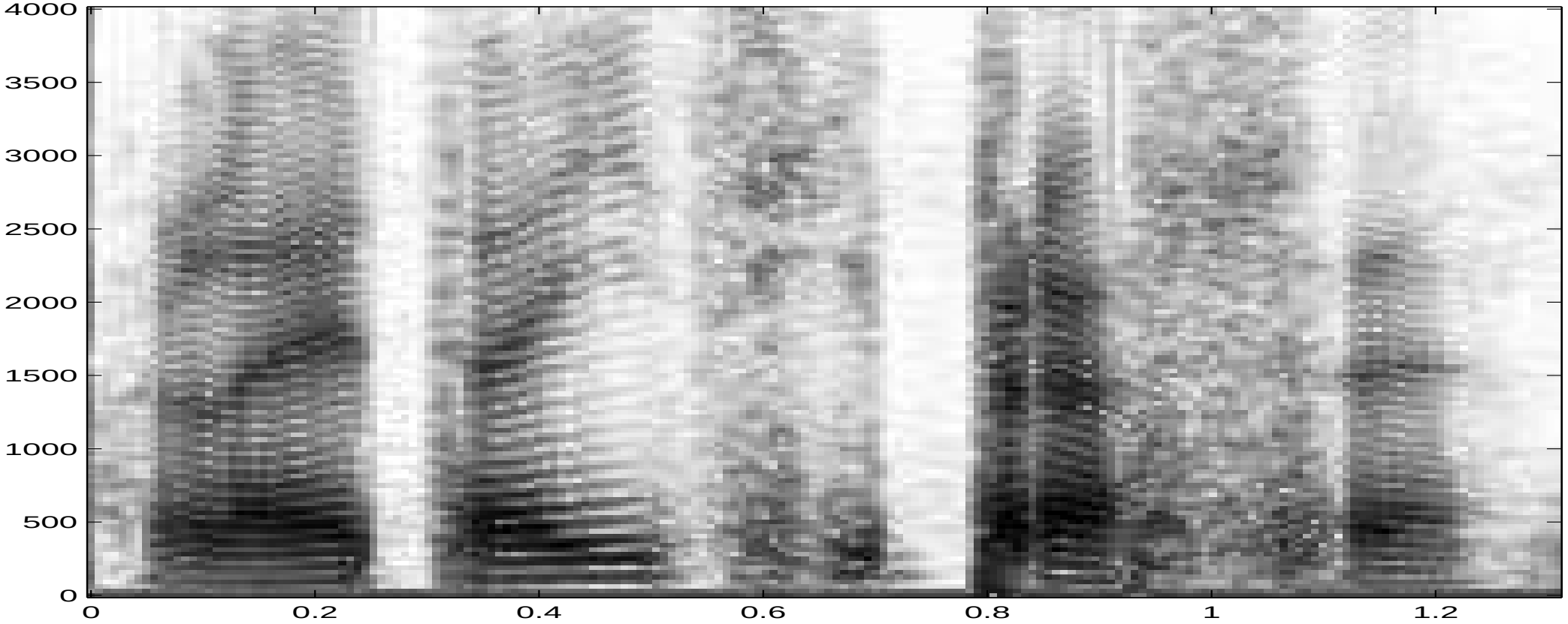
- řeč dělíme na rámce.
- v každém rámci odhadneme PSD, nejčastěji pomocí DFT.
- zobrazení:
  - vodorovná osa čas ("hrubý" čas v rámcích).
  - svislá osa frekvence.
  - stupně šedi nebo barvičky udávají energii.

Podle délky rámce dělíme na:

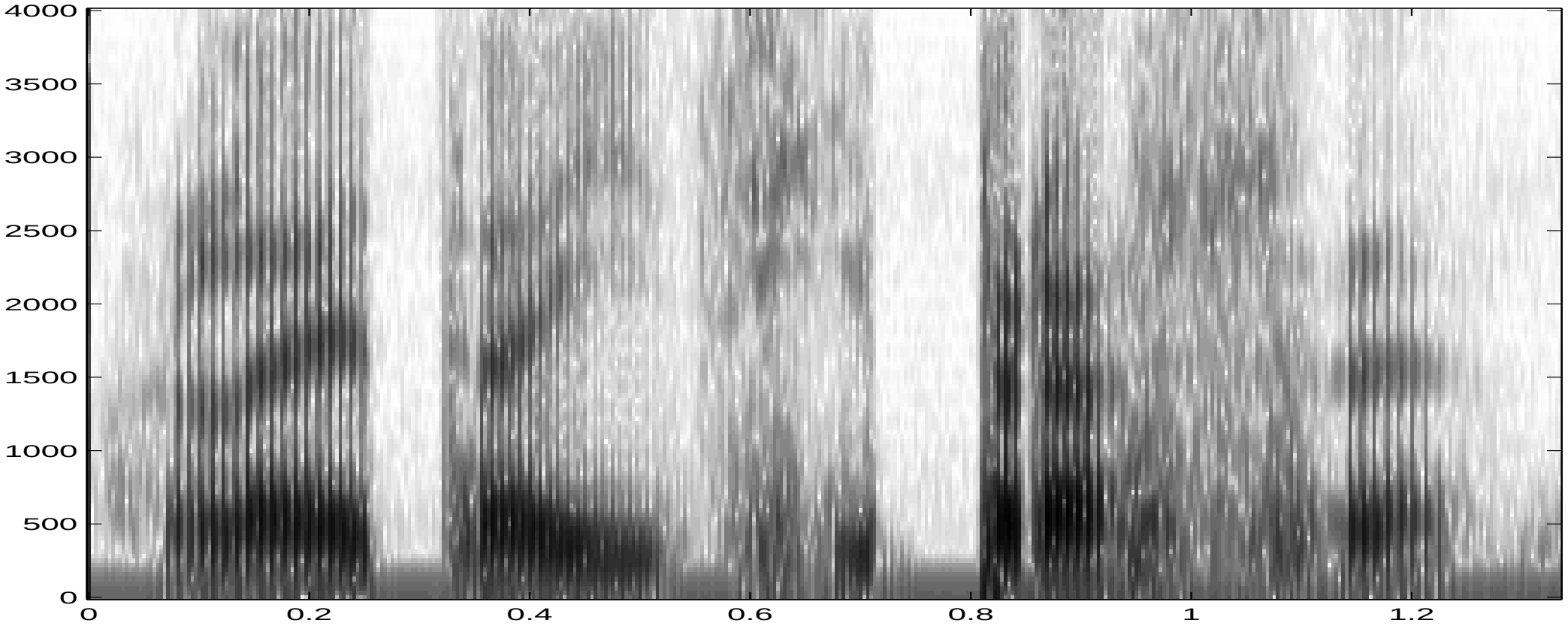
- long-term spectrogram.
- short-term spectrogram.

☹ Nevýhoda DFT: Přesnost ve frekvenci a v čase se nedá získat zároveň – wavelet analysis...

