Interakce člověk–počítač v přirozeném jazyce (ICP)

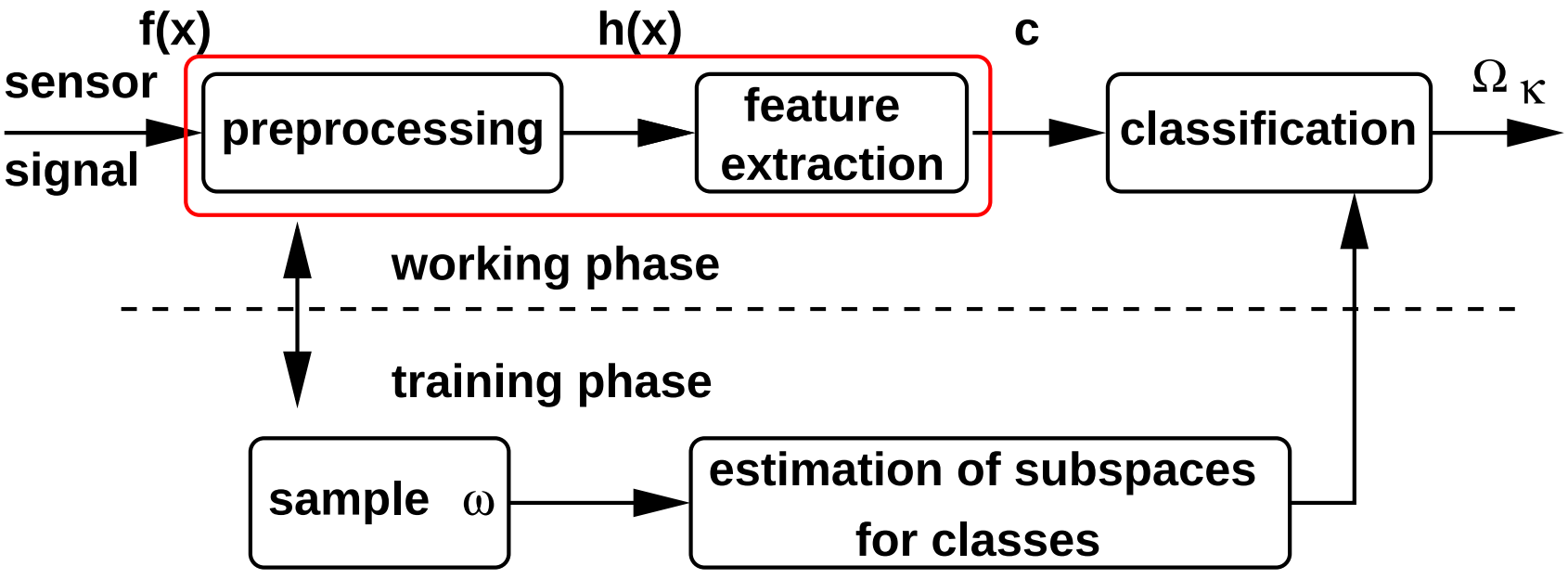
LS 2013 — Příznaky

Tino Haderlein, Elmar Nöth

Katedra informatiky a výpočetní techniky (KIV)
Západočeská univerzita v Plzni

Lehrstuhl für Mustererkennung (LME)
Friedrich-Alexander-Universität Erlangen-Nürnberg

Klasifikační systém



Diskrétní Fourierova Transformace (DFT)

$$F_\mu = \sum_{j=0}^{M-1} f_j exp(\frac{-i2\pi}{M} j\mu), \quad kde \;\; W_M = exp(\frac{-i2\pi}{M})$$

$$F_\mu = \sum_{j=0}^{M-1} f_j W_M^{j\mu}, \qquad kde \;\; M = 8 \quad \Rightarrow$$

$$\begin{aligned}
F_0 &= f_0 W_8^0 + f_1 W_8^0 + f_2 W_8^0 + f_3 W_8^0 + \\
&\quad f_4 W_8^0 + f_5 W_8^0 + f_6 W_8^0 + f_7 W_8^0 \\
F_1 &= f_0 W_8^0 + f_1 W_8^1 + f_2 W_8^2 + f_3 W_8^3 + \\
&\quad f_4 W_8^4 + f_5 W_8^5 + f_6 W_8^6 + f_7 W_8^7 \\
F_2 &= f_0 W_8^0 + f_1 W_8^2 + f_2 W_8^4 + f_3 W_8^6 + \\
&\quad f_4 W_8^8 + f_5 W_8^{10} + f_6 W_8^{12} + f_7 W_8^{14} \\
&\ldots, \\
kde \;\; W_8^n &= W_8^{(n \bmod 8)}
\end{aligned}$$

Diskrétní Fourierova Transformace (DFT)

000	0	f0	0	f1	0	f2	0	f3	0	f4	0	f5	0	f6	0	f7	0
001	1	f0	0	f1	1	f2	2	f3	3	f4	4	f5	5	f6	6	f7	7
010	2	f0	0	f1	2	f2	4	f3	6	f4	0	f5	2	f6	4	f7	6
011	3	f0	0	f1	3	f2	6	f3	1	f4	4	f5	7	f6	2	f7	5
100	4	f0	0	f1	4	f2	0	f3	4	f4	0	f5	4	f6	0	f7	4
101	5	f0	0	f1	5	f2	2	f3	7	f4	4	f5	1	f6	6	f7	3
110	6	f0	0	f1	6	f2	4	f3	2	f4	0	f5	6	f6	4	f7	2
111	7	f0	0	f1	7	f2	6	f3	5	f4	4	f5	3	f6	2	f7	1
užívá bit reversal a $W_8^{4+k} = -W_8^k, k = 0, 1, 2, 3$															
000	0	f0	0	f1	0	f2	0	f3	0	f4	0	f5	0	f6	0	f7	0
100	4	f0	0	f1	-0	f2	0	f3	-0	f4	0	f5	-0	f6	0	f7	-0
010	2	f0	0	f1	2	f2	-0	f3	-2	f4	0	f5	2	f6	-0	f7	-2
110	6	f0	0	f1	-2	f2	-0	f3	2	f4	0	f5	-2	f6	-0	f7	2
001	1	f0	0	f1	1	f2	2	f3	3	f4	-0	f5	-1	f6	-2	f7	-3
101	5	f0	0	f1	-1	f2	2	f3	-3	f4	-0	f5	1	f6	-2	f7	3
011	3	f0	0	f1	3	f2	-2	f3	1	f4	-0	f5	-3	f6	2	f7	-1
111	7	f0	0	f1	-3	f2	-2	f3	-1	f4	-0	f5	3	f6	2	f7	1

Diskrétní Fourierova Transformace (DFT)

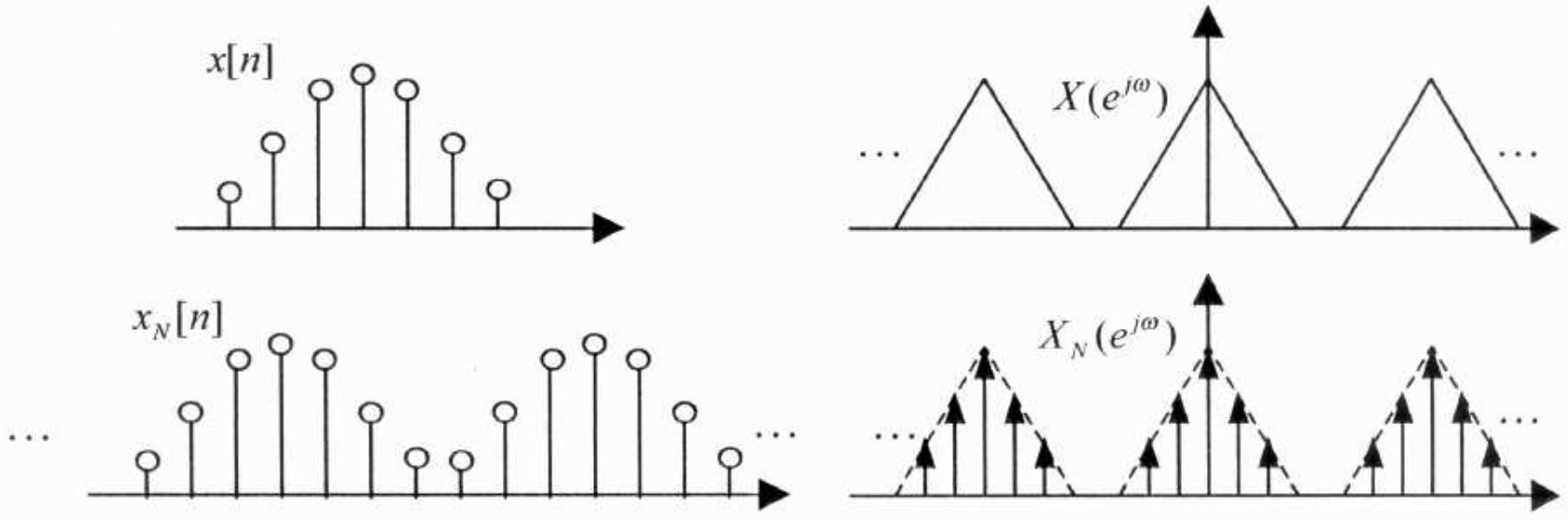


Figure 5.14 Relationships between finite and periodic signals and their Fourier transforms. On one hand, $x[n]$ is a length N discrete signal whose transform $X(e^{j\omega})$ is continuous and periodic with period 2π. On the other hand, $x_N[n]$ is a periodic signal with period N whose transform $X_N(e^{j\omega})$ is discrete and periodic.

Rekapitulace

- buď $F(e^{i\omega})$ Fourierův obraz funkce f_n
- f_n se jmenuje N-periodické, když $f_n = f_{n+jN}$ pro libovolné n a j

$\rightarrow$ všechny spektrální komponenty mimo u $f_A k/N$, $k = 0, \pm 1, \pm 2, \ldots$ zaniknou

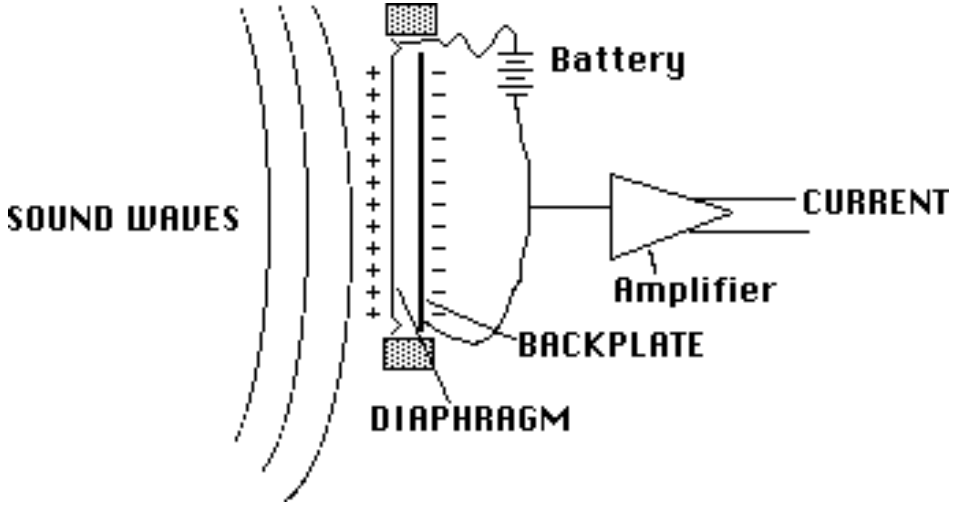
- výsledné čárové spektrum vznikne z Diskrétní Fourierové transformace (DFT)
  $F_\nu = \sum_{n=0}^{N-1} f_n e^{-2\pi i \nu n/N}$
- rekonstrukce pomocí inverzní DFT
  $f_n = \frac{1}{N} \sum_{\nu=0}^{N-1} F_\nu e^{2\pi i \nu n/N}$

Rekapitulace

- f_n neperiodické $\rightarrow$ periodické pokračování v časové oblasti = vzorkování kontinuálního spektra ve frekvenční oblasti
- ν-tý Fourierův koeficient popisuje spektrální hustotu u $\nu(NT)^{-1} = f_A\nu/N$ Hz
- frekvenční rozlišení je f_A/N Hz
- f_n je reálné $\Rightarrow$ symetrie $|F_\nu|^2 = |F_{-\nu}|^2$
- DFT vyžaduje N^2 komplexních násobení
  rychlá Fourierova transformace (FFT) vyžaduje jen $\mathcal{O}(\mathcal{N}\log\mathcal{N})$
  rychlá Hartleyova transformace (FHT) využívá, že f_n má jen reálné hodnoty