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



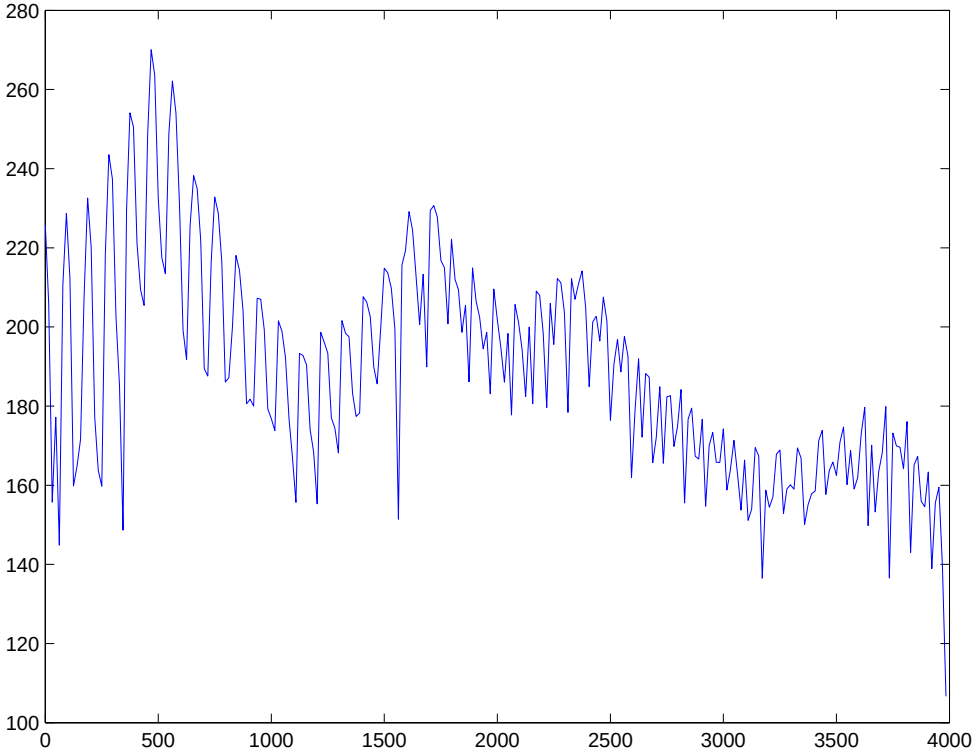
short-term: `specgram(s,256,8000,hamming(50));`



CEPSTRUM

pro oddělení buzení a modifikace – v kódování se nám s nimi pracuje lépe samostatně, v rozpoznávání buzení zahazujeme úplně (příliš závislé na řečníkovi, náladě, . . .).

Návrh č. 1: odfiltrujeme část pod 400 Hz a zbavíme se základního tónu . . . **BAD IDEA:**



- násobky základního tónu jsou "rozesety" po celém spektru.
- mohli bychom přijít o první formant.
- telefonní pásmo začíná na 300 Hz a také perfektně slyšíme, s jakou výškou hlasu člověk mluví.
- . . . takže budeme potřebovat něco lepšího.

$\Rightarrow$ **Cepstrum**

The problem

Buzení $e(t)$ je konvoluováno s impulsní odezvou filtru:

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

v kmitočtové oblasti *součin*:

$$S(f) = G(f)H(f). \tag{19}$$

ani v jedna z oblastí nelze dvě složky dobře oddělit. Řešení: **nelinearita**, která dokáže převést součin na součet.

Definice cepstra

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

Hodnoty $c(n)$ jsou **cepstrální koeficienty**. Jelikož $G(f)$ je sudá funkce, jsou $c(n)$ reálné a platí:

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

Suma v rovnici je definicí DFT, proto můžeme $c(n)$ vypočítat jako:

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

DFT-cepstrum

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

- spectrum $\longrightarrow$ cepstrum.
- frekvence $\longrightarrow$ kvefrence.
- filtrování $\longrightarrow$ liftrování.
- atd.

Opravdu to dokáže "rozetnout" konvoluci ?

$$s(n) = e(n) \star h(n), \tag{24}$$

$$S(f) = E(f)H(f) \quad \text{a tedy} \quad |S(f)|^2 = |E(f)|^2 \, |H(f)|^2. \tag{25}$$

Při výpočtu cepstra využíváme linearity zpětné Fourierovy transformace:

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

Dostáváme:

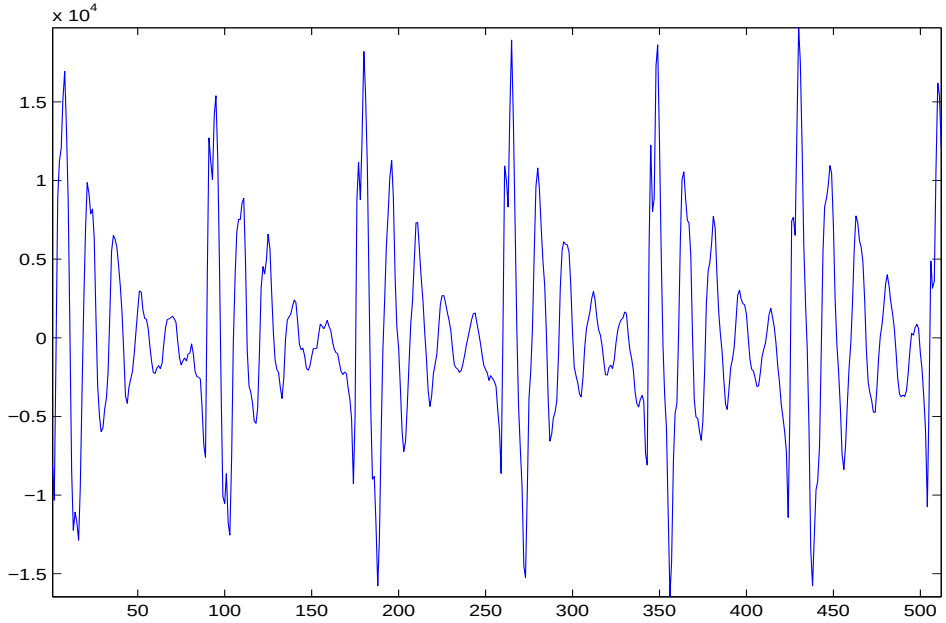
$$c(n) = \mathcal{F}^{-1}\left\{\ln[|E(f)|^2 \, |H(f)|^2]\right\} = \mathcal{F}^{-1}\left\{\ln|E(f)|^2 + \ln|H(f)|^2\right\} = \tag{26}$$

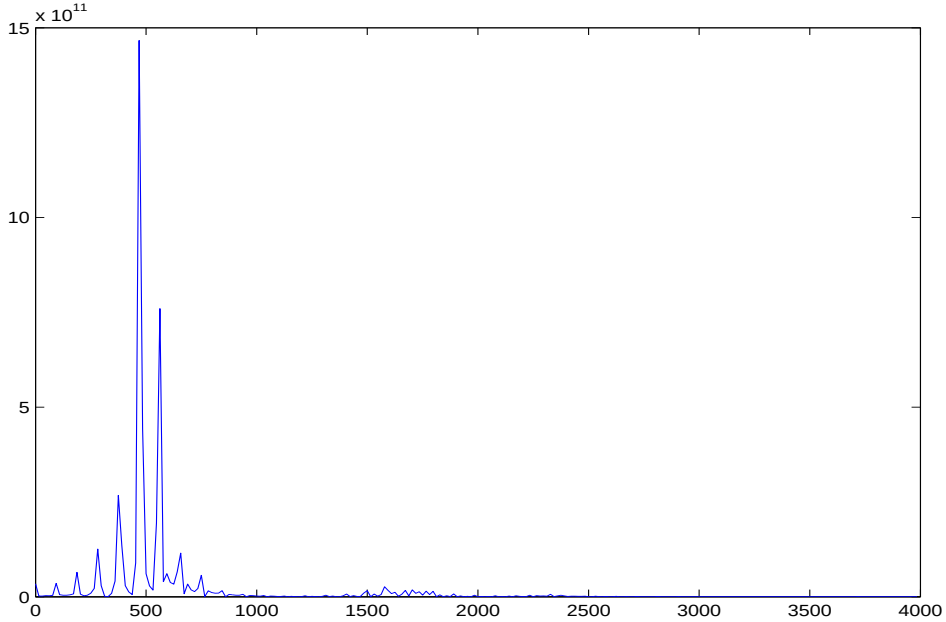
$$= \mathcal{F}^{-1}\left\{\ln|E(f)|^2\right\} + \mathcal{F}^{-1}\left\{\ln|H(f)|^2\right\} = c_e(n) + c_h(n) \tag{27}$$

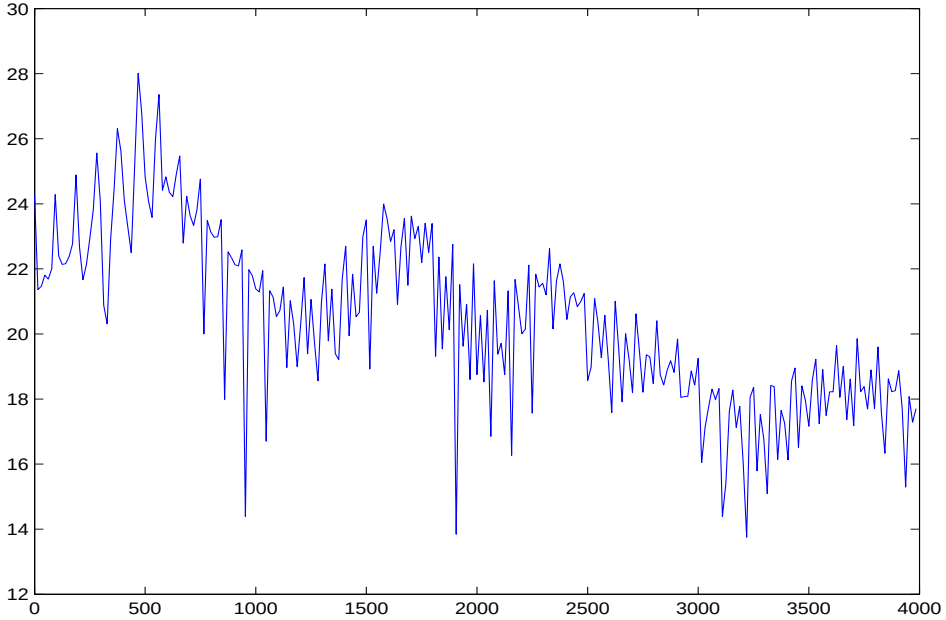
$$\tag{28}$$

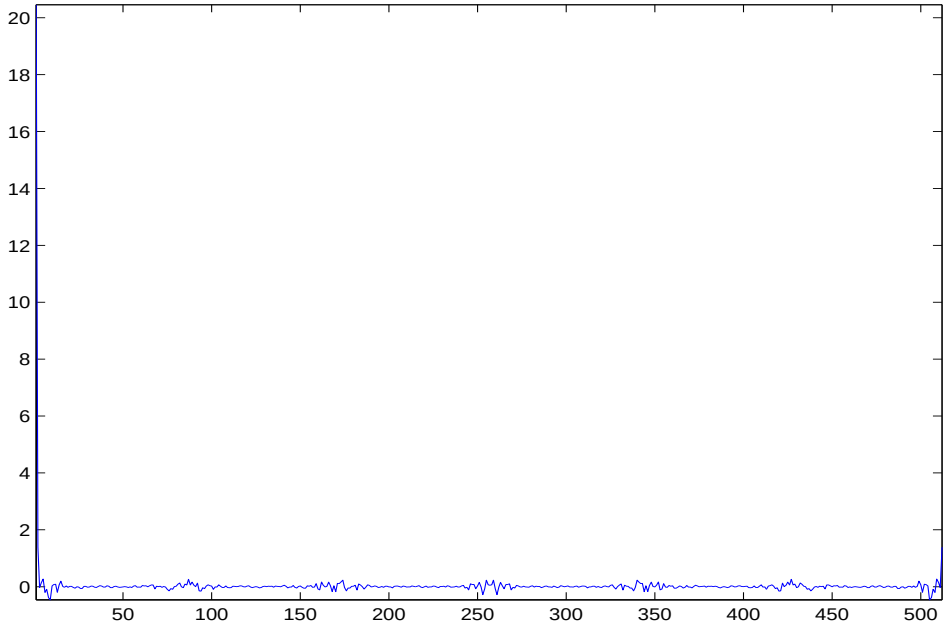
Z konvoluce se stal **součet**. Pokud jsou koeficienty $c_e(n)$ a $c_h(n)$ odděleny na kvefrenční ose, je možné je separovat jednoduchým oknem — naštěstí **jsou**.