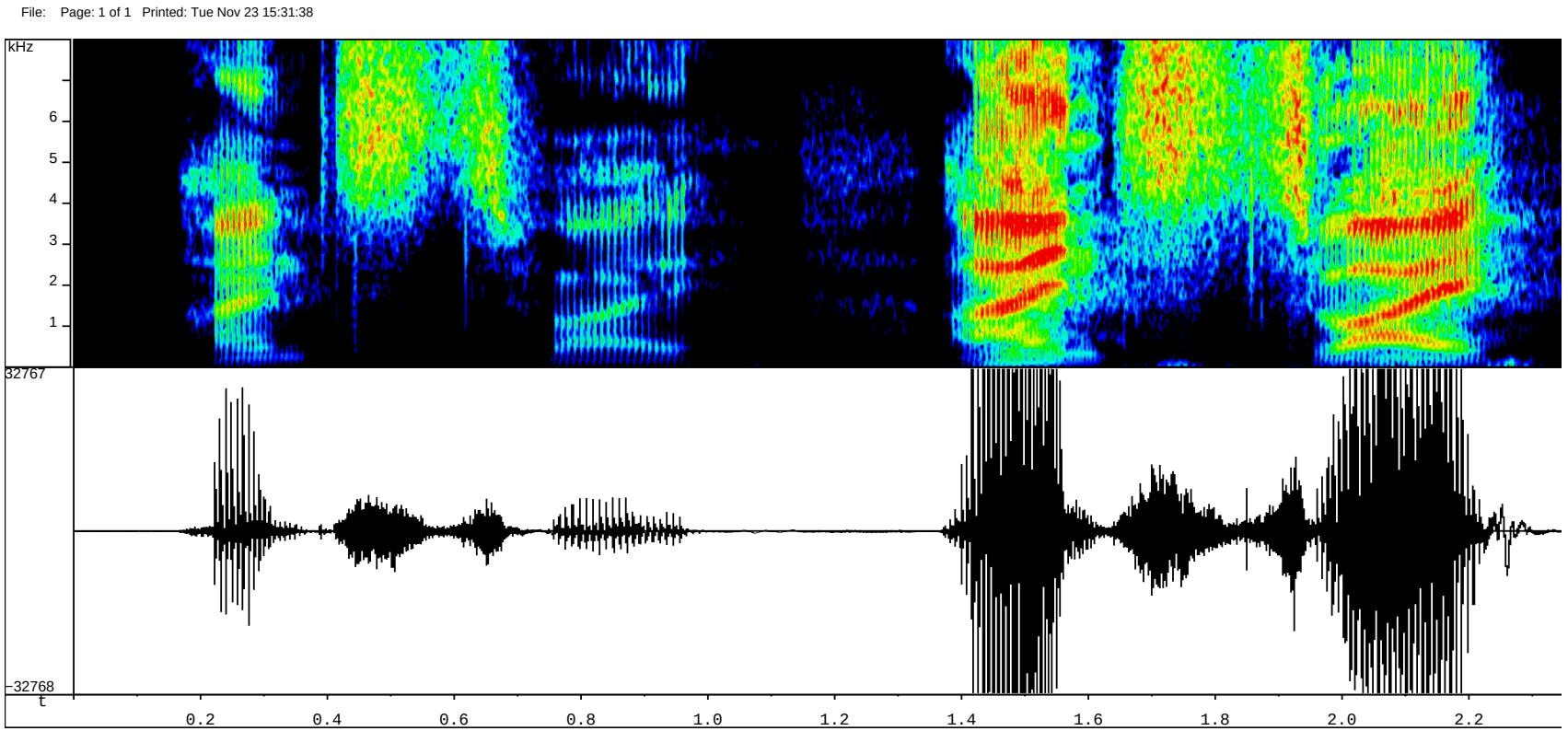
Recording Speech

Condenser Microphone:



- battery produces electrical potential between backplate and diaphragm
- amount of charge is determined by the voltage of the battery, the area of the diaphragm and backplate, and the distance between the two
- when the distance changes, current flows in the wire as the battery maintains the correct charge
- variant: material with a permanently imprinted charge for the diaphragm (= electret)

Přebuzení



přebuzené

Zavedení vzorkovacího teorému

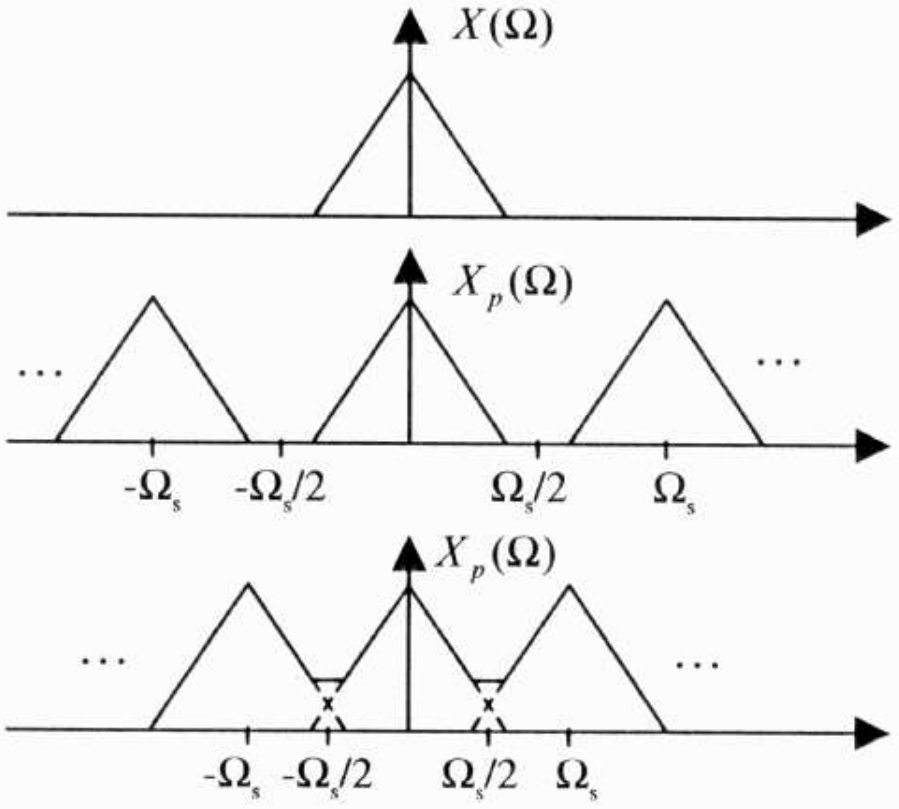
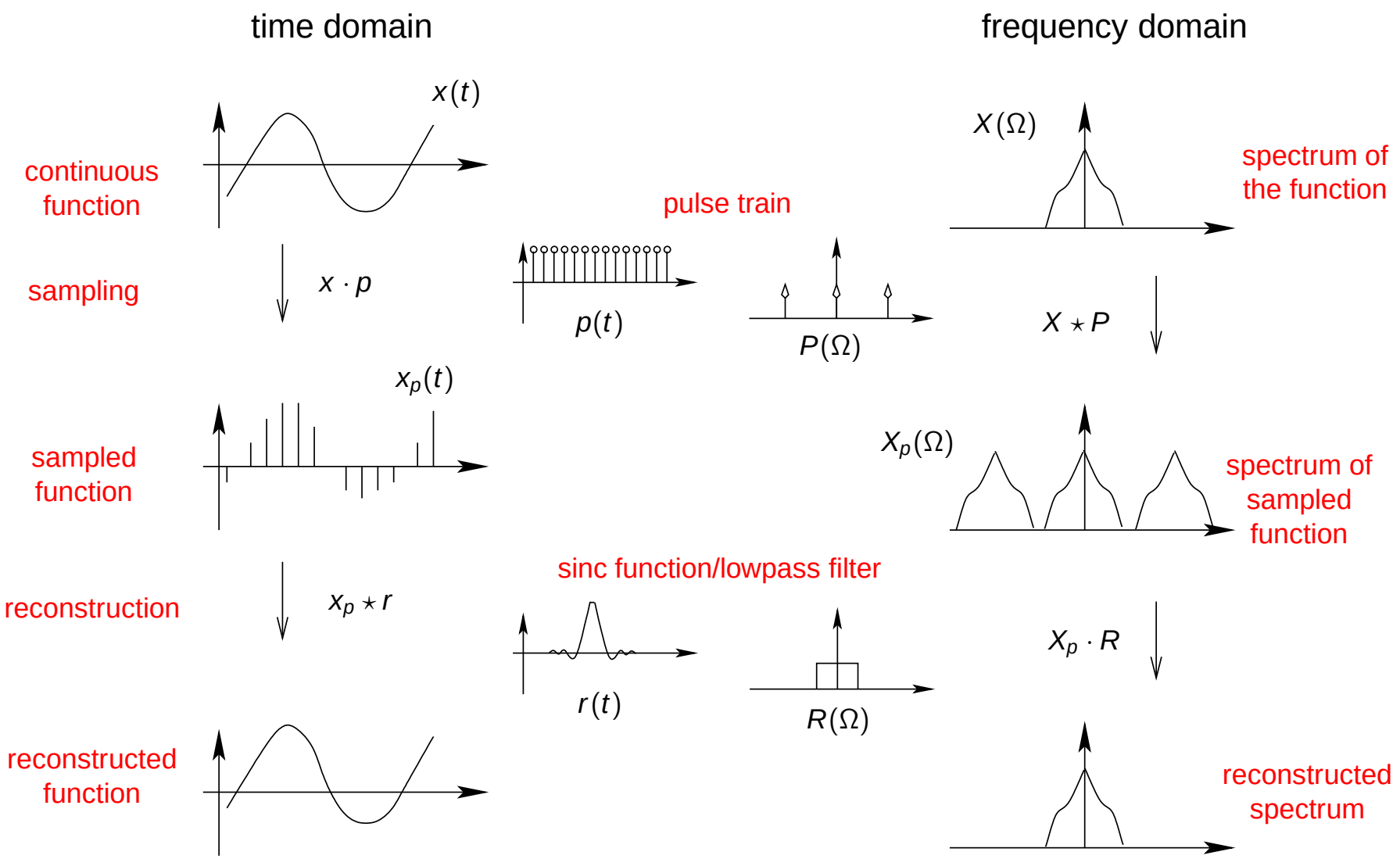


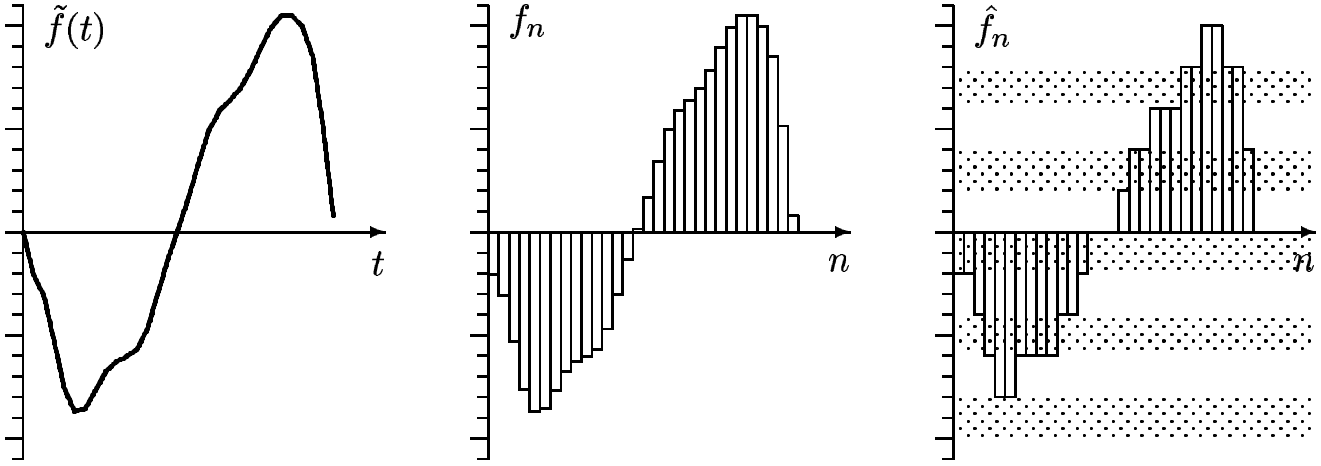
Figure 5.27 $X(\Omega)$, $X_p(\Omega)$ for the case of no aliasing and aliasing.

Zavedení vzorkovacího teorému



Vzorkování

- diskretizace spojitého (analogického) signálu $\tilde{f}(t)$
- diskretizace v definičním oboru: vzorkování $\longrightarrow f_n$
- diskretizace v rozsahu hodnot: kvantizace $\longrightarrow \hat{f}_n$
- vzorkovací perioda T = rozestup mezi vzorkovacími body
- vzorkovací frekvence $f_A = 1/T$ [Hz]



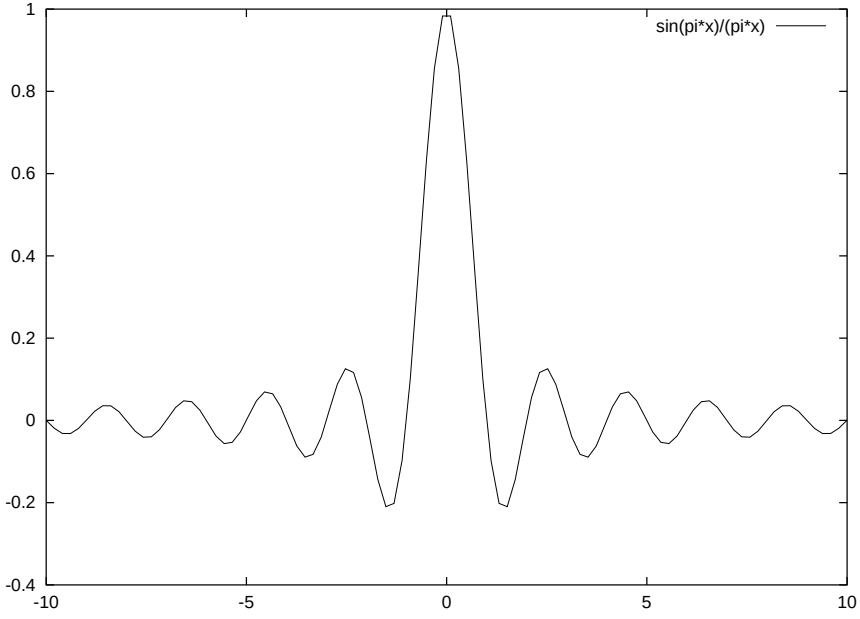
Vzorkovací teorém

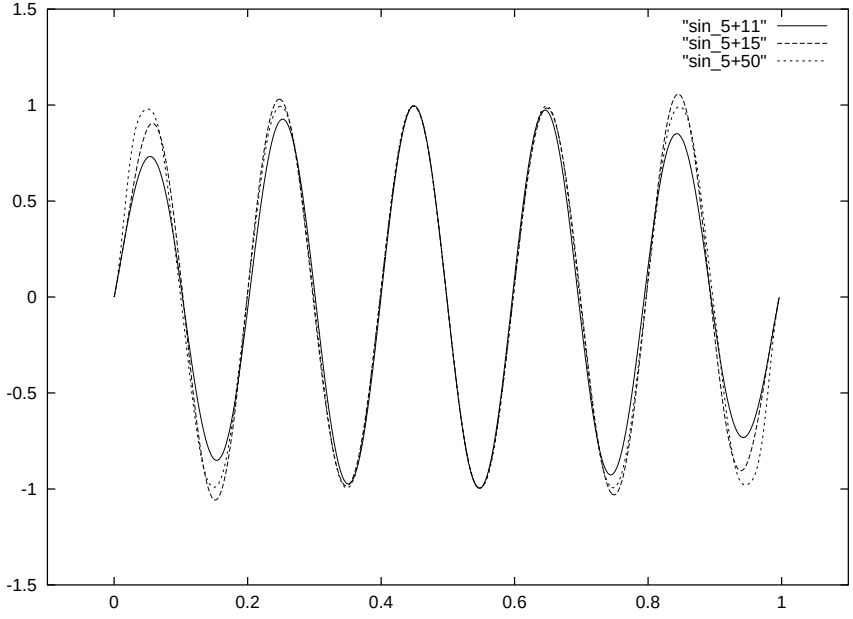
- Když signál $\tilde{f}(t)$ má omezení pásma a vzorkovací frekvence je dost vysoká, je možné signál úplně rekonstruovat ze vzorkovačních hodnot a nejsou u toho žádné artefakty.
- **(Nyquistův) vzorkovací teorém:**
  Buď pro Fourierův obraz $\tilde{F}(2\pi\omega)$ funkce $\tilde{f}(t)$ omezení pásma $\tilde{F}(2\pi\omega) = 0$ pro $|\omega| \geq f_G$:
  Pro vzorkovací frekvence vyšší než dvojitá hraniční frekvence ($f_A = 1/T \geq 2f_G$) se může $\tilde{f}(t)$ úplně rekonstruovat ze vzorkovačních hodnot $f_n = \tilde{f}(nT)$, $n = 0, \pm 1, \pm 2, \ldots$ interpolační rovnicí $\tilde{f}(t) = \sum_{n=-\infty}^{\infty} f_n \frac{\sin(\pi(t-nT)/T)}{\pi(t-nT)/T}$.