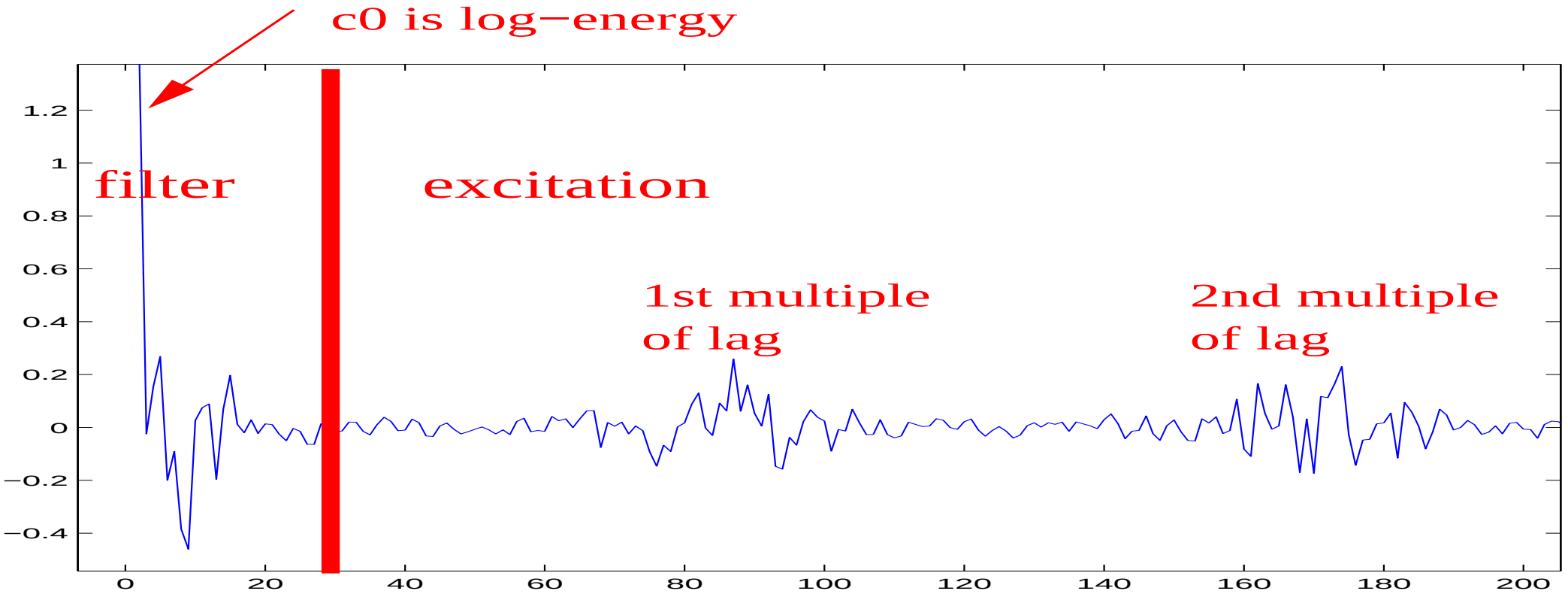
signál, $|\mathcal{F}[s(n)]|^2$, $\ln|\mathcal{F}[s(n)]|^2$, $\mathcal{F}^{-1}\left\{\ln|\mathcal{F}[s(n)]|^2\right\}$.







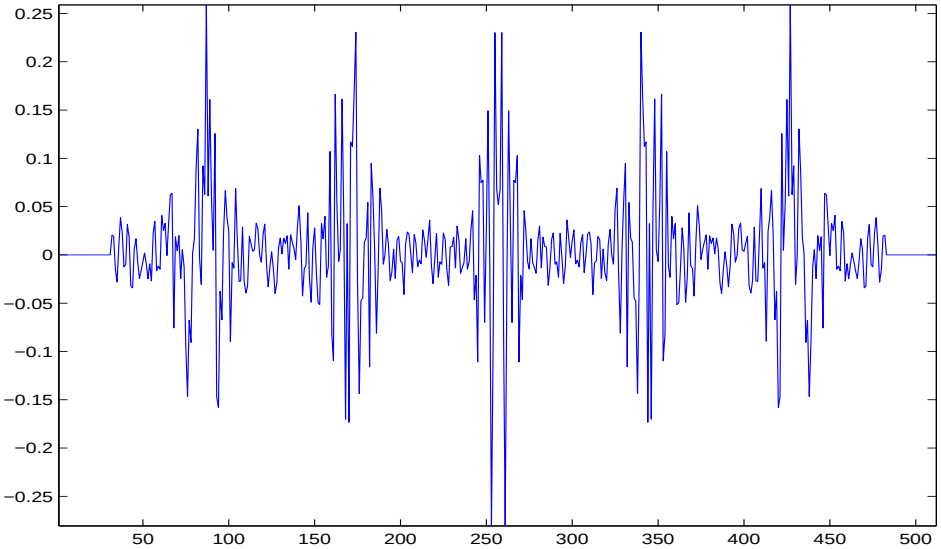


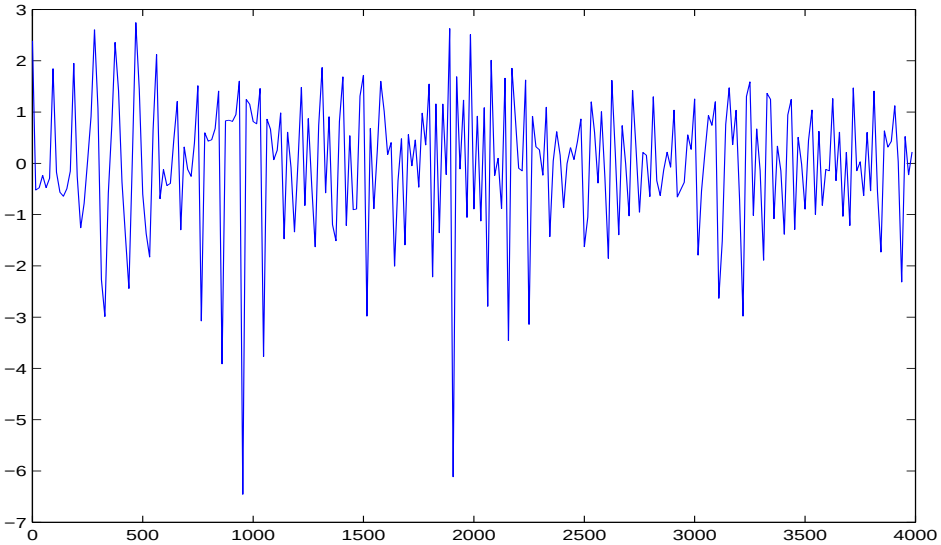


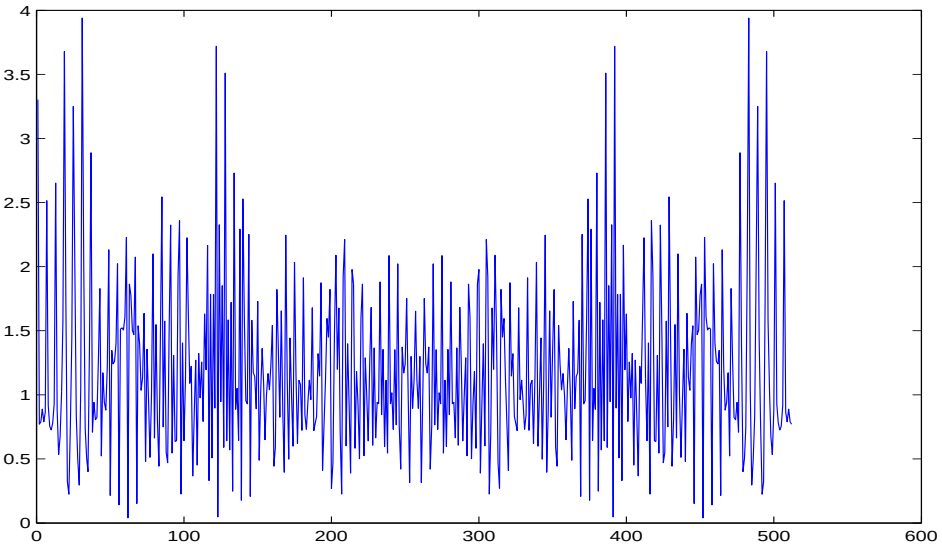
Pro vzorkovací frekvenci F_s =8000 Hz můžeme na kvefrenční ose oddělit vliv buzení a filtru separací s hranicí 30.

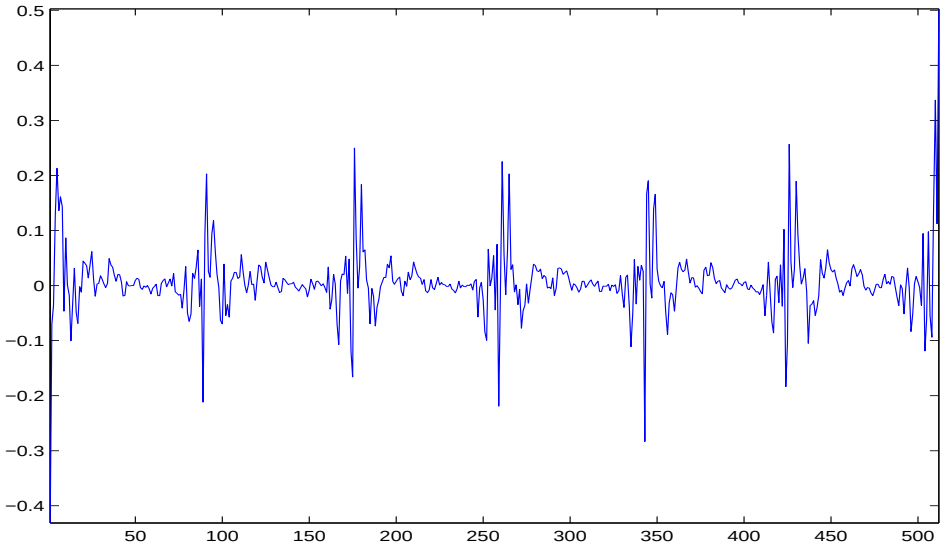
Pouze buzení – vynulujeme cepstra patřící filtru:

vynulované cepstrum, $\ln |\mathcal{F}[s(n)]|^2$, $|\mathcal{F}[s(n)]|$, signál (při IDFT byly použity fáze původního signálu).