Vzorkovací teorém

- když se zanedbá vzorkovací teorém: spectral aliasing
- typické hodnoty:
  mikrofon: vzorkování 16 – 20 kHz, omezení pásma 8 – 10 kHz
  telefon: omezení pásma na 300 Hz – 3,4 kHz $\Rightarrow$ vzorkování s 8 kHz

Vzorkovací teorém: příklady

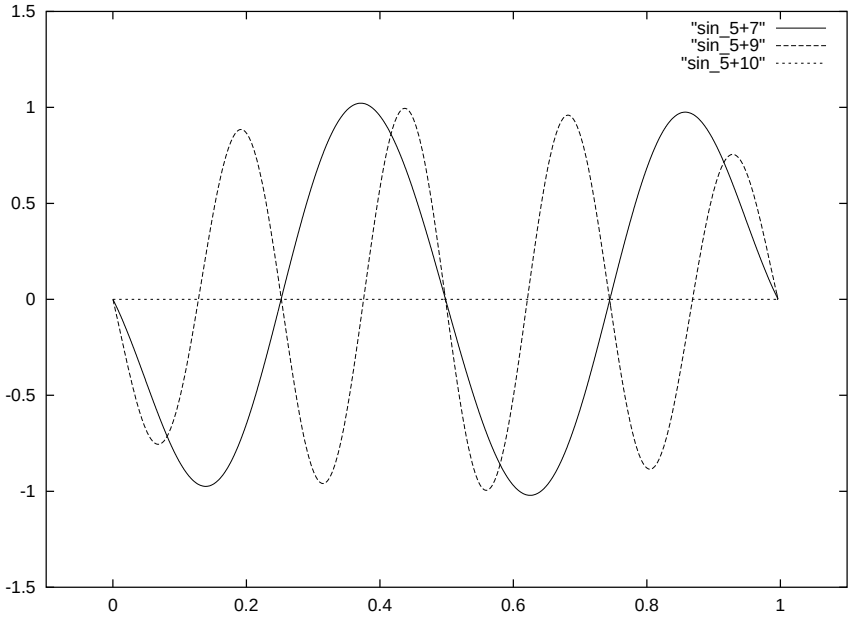


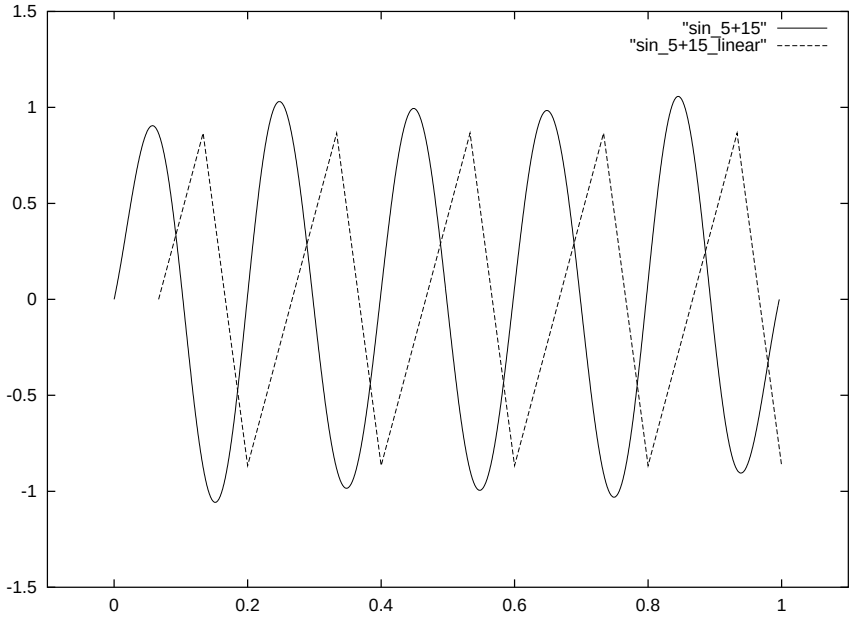


vlevo: „sinc“ funkce,

vpravo: správná vzorkovací funkce, narušení vzorkovacího teorému na krajích

Vzorkovací teorém: příklady



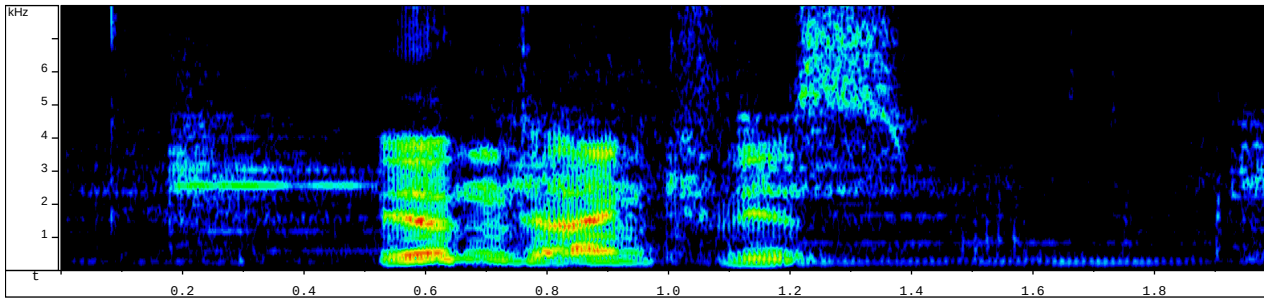


vlevo: příliš nízká vzorkovací frekvence,

vpravo: správná vzorkovací frekvence, rekonstrukce linearní interpolací

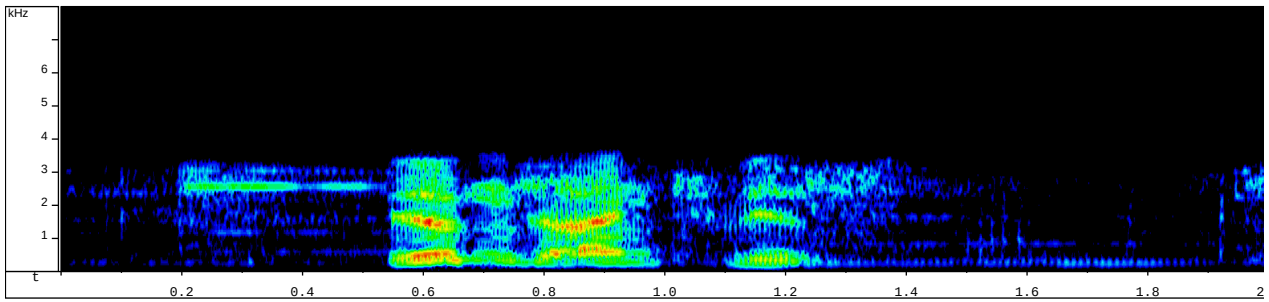
Příliš nízká vzorkovací frekvence

File: quiz−16kHz.wav Page: 1 of 1 Printed: Tue Nov 23 09:20:33



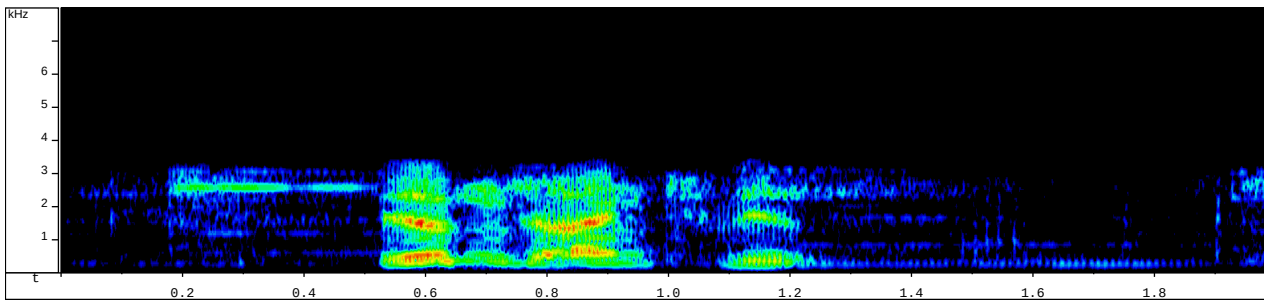
16 kHz

File: /users/noeth/vortraege/medav/schulung/AUDIO/quiz−8kHz−16.wav Page: 1 of 1 Printed: Thu Nov 25 22:31:30



8 kHz, bez dolního propustu

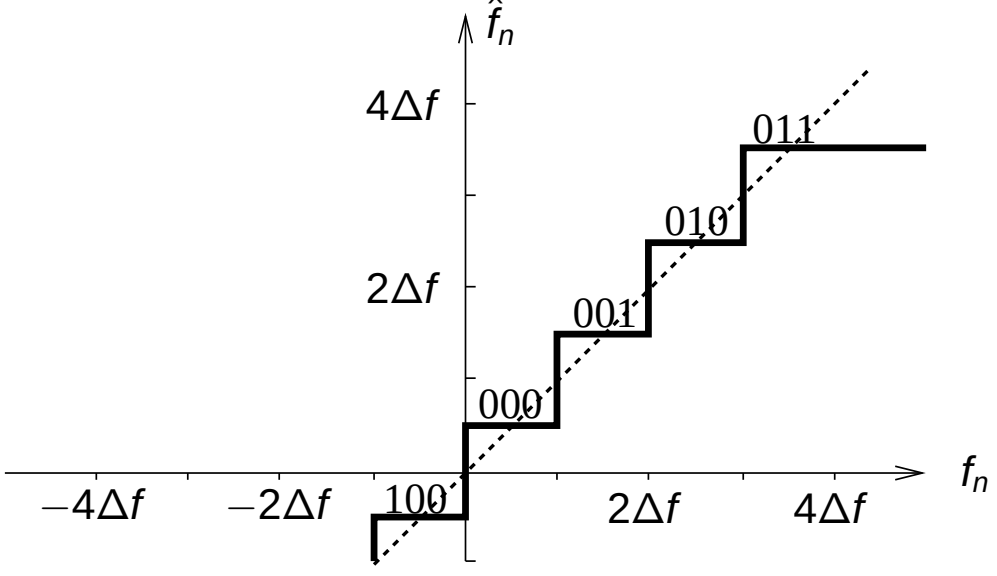
File: /users/noeth/vortraege/medav/schulung/AUDIO/quiz−16kHz−8−16.wav Page: 1 of 1 Printed: Thu Nov 25 22:33:12



8 kHz, s dolním propustem

Kvantizace: uniformní kvantizace

- rozklad rozsahu hodnot do 2^B pruhů se šířkou $\Delta f = 2f_{max}/2^B$
- f_n je zobrazen na nejbližší střed intervalu $\hat{f}_n$



- chyba kvantováním $e_n = \hat{f}_n - f_n$
- odstup od signálu k šumu (signal to noise ratio, SNR):
  $r = 10 \log_{10} \frac{E[f_n^2]}{E[e_n^2]}$ [dB]
- $E[\cdot]$: očekávaná hodnota

Aproximace SNR

Předpoklady:

- ustálený (stacionární) bílý (= rovnoměrně rozdělený) šum
- buď šum nezávislý od vzorkovacích hodnot
- buď šum rovnoměrně rozdělený v $[-\Delta f, \Delta f]$

$\Rightarrow$ uzavřené řešení pro SNR:

$$r = 6B + 4.77 - 20 \log_{10} \frac{f_{max}}{\sigma_f} ,$$

kde $\sigma_f^2 = \sqrt{E[f_n^2]}$ = standardní odchylka vstupních hodnot

Další zjednodušení:

- stanov $f_{max} = 4 \cdot \sigma_f \qquad \Rightarrow r = 6B - 7.2$
- dynamický rozsah řeči u 50–60 dB
  při optimální modulaci stačí 10–12 bit