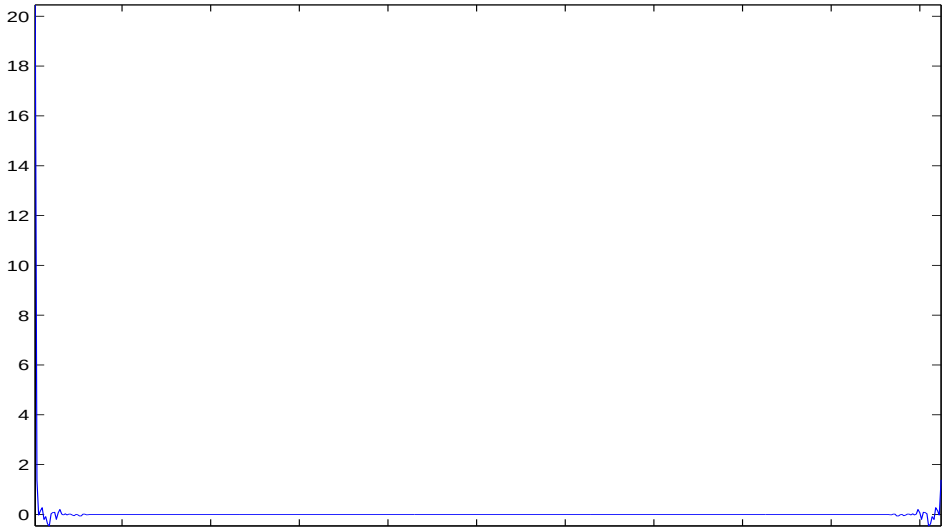


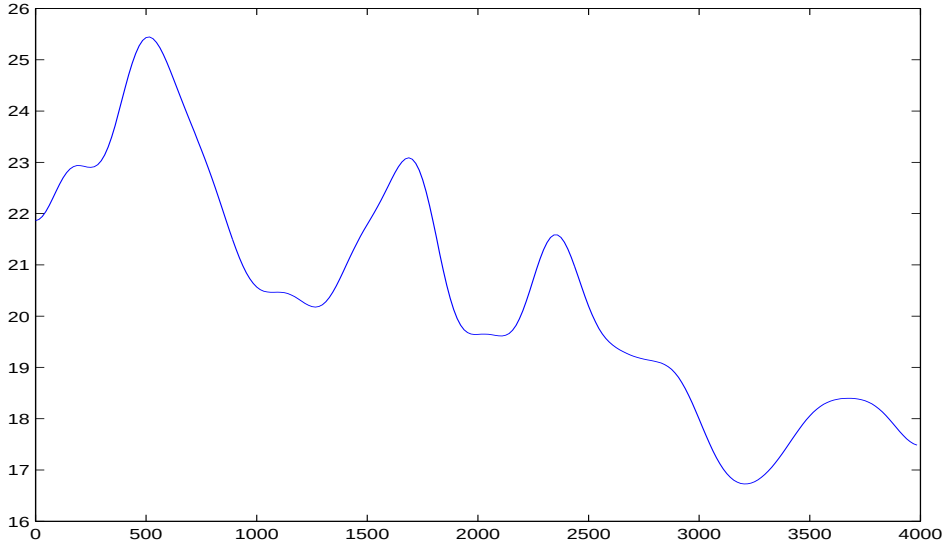


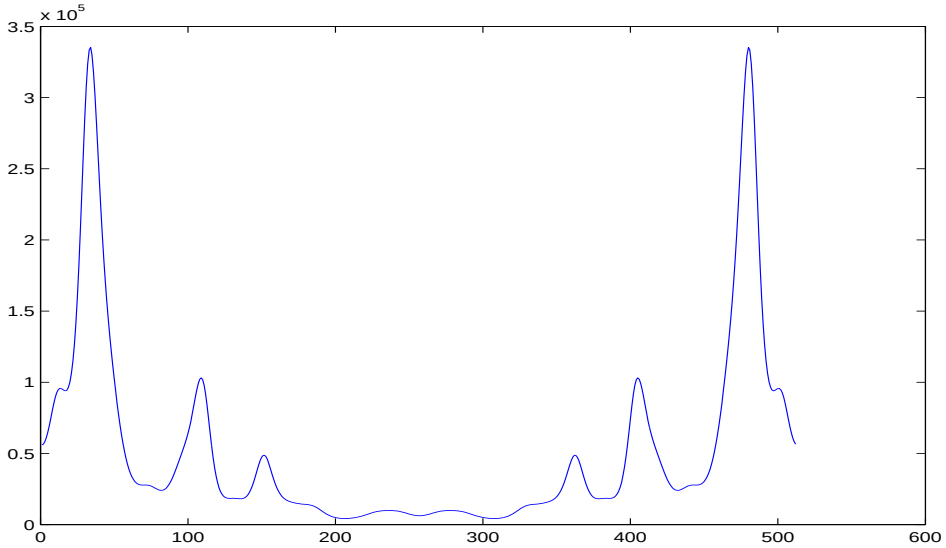


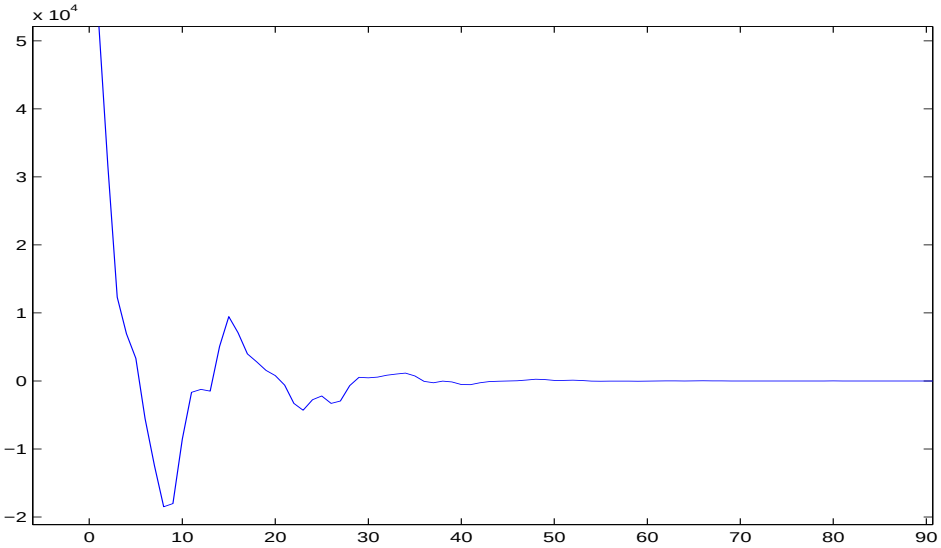
Pouze filtr – vynulujeme cepstra patřící buzení:

vynulované cepstrum, $\ln|\mathcal{F}[s(n)]|^2$, $|\mathcal{F}[s(n)]|$, signál (při IDFT byly použity fáze nula).









Mel-frequency cepstrum – MFCC

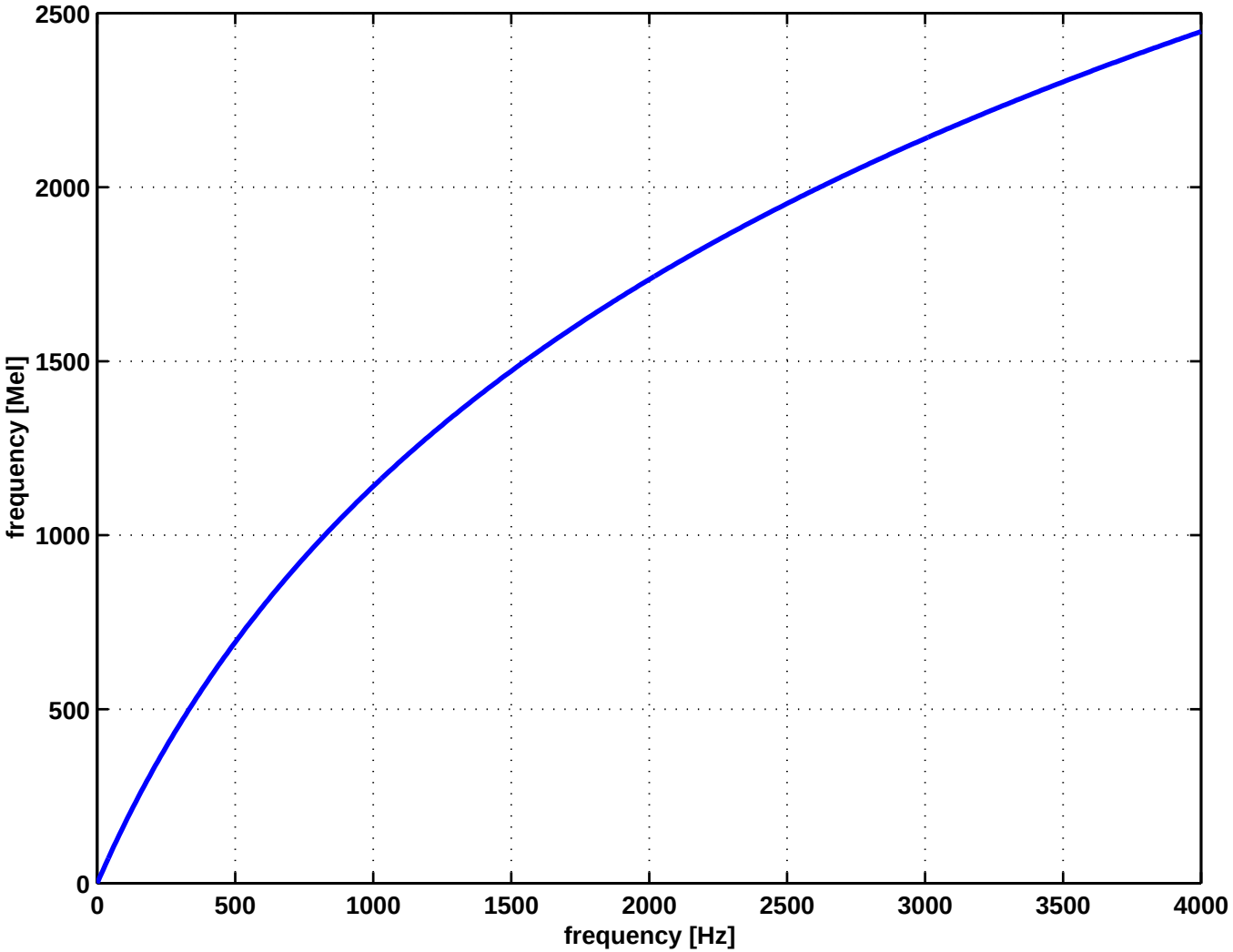
- DFT má všude stejné frekvenční rozlišení.
- Lidské ucho má na nízkých frekvencích větší rozlišení než na vysokých.
- Pro rozpoznávače řeči chceme přiblížit cepstrum slyšení.

Jak na to ?

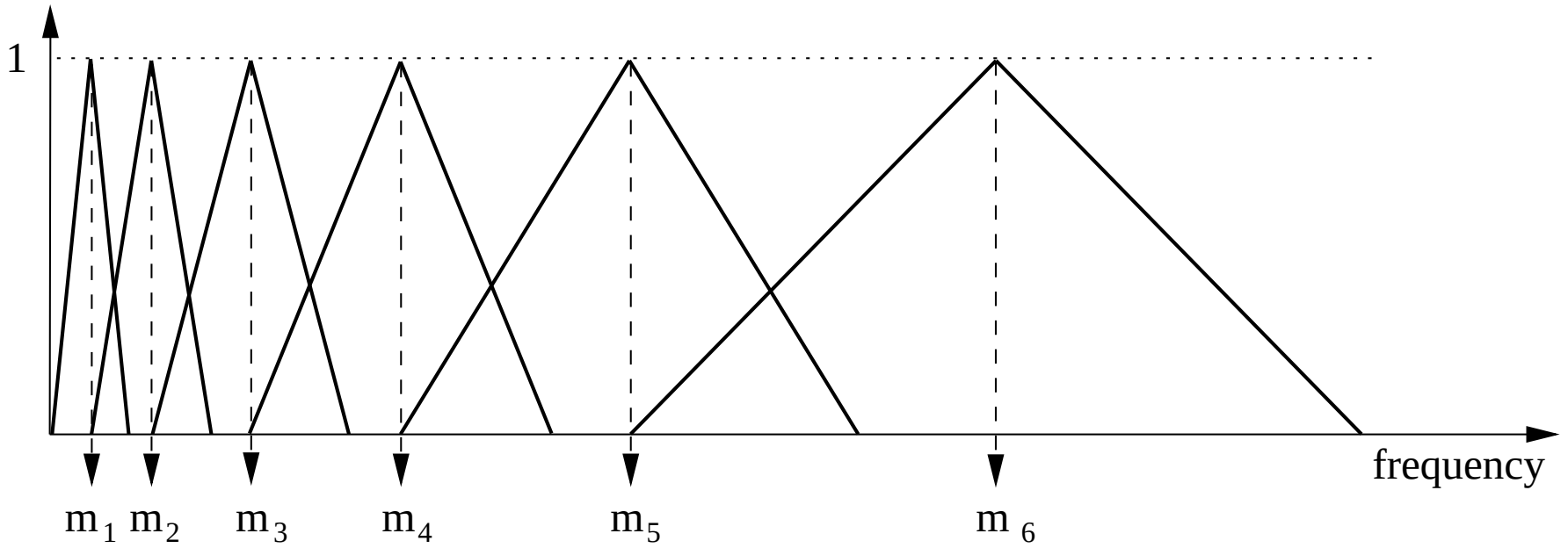
- Na frekvenční osu rozmístíme **nelineárně** filtry, měříme energii na jejich výstupu, použijeme je místo DFT při výpočtu cepstra.
- Frekvenční osu můžeme nelineárně upravit a na upravené ose pak filtry rozmístit rovnoměrně.

Používaná nelineární úprava využívá převodu Hertzů na Mely:

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



Lineární rozmístění filtrů na Mel-ové ose má za následek nelineární rozmístění na stadardní kmitočtové ose v Hz:



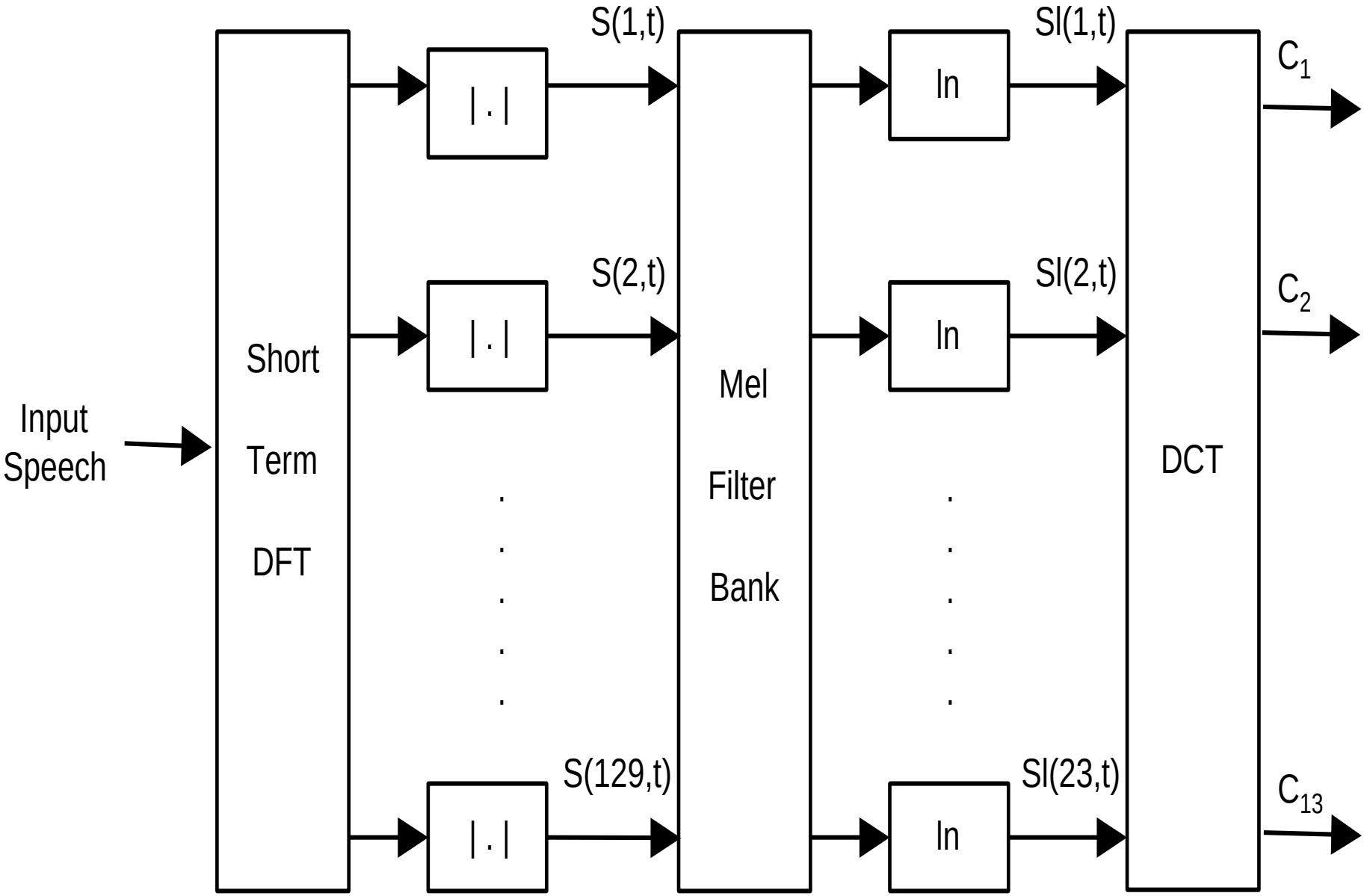
Výpočet energií:

1. zkonstruujeme banku filtrů, vstupní signál filtrujeme v časové oblasti a počítáme energie: $\sum_n s_i^2(n)$ ... MOC SLOŽITÉ.
2. provedeme DFT, umocníme, vynásobíme trojúhelníkovým oknem a sečteme. (použito v toolkitu pro rozpoznávání řeči HTK Hidden Markov Model ToolKit).