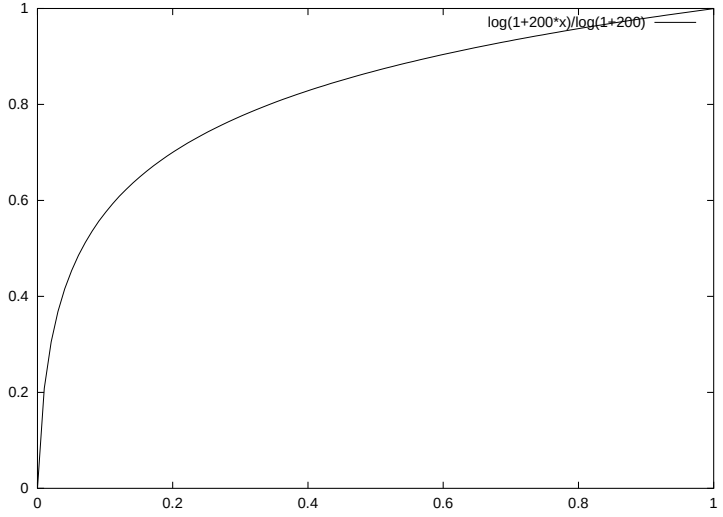
μ-law kódování

- hodnoty jsou v řeči rozdělovány exponenciálně

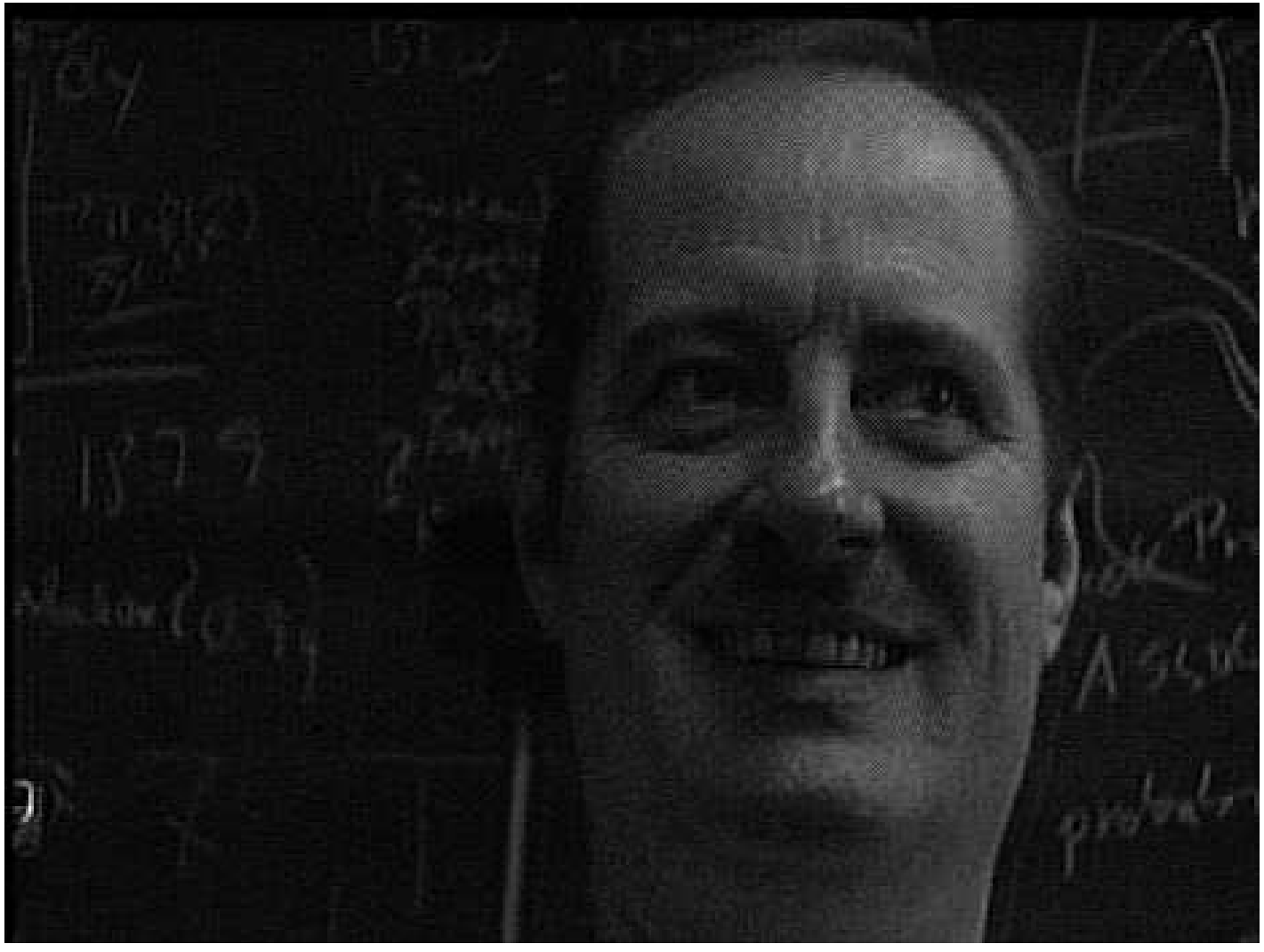
$\Rightarrow$ logaritmické zkřivení zvýší efektivnost kódování: rozlišení malých amplitud se zlepší, oblasti nad tím jsou komprimované

- logaritmická kompanze (μ-*law*) způsobí hodnoty s rovnoměrně rozdělenými hodnotami

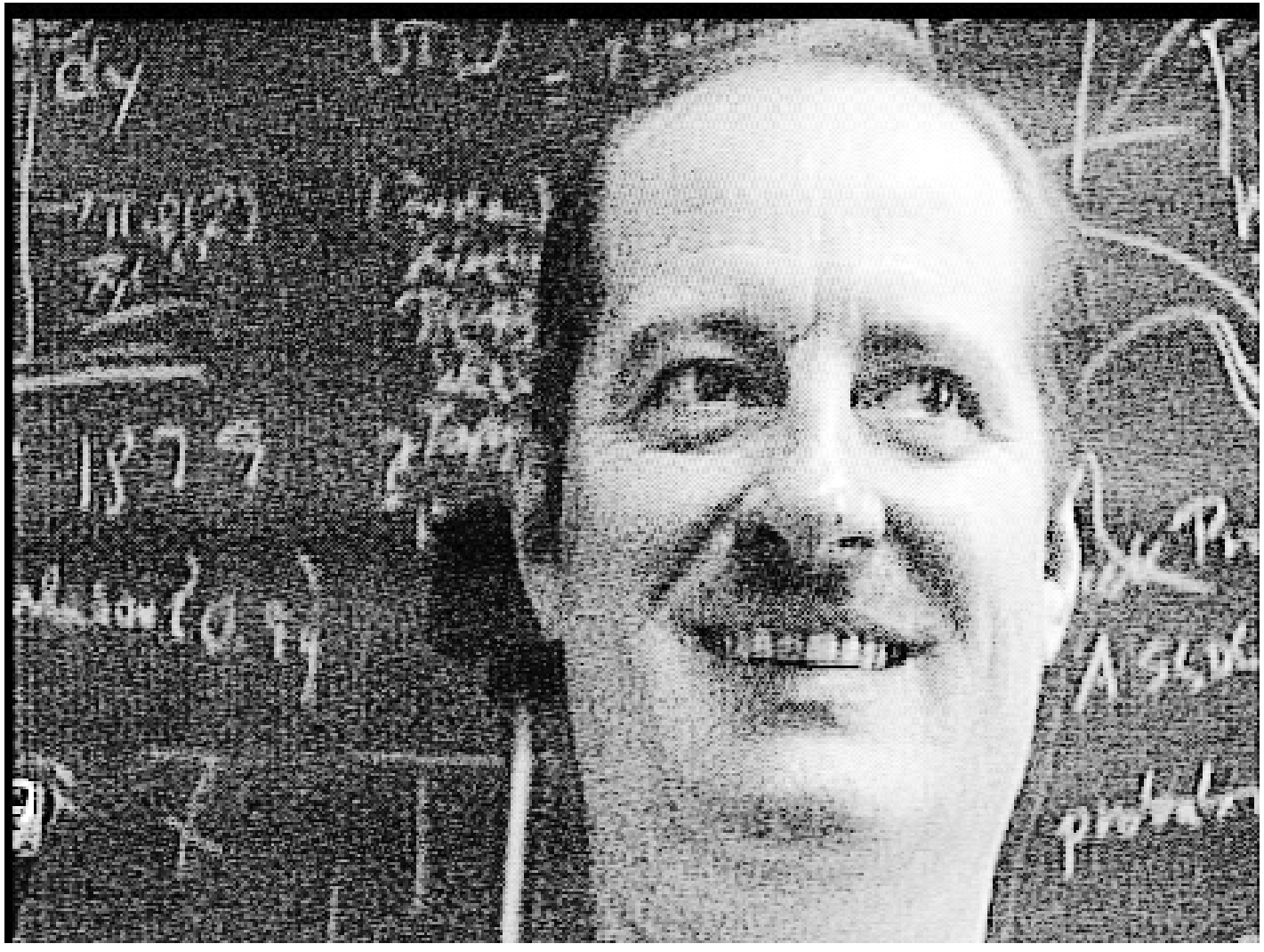
$$f_n^{(\mu)} = f_{max} \cdot \mathrm{sgn}(f_n) \cdot \frac{\log(1+\mu\frac{|f_n|}{f_{max}})}{\log(1+\mu)}, \quad \mu = 100, \ldots, 500$$



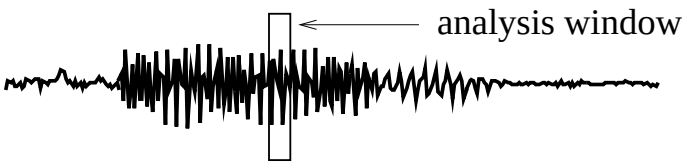
Normalizace



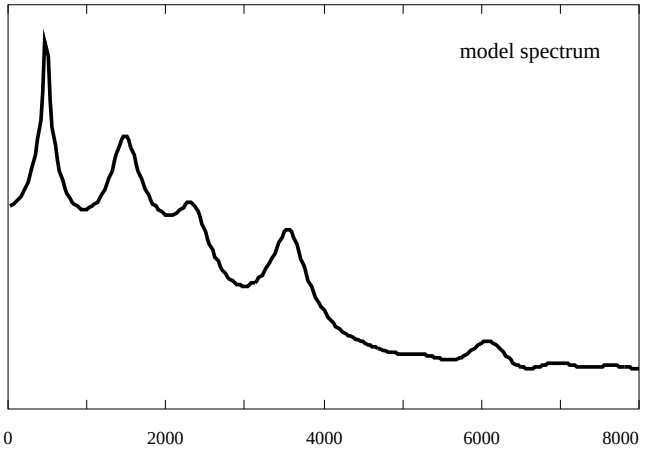
Normalizace



Akustické příznaky



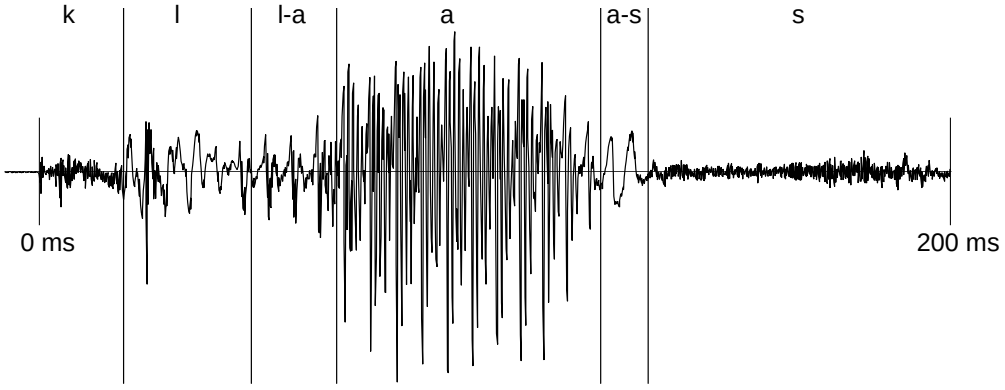
FFT

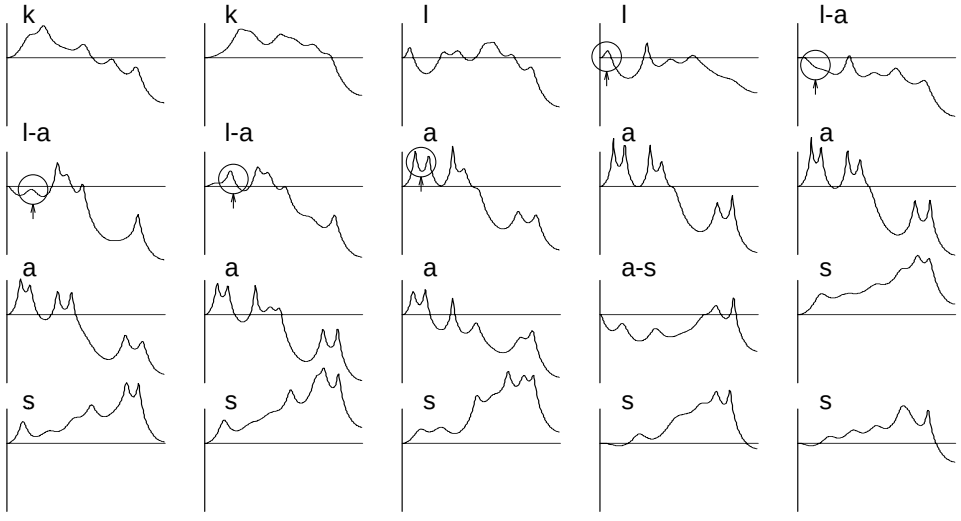


cepstrum

$$\begin{bmatrix} -0.986 \\ 1.000 \\ \vdots \\ -0.333 \end{bmatrix}$$

Was kostet eine Rückfahrkarte zweiter] **Klass** [e nach Hamburg?





Feature Extraction

Acoustic features, computed typically every 10 msec:

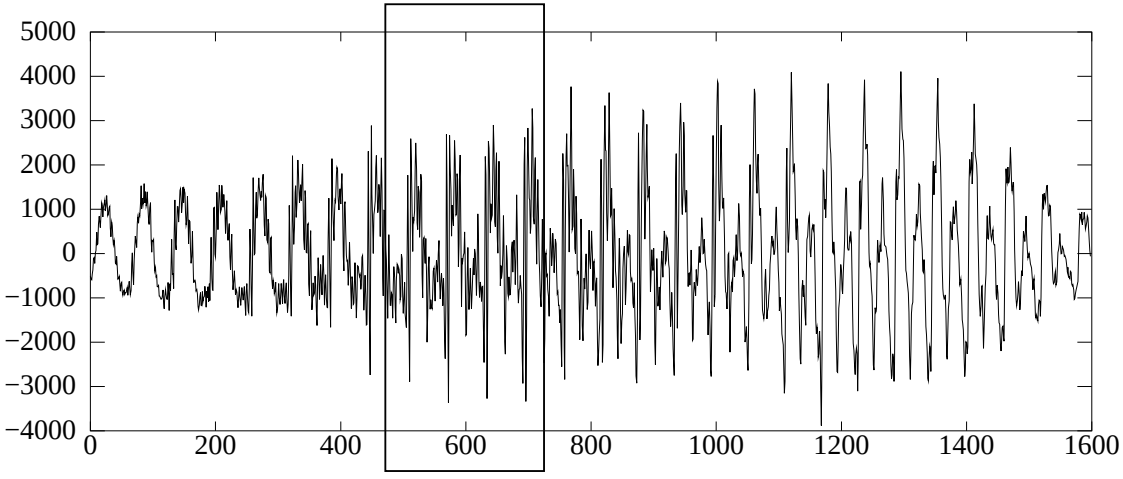
- mel cepstrum coefficients,
  - take 10 to 30 msec (80 – 500 speech samples)
  - apply Hamming window
  - compute **d**iscrete **f**ourier **t**ransform (DFT)
  - smooth spectrum, i.e. integrate over mel scaled bands
  - take logarithm
  - apply inverse DFT or cosine transform
    (separation of source and filter)
- subtract long term average (cepstral mean subtraction, i.e. preprocessing for channel adaptation)
- add energy, pitch, first (and second) order derivative(s)

$\Rightarrow$ observation $\boldsymbol{O}$: sequence of $\boldsymbol{c} \in \mathbb{R}^n$,
$n \approx 20$, i.e. reduction by one order of magnitude

Krátkodobá analýza

Účel: popis spektrální sestavy řečového signálu

- Problém: Diskrétní Fourierova transformace je vhodná jen pro periodické signály.
- Řečový signál není periodický, protože se promění v časovém průběhu.
- Pozorování: Pro velmi krátké časové úseky jsou řečové signály přibližně stacionární, tj. spektrální soustava je pro malý časový interval neměnná:



Krátkodobá analýza

Postup:

- vyřízni v každém čase *m* okno ze signálu
- analyzuj tohle okno pomocí DFT
- implicitně předpokládáme, že se vyříznutý úsek opakuje periodicky: periodické pokračování okna řečového signálu

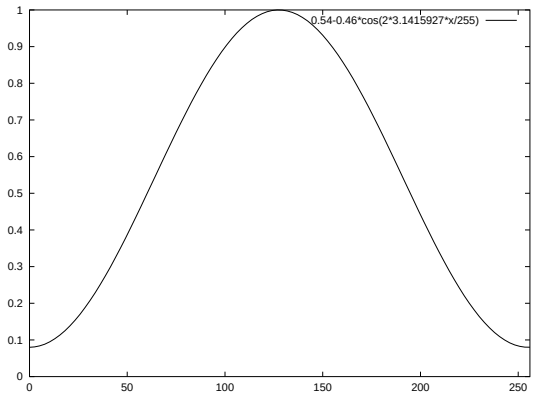
Otázky:

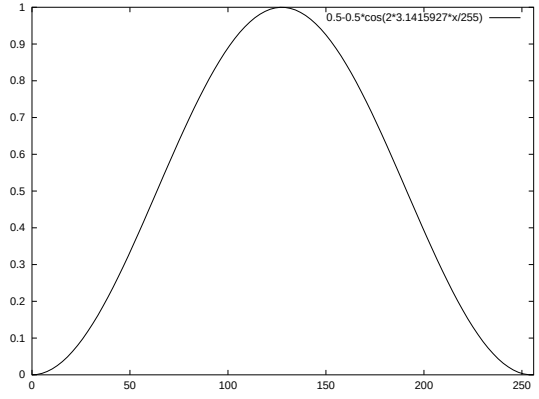
- Vyříznutí odpovídá násobení řečového signálu s okénkovou funkcí, která se přesně v konkrétním intervalu nerovná 0: Jaký vliv má to na spektrum?
- Jakou okénkovou funkci máme vybrat?
- Jak velké má být okno?
- Mají se okna překrývat a popř. jak moc?

Výběr okénkové funkce

- modifikovaný signál $f^{(m)}$ v čase m:
  $f_n^{(m)} = f_n \cdot w_{m-n}$
- příklady pro často užívané okénkové funkce:

časová funkce	útlum	
$w_n^R = 1$	13 dB	(pravoúhlé okno)
$w_n^M = 0.54 - 0.46\cos(\frac{2\pi n}{N-1})$	43 dB	(Hammingovo okno)
$w_n^N = 0.50 - 0.50\cos(\frac{2\pi n}{N-1})$	32 dB	(Hannovo okno)
$w_n^G = e^{-0.5(\frac{n-N/2}{\sigma N/2})^2}$	58 dB	(Gaußovo okno, $\sigma = 3$)
$w_n^P = 4\frac{n}{N}(1 - \frac{n}{N})$	22 dB	(parabola)





Výběr okénkové funkce

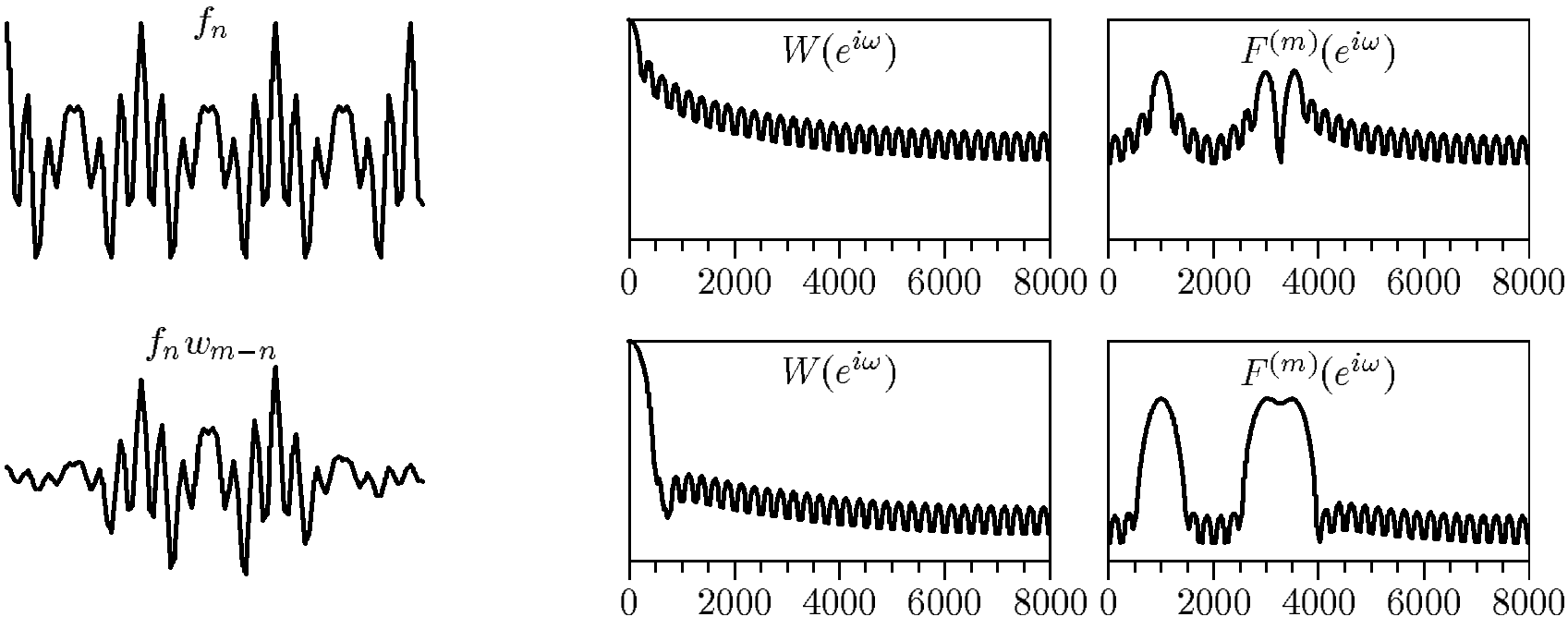
- násobení v časové oblasti = konvoluce ve frekvenční oblasti

→ krátkodobé spektrum:
$F^{(m)}(e^{i\omega}) = \sum_{n=-\infty}^{\infty} f_n w_{m-n} e^{-i\omega n} =$
$\frac{1}{2\pi} \int_{-\pi}^{\pi} W(e^{-i\phi}) e^{-i\phi m} F(e^{i(\omega-\phi)}) d\phi$
kde $F(e^{i\omega})$ a $W(e^{i\omega})$ jsou Fourierovy obrazy funkcí f_n a w_n

- problém: $W(e^{i\omega})$ rozmaže originální spektrum! → okénková funkce má být koncentrovaná kompaktně kolem oblasti $\omega = 0$, v časové oblasti má ale také být kompaktní → kompromis

Výběr okénkové funkce

pravoúhelníkové okno (nahoře) ↔ Hammingovo okno (dole)



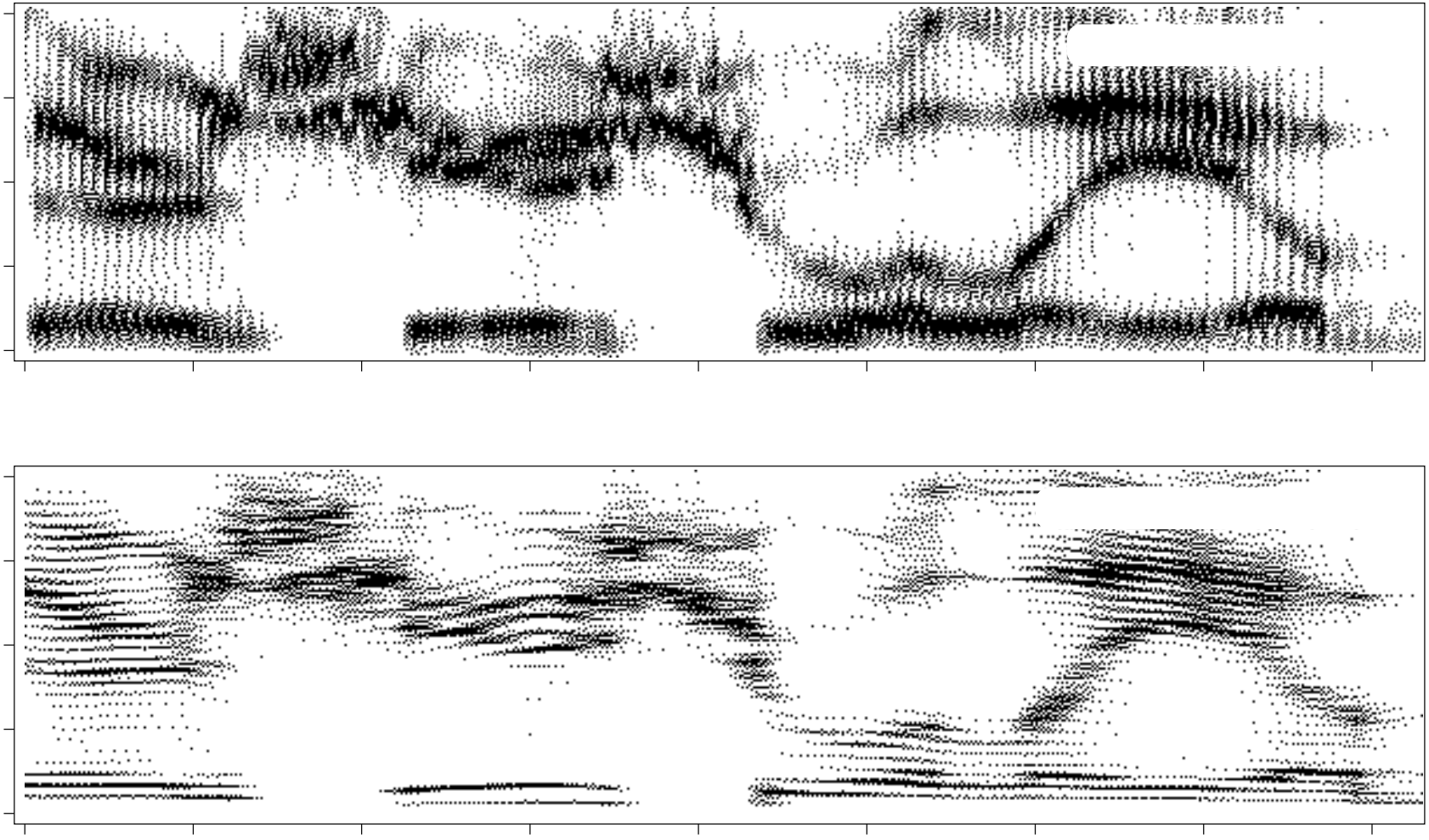
Výběr velikosti okna

- Čím větší je okno v časové oblasti, tím větší je frekvenční rozlišování DFT.
  Pro okna s velikostí 256 vzorkovacích hodnot je frekvenční rozlišení DFT u $f_A = 16\,\text{kHz}$; $16000/256 = 62,5\,\text{Hz}$;
  pro okna s velikostí 64 vzorkovacích hodnot je frekvenční rozlišení DFT jen $16000/64 = 250\,\text{Hz}$.
- Když je okno příliš velké, řečový signál v okně už není stacionární.
- Princip neurčitosti: Čím lepší je časové rozlišení, tím horší je frekvenční rozlišení a naopak.
- Kompromis: typický posuv 10 ms, šířka okna 25 ms

Spektrogramy

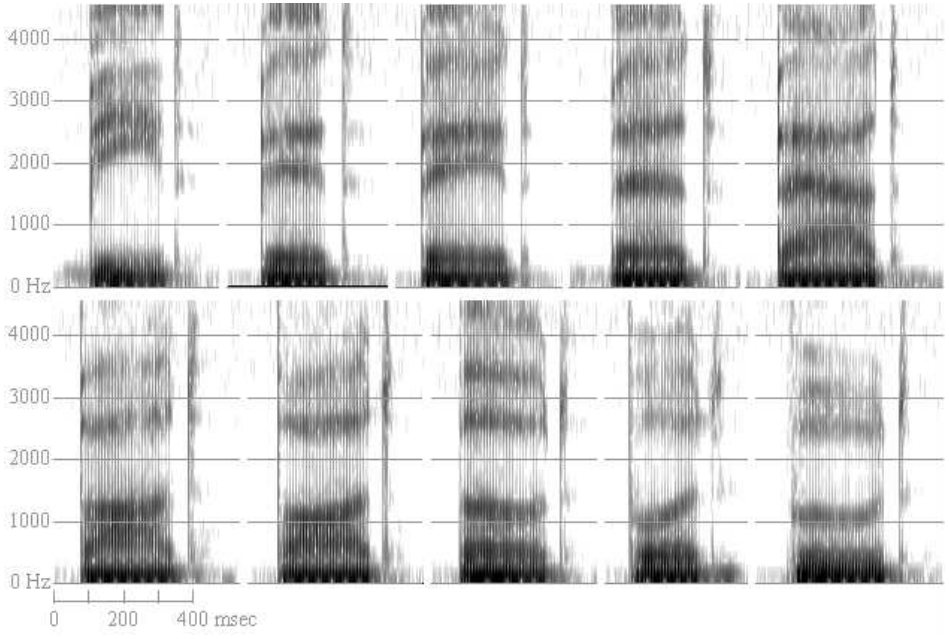
- ignoruj fázi signálu $\rightarrow$ **kvadradické spektrum** $|F(e^{i\omega})|^2$
- čas se zobrazuje na ose x, frekvence (Hz) na ose y; intenzita je reprezentována zčernáním nebo barvou
- širokopásmový spektrogram:
  - malé frekvenční rozlišení & velké časové rozlišení
  - svislé pruhy s rozestupem základní periody
  - detekce krátkých fáz ploziv
- úzkopásmový spektrogram:
  - velké frekvenční rozlišení & malé časové rozlišení
  - vodorovné pruhy s rozestupem základní frekvence
  - rozlišování blízko u sebe ležících formantů

Spektrogramy



Co vidíme na spektrogramu? Formanty.