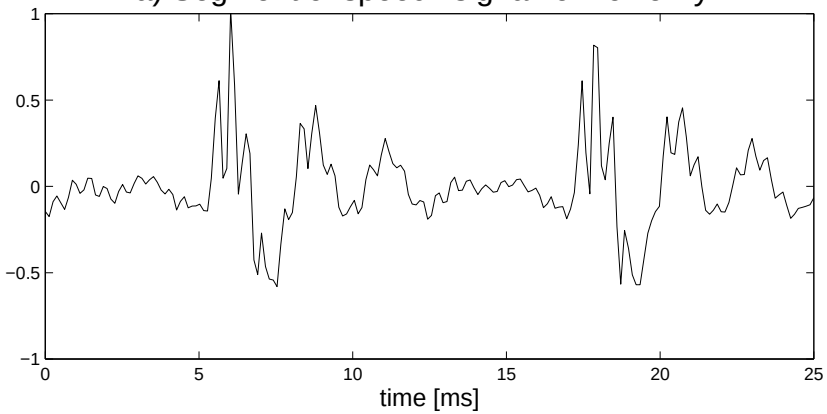
Zpětnou FT můžeme realizovat pomocí diskrétní cosinové transformace (DCT) ... (bez odvození: využíváme symetrie spektra a toho, že výsledek musí vyjít reálný):

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

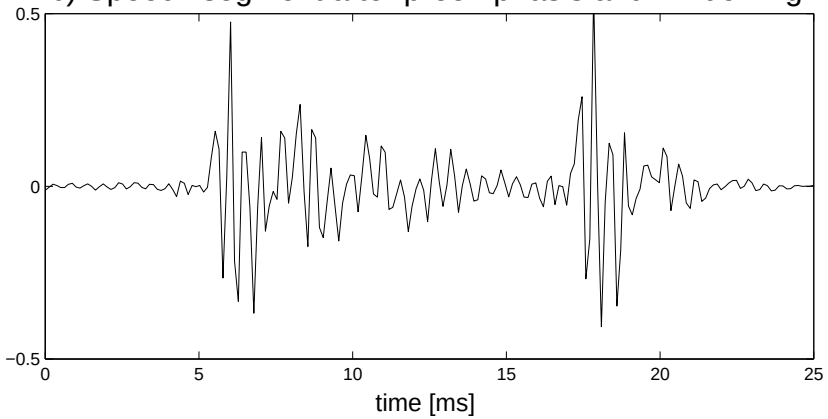
$\Rightarrow$ **Mel-frekvenční cepstrální koeficienty (MFCC)**


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

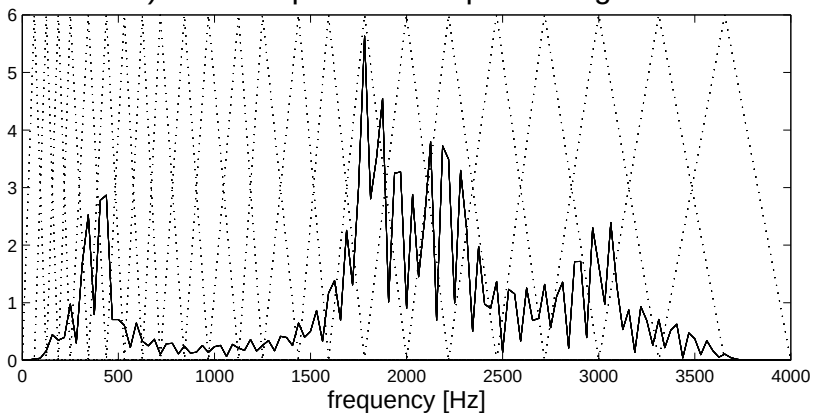
a) Segment of speech signal for vowel 'iy'



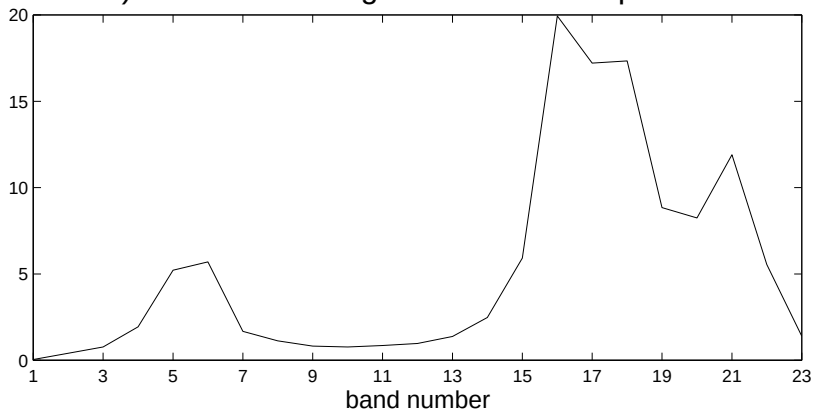
b) Speech segment after preemphasis and windowing



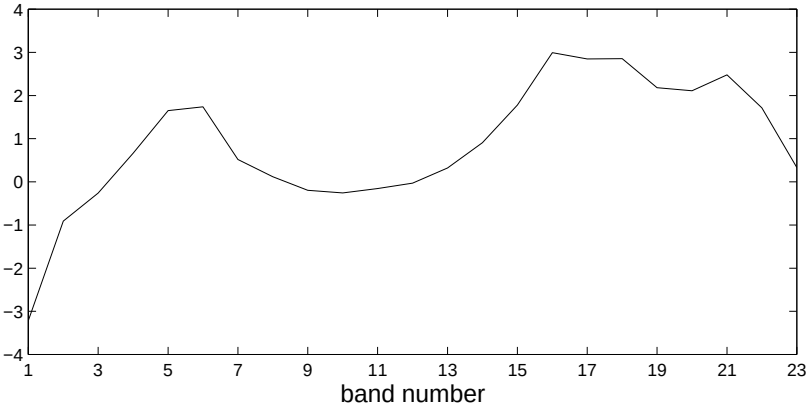
c) Fourier spectrum of speech segment



d) Filter bank energies − smoothed spectrum



e) Log of filter bank energies



f) Mel frefuency cepstral coefficients