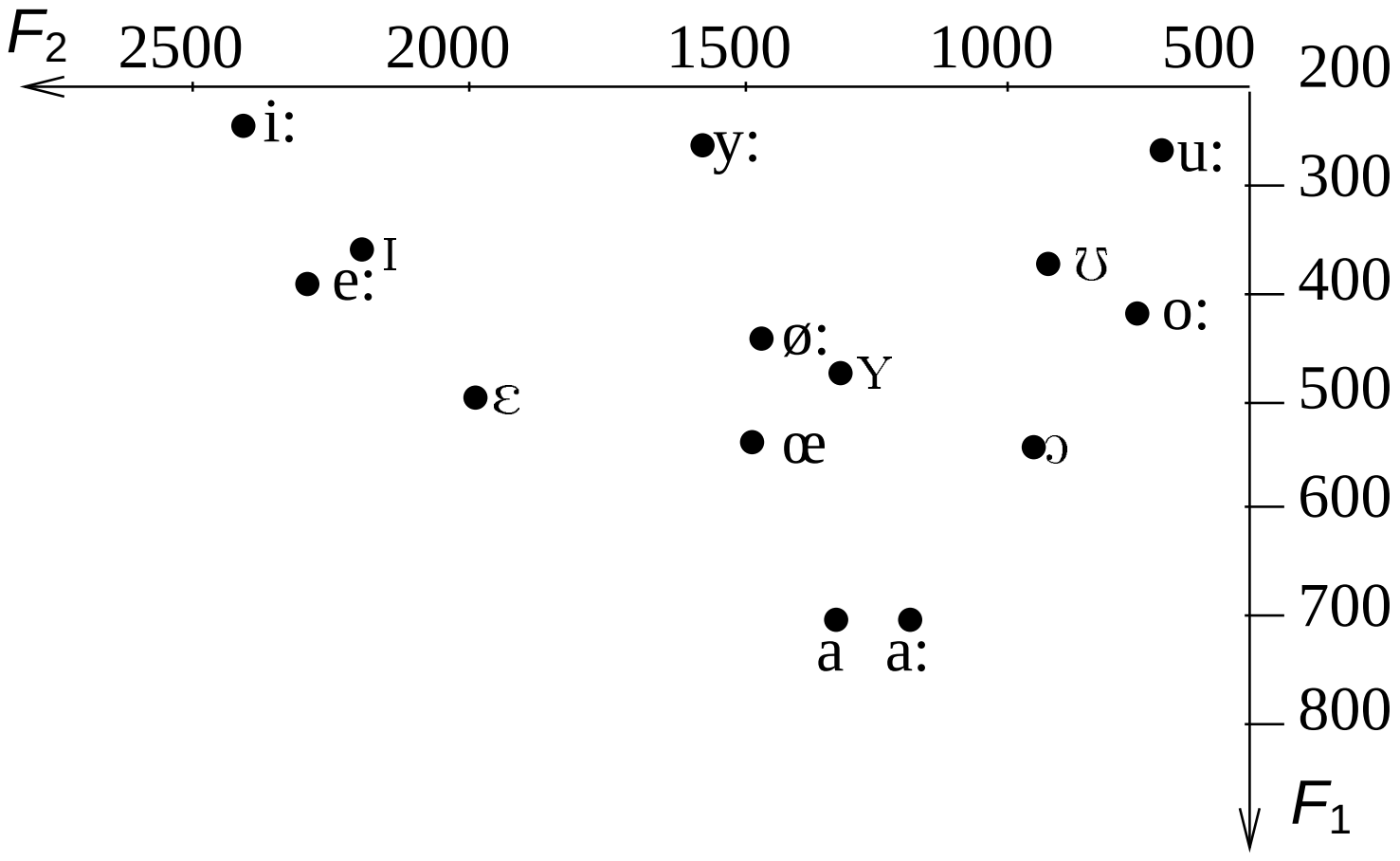
- úseky ve spektru, ve kterých frekvence mají obzvlášť silnou amplitudu
- v samohláskách; poloha formantů charakterizuje různé samohlásky
- např. samohlásky v American English (kontext je vždycky b, d; od vlevo nahoře do vprava dole): *bead*, *bid*, *bade*, *bed*, *bad*, *bod*, *bawd*, *bode*, *buhd*, *booed*



Formanty: fyzikální tvoření

- Formanty vzniknou resonancí (vlastní kmitočet) hlasového traktu, tj. hlasový trakt zesiluje určité frekvence a filtruje jiné ze signálu.
- Resonance jsou určeny délkou hlasového traktu.

Formanty: klasifikace samohlásek



Co vidíme na spektrogramu?

harmonické (frekvence) prvotní frekvence

- Při periodické aktivaci je poměr mezi dílčími kmity harmonický; tyhle kmity jsou celočíselné násobky prvotní frekvence (základní frekvence).
  Příklad: Při základní frekvenci 150 Hz jsou ve spektru viditelné harmonické u 300, 450, 600, ... Hz.

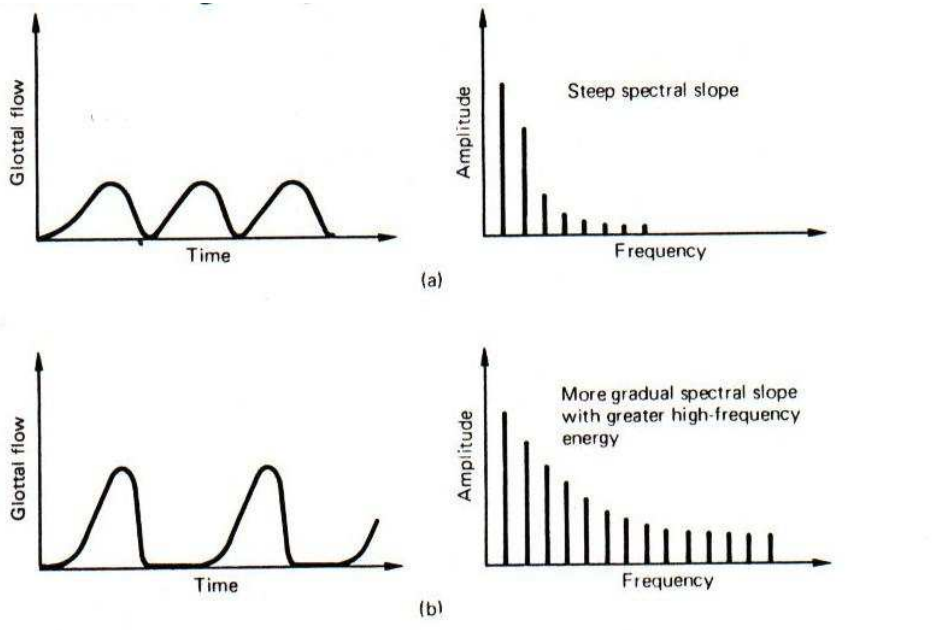
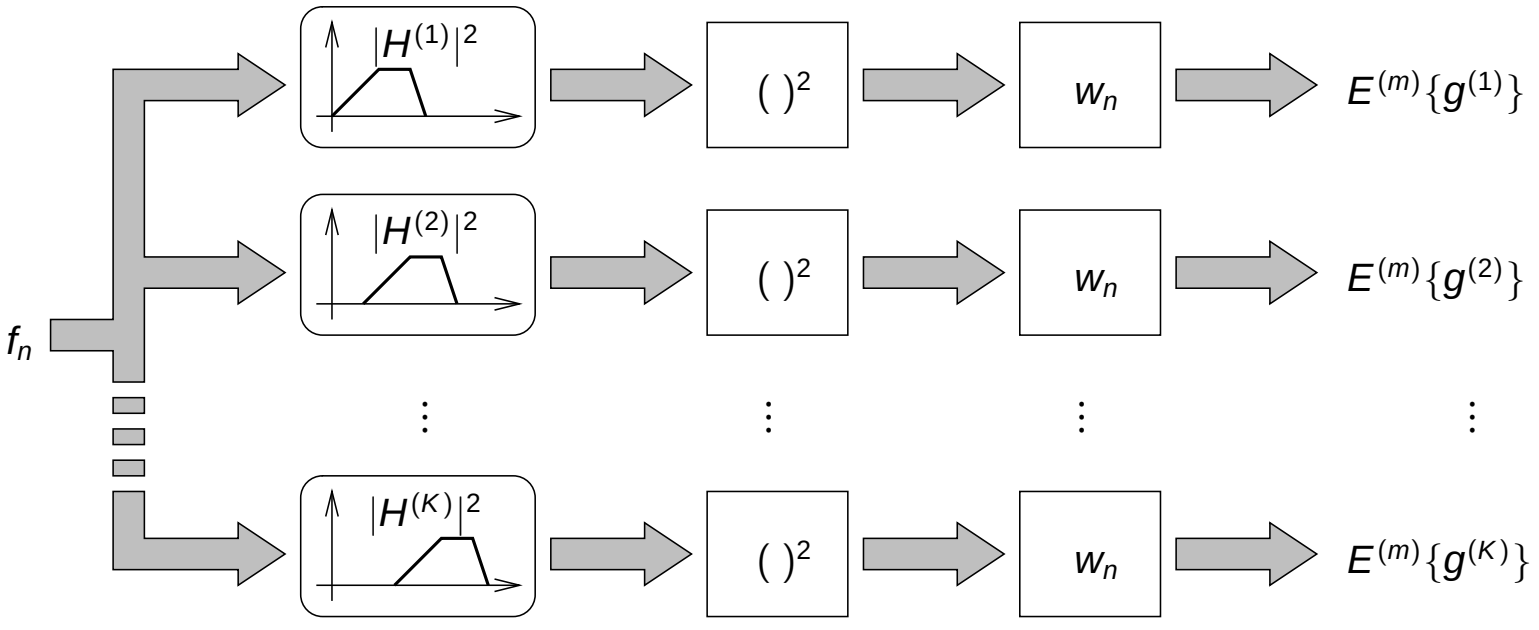


FIGURE 7.11.2 **Phonation airflow (volume velocity) waveforms and their corresponding spectra: (a) breathy voice; (b) bright voice**

Pásmová spektra

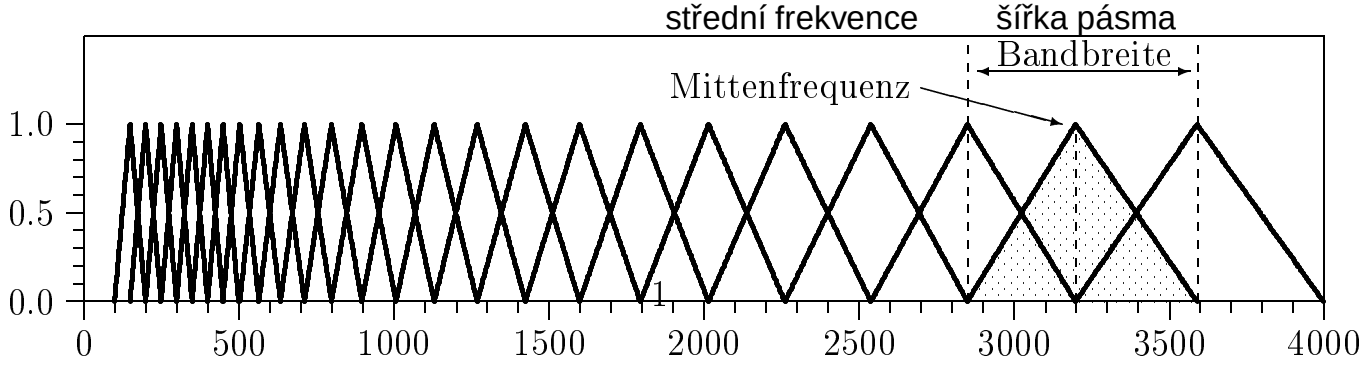
- sluchově orientovaná integrace energie v kritických frekvenčních pásmech je modelována bankou pásmových propustí s identickými okny krátkodobé energie
- vyfiltrování částí signálu f_n filtrem (lineárními systémy) s impulzovými odpověďmi $h_n^{(k)}$ a frekvenčními přenosy $H^{(k)}(e^{i\omega})$



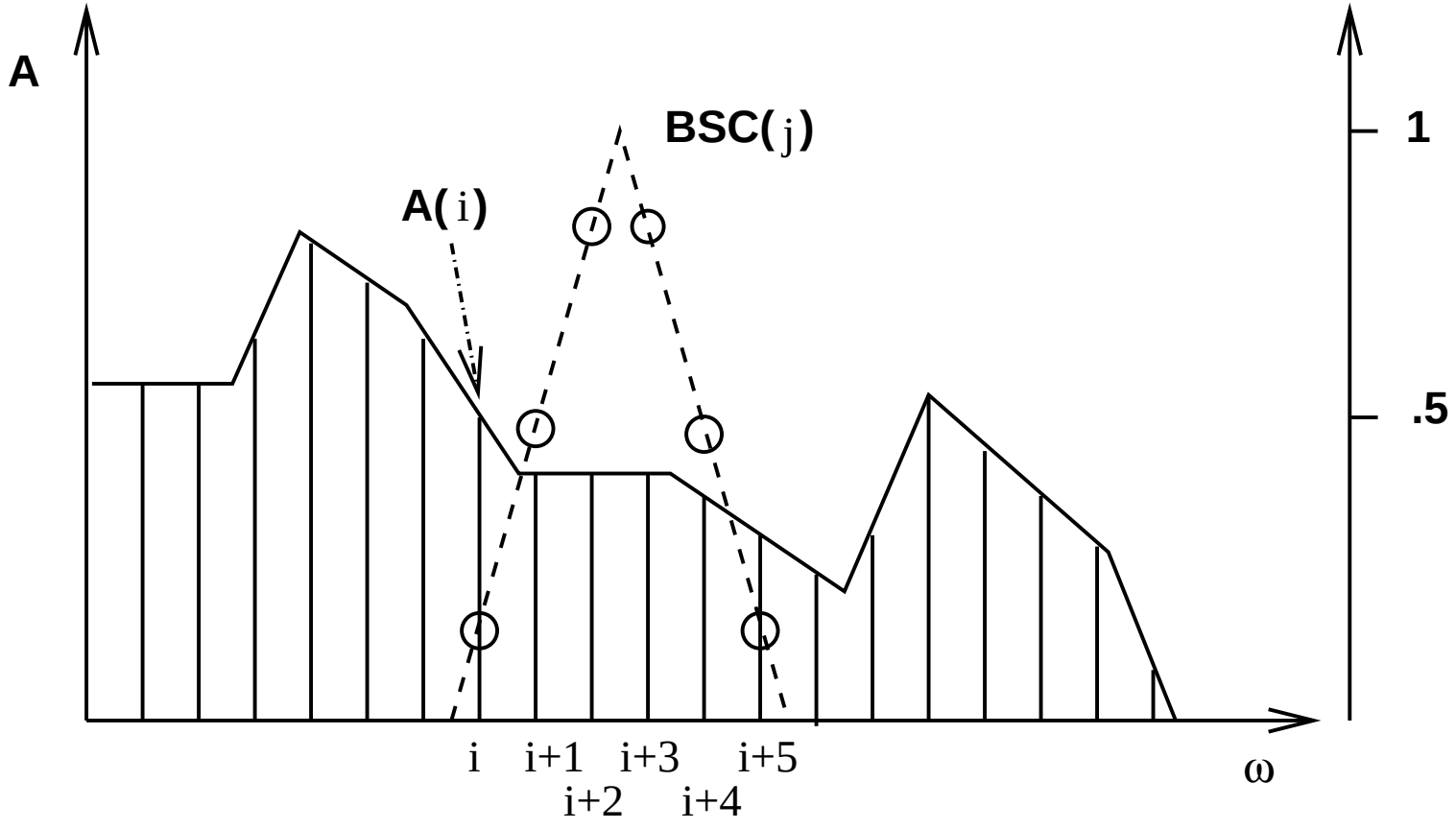
Pásmová spektra

- hledáme: krátkodobé energie $E^{(m)}\{g^{(k)}\}$ filtrovaných komponent signálu $g^{(k)} = f \star h^{(k)}$
- věta o konvoluci $\rightarrow$ dobrá aproximace energií pásmové propusti: $e_k^{(m)} = \sum_{\nu=0}^{N-1} \eta_{k\nu} |F_\nu^{(m)}|^2$ s váhami $\eta_{k\nu} = |H^{(k)}(e^{i2\pi\nu/N})|^2$
- banky filtrů s tvarem trojúhelníku, pravoúhelníku nebo lichoběžníku napodobují rozdělování vzruchu na basilární membráně; leží ekvidistantně na mel- nebo barkově stupnici frekvencí se šířkou pásma 1 bark

$\rightarrow$ **koeficienty mel(ody)-spektra**, např. s 25 frekvenčními skupinami a bankou trojúhelníkových filtrů:



(Mel-)frekvenční pásma



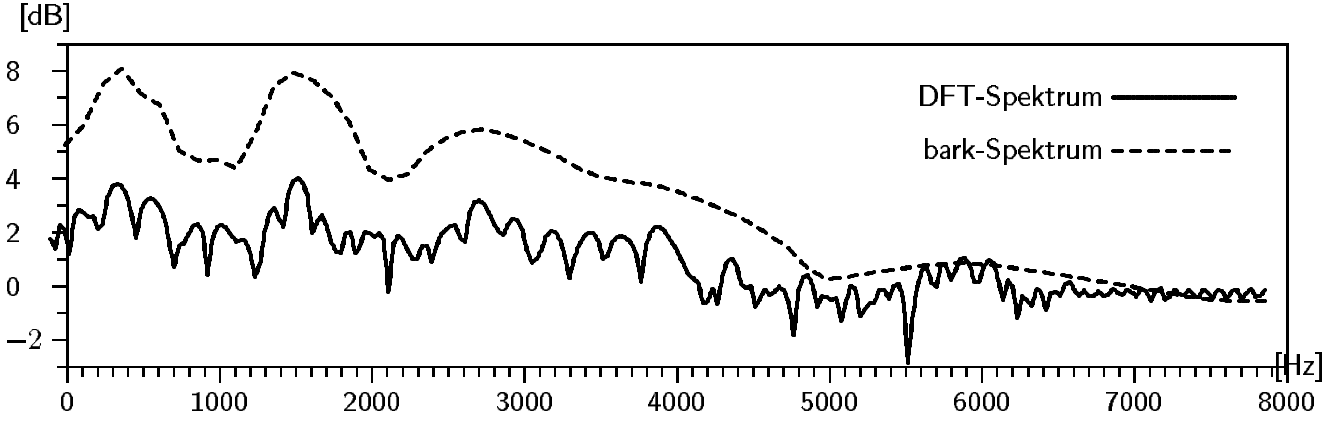
Band spectrum coefficient $BSC(j) =$

$$.15 * A(i) + .5 * A(i+1) + .85 * A(i+2) + .85 * A(i+3) + .5 * A(i+4) + .15 * A(i+5)$$

Mel-spektrum

- 7 trojúhelníkových filtrů kolem středních frekvencí $150, 200, \ldots, 450$ Hz
- 3 oktávy od 500 Hz do 4000 Hz; každá má 6 pásem
- každé pásmo skončí u středních frekvencí jeho sousedních pásem
- průběh spektra je vyhlazen
- harmonická struktura zanikne
- resonance hlasového traktu se objeví

Barkovo spektrum samohlásky ə (schwa; střední frekvence ekvidistantní na frekvenční stupnici):



Cepstrální koeficienty

- východisko: Fantův source-filter model generace řeči
- účel: Odděluj signál buzení od modelu pro hrtanovou záklopku (glottis), hlasový trakt a rty:

$$\begin{aligned} f_n &= e_n \star h_n \\ FT\{f_n\} &= FT\{e_n\} \cdot FT\{h_n\} \\ \log FT\{f_n\} &= \log FT\{e_n\} + \log FT\{h_n\} \\ FT^{-1}\{\log FT\{f_n\}\} &= FT^{-1}\{\log FT\{e_n\}\} + FT^{-1}\{\log FT\{h_n\}\} \end{aligned}$$

komplexní a reálné cepstrum: (homomorfní analýza)
$FT^{-1}\{\log FT\{f_n\}\}$ a $FT^{-1}\{\log |\{FT\{f_n\}\}|\}$
inverzní DFT na logaritmovaném absolutním spektru $\rightarrow$ **cepstrální koeficienty**
$c_q^{(m)} = \frac{1}{N} \sum_{\nu=0}^{N-1} \log |F_\nu^{(m)}| e^{i2\pi\nu q/N}, \quad q = 0, \ldots, N-1$

Cepstrální koeficienty

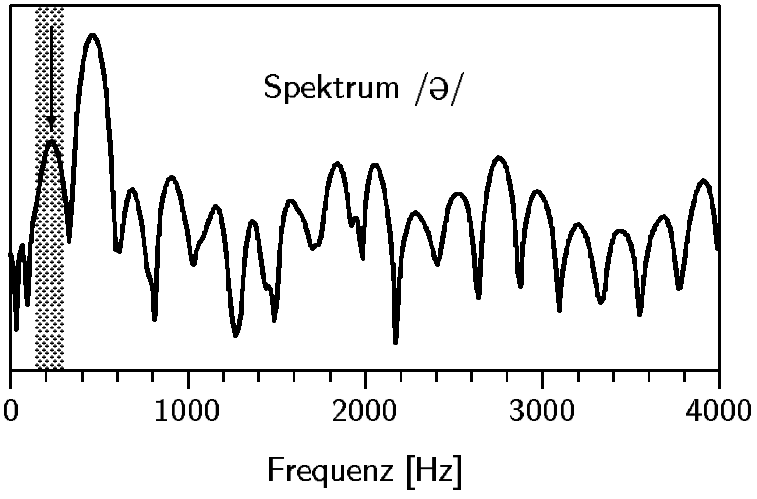
absolutní spektrum a cepstrum jsou reálná a symetrická $\rightarrow c_q^{(m)}$ také $\rightarrow$ **diskrétní kosinová transformace (DCT)** stupně $N/2$

$$c_0^{(m)} = \sqrt{2/N} \sum_{\nu=0}^{N/2-1} \log |F_\nu^{(m)}|$$

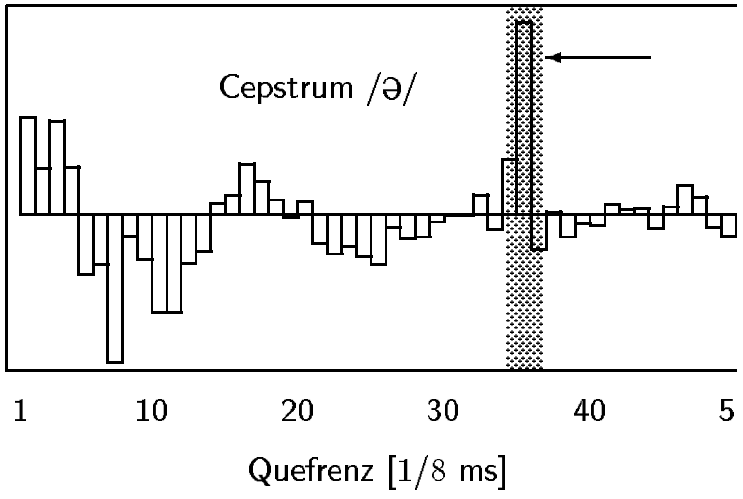
$$c_q^{(m)} = \sqrt{4/N} \sum_{\nu=0}^{N/2-1} \log |F_\nu^{(m)}| \cos \frac{\pi q(2\nu+1)}{N} \quad \text{für } q = 1, \ldots, N/2$$

Cepstrální koeficienty

výkonové spektrum $\log |F_\nu^{(m)}|$ a cepstrální koeficienty $c_q^{(m)}$ samohlásky



frekvence



kvefrence

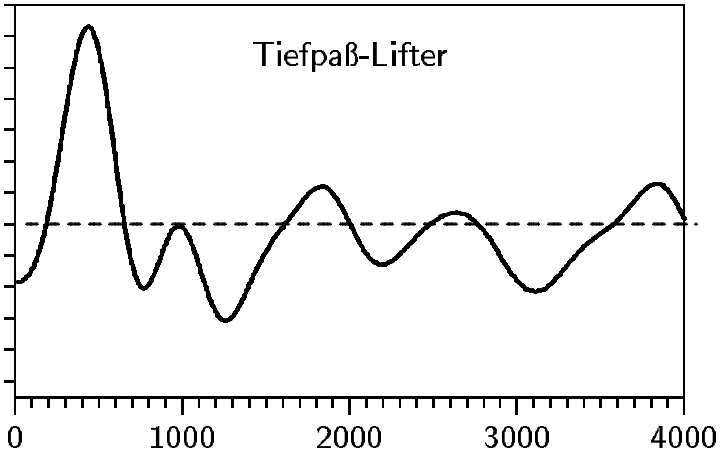
Cepstrální koeficienty

- resonance ve výkonovém spektru u 500 Hz, 2000 Hz a 2700 Hz; přikrývány špičkami harmonických (frekvencí)
- cepstrum je „spektrum spektra“ $\rightarrow$ vrchol u cepstrálního koeficientu $c_q^{(m)}$ poukazuje na spektrální kmit
- pomalé části $\Rightarrow$ dolní cepstrální koeficienty
  rychlé části, zejména harmonické $\Rightarrow$ horní cepstrální koeficienty
- tady u *kvefrence* 35 jednotek na $1/8$ [ms] (f_A = 8000Hz)
  $35 \cdot 1/8 = 4.375$ [ms]
  $1/0.004375 = 228.6$ [Hz] (= základní frekvence)

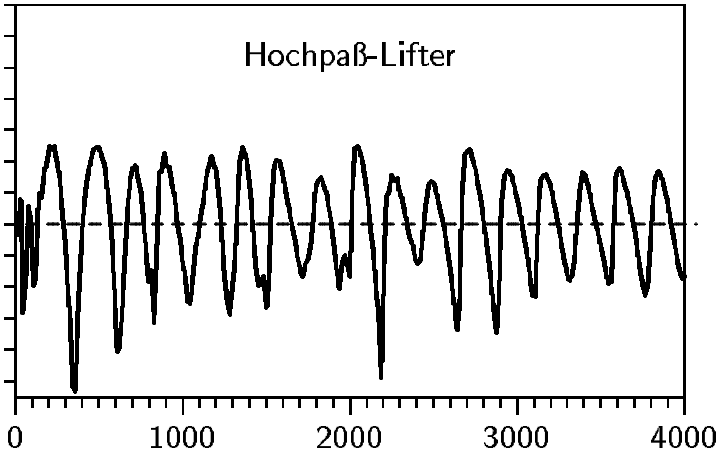
Cepstrální koeficienty

1. jen dolní koeficienty jsou důležité
2. části s vyšší kvefrencí: stanovit na nulu
3. zpětná transformace do spektrální oblasti
   $\hat{C}_\nu^{(m)} = DFT\{\hat{c}_q^{(m)}\}$

liftering = [1]+[2] analogicky k filtraci prostřednictvím dolní propusti



liftr typu dolní propust

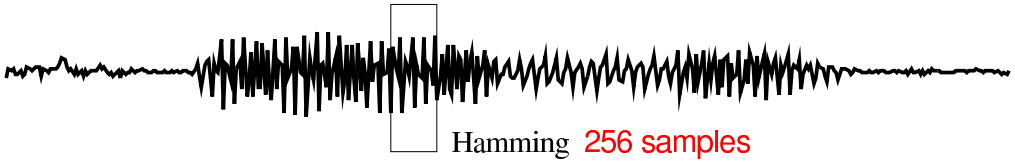


liftr typu horní propust

Mel-cepstrum

- Mel-frequency ceptral coefficients (MFCC): nejčastěji užívané příznaky v automatickém rozpoznávaní řeči
- kosinová transformace logaritmovaného mel-spektra; jako nahoře:
  $e_k^{(m)} = \sum_{\nu=0}^{N-1} \eta_{k\nu} |F_\nu^{(m)}|^2$
  kosinová transformace $e_k^{(m)}$:
  $c_q^{(m)} = \sum_{k=1}^{K} \log e_k^{(m)} \cos \frac{\pi q(2k+1)}{2K}$, $q = 1, 2, \ldots$
- důležité: mel-spektrum neobsahuje harmonické $\rightarrow$ tady žádný liftering!
- kosinová transformace dekoreluje příznaky podobně jako analýza hlavních komponent (principal component analysis, PCA)
- běžně je koeficient $c_0^{(m)}$ nahrazován mírou hlasitosti nebo mírou subjektivní hlasitosti
- logaritmování způsobí kompresi stupnice energie
- při velmi malém SNR je logaritmus často nahráván kořenovou funkcí $\rightarrow$ root-mel-cepstrum:
  $c_q^{(m)} = \sum_{k=1}^{K} (e_k^{(m)})^{\frac{1}{3}} \cos \frac{\pi q(2k+1)}{2K}$, $q = 1, 2, \ldots$

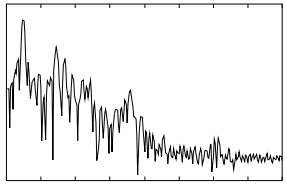
Vypočítání příznaků mel-cepstra



Hamming 256 samples

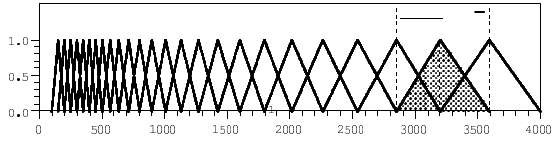
$|DFT|^2$

power spectrum



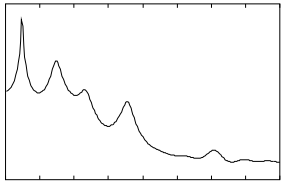
128 coefficients

Mel



25 triangles

Mel spectrum



25 coefficients

log

cosine transform

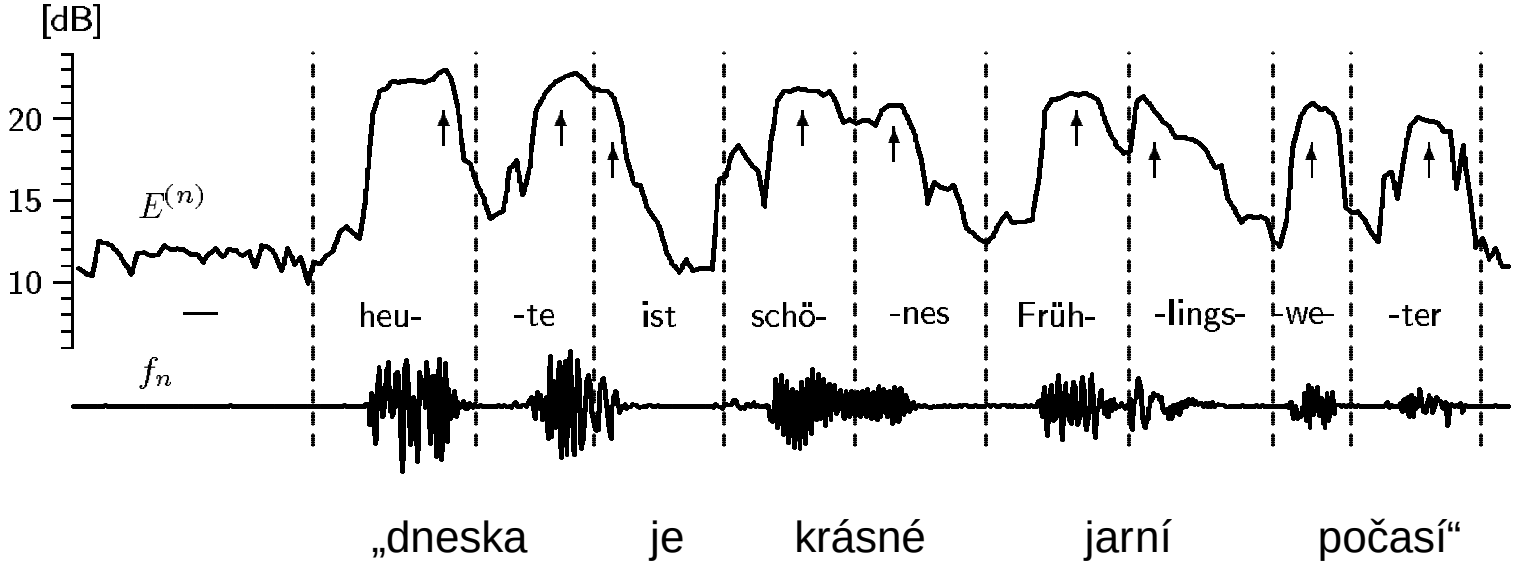
take only the first 12 coefficients

Mel Cepstrum

one 12–dimensional feature vector

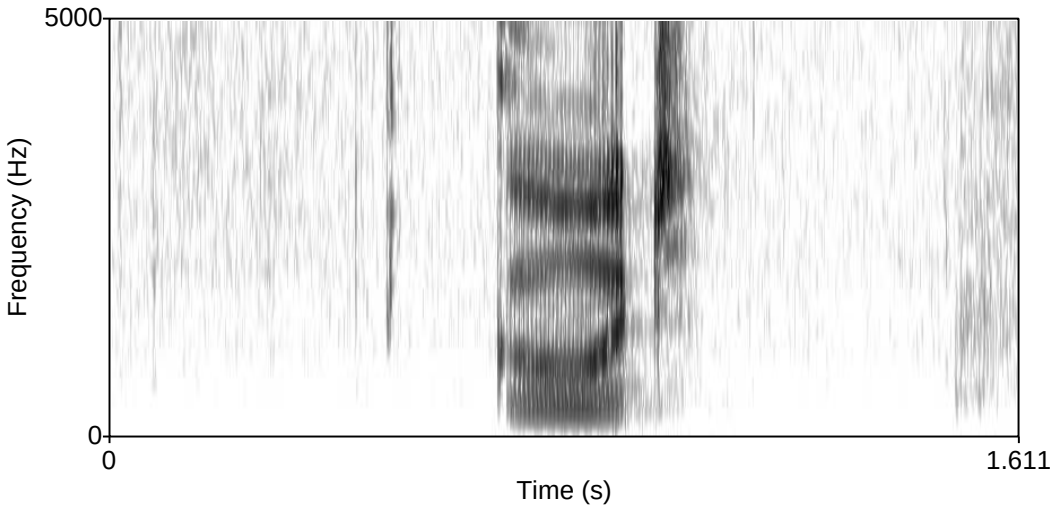
Krátkodobá energie

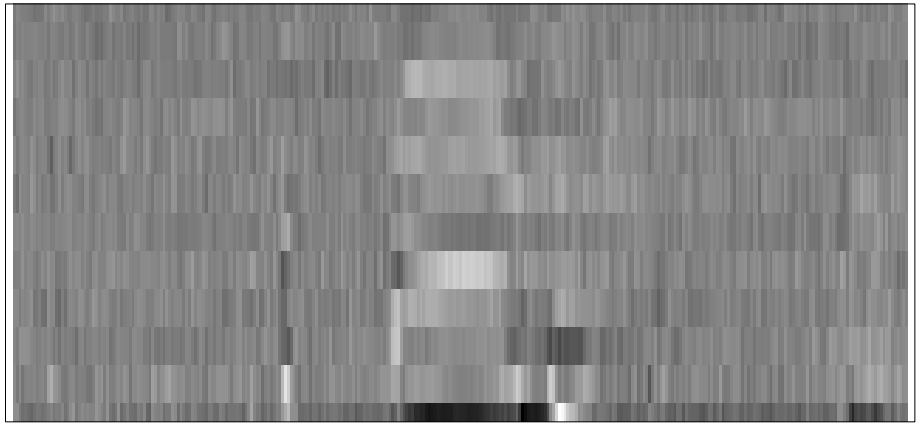
- „Standardní“ mel-frekvenční cepstrální koeficienty (MFCC) jsou často doplněny krátkodobou energií jako příznak hlasitosti:
  $E^{(m)} = \sum_{n=-\infty}^{\infty} |f_n^{(m)}|^2 = \sum_{n=-\infty}^{\infty} |f_n w_{m-n}|^2 , \quad m = 0, \pm 1, \pm 2, \ldots$
  s oknem délky N, $\alpha_n = w_{-n}^2 \geq 0$:
  $E^{(m)} = \sum_{n=0}^{N-1} \alpha_n |f_{m+n}|^2$



Části řeči a ticha jsou viditelné v profilu energie, hrubá detekce hranic a jáder slabik je možná.

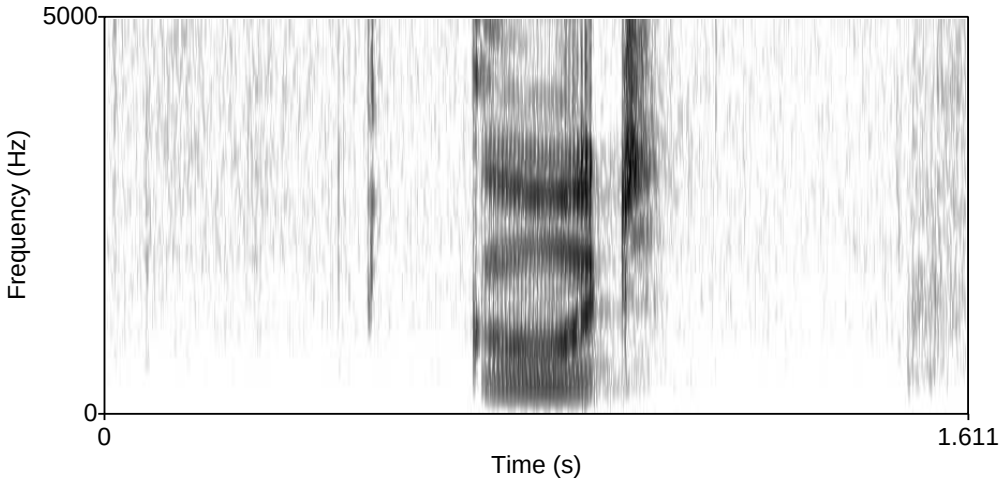
Spektrogram a 12 MFCC





Mel-cepstrum

Spektrogram a mel-cepstrální koeficient c_0:



Mel-cepstrum

Hlásky se rozlišují podle způsobu, jak jsou reprezentovány prostřednictvím MFCC (Mel-Frequency Cepstral Coefficients):

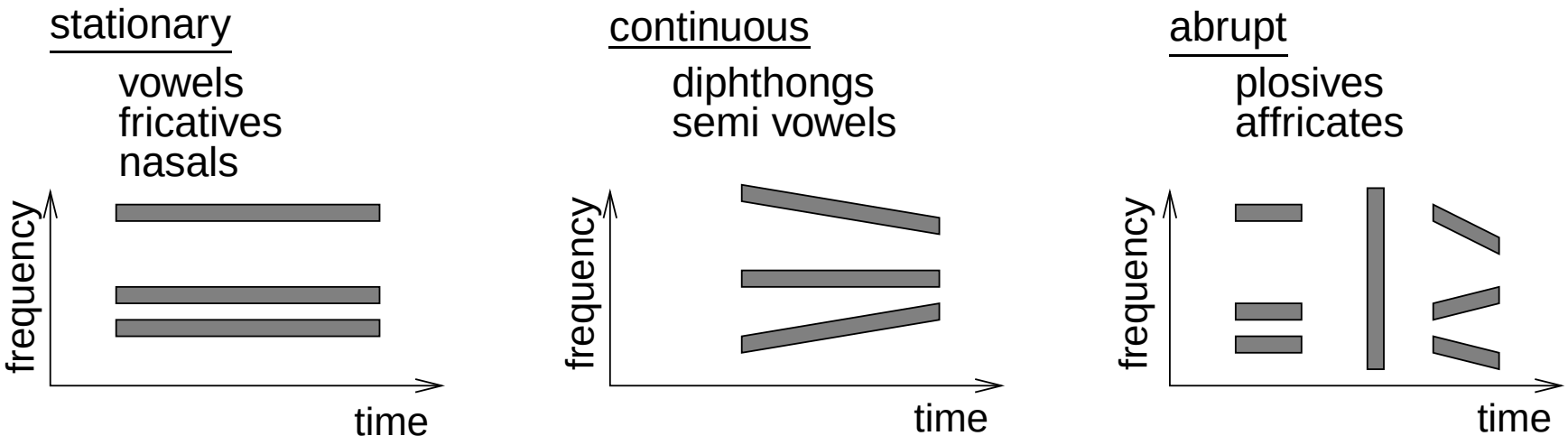
- místo v příznakové oblasti
- rozsah v příznakové oblasti
- dynamické chování přes několik časových oken

$\rightarrow$ dynamické chování se může popisovat dalšími příznaky, které obsahují informace o změnách MFCC v časovém průběhu

- Dlouhodobé spektrální změny se musí rozlišovat od krátkodobých.
- Pro identifikaci hlásek jsou důležité jen náhlé, krátkodobé změny.

Časové změny řečového signálu

Stacionární, souvislé (kontinuální) a náhlé (abruptní) spektrální chování:



- statické příznaky: momentální vlastnosti řečového signálu
- dynamické příznaky: časový průběh krátkodobých parametrů
  - v modelu můžeme rychlé změny zohledňovat
  - pomalé změny mohou být modelem ignorovany

Dynamické příznaky

- východisko: d-dimenzionální, statický, krátkodobý, příznakový vektor m-tého okna analýzování:
  $\boldsymbol{x}^{(m)} = (x_1^{(m)}, \ldots, x_d^{(m)})^\top$
- nyní: analýza příznakových řad
  $\ldots, x_k^{(m-\tau)}, \ldots, x_k^{(m)}, \ldots, x_k^{(m+\tau)}, \ldots, \quad k = 1, \ldots, d$
- nejjednodušší možnost: statický příznakový vektor je rozšířen sousedními příznakovými vektory:
  $\boldsymbol{y}^{(m)} = (\boldsymbol{x}^{(m)}, \boldsymbol{x}^{(m-\tau)})$ anebo $\boldsymbol{y}^{(m)} = (\boldsymbol{x}^{(m-\tau)}, \boldsymbol{x}^{(m)}, \boldsymbol{x}^{(m+\tau)})$
- jiné metody:
  „standard“: časové derivace (mel-)cepstrálních parametrů
  často: filtrace (např. RASTA)
  málo: Fourierova transformace
- většinou statický příznakový vektor je rozšířen dynamickými:
  $\boldsymbol{y}^{(m)} = (\boldsymbol{x}^{(m)}, \delta\boldsymbol{x}^{(m)})$

Derivace (mel-)cepstrálních parametrů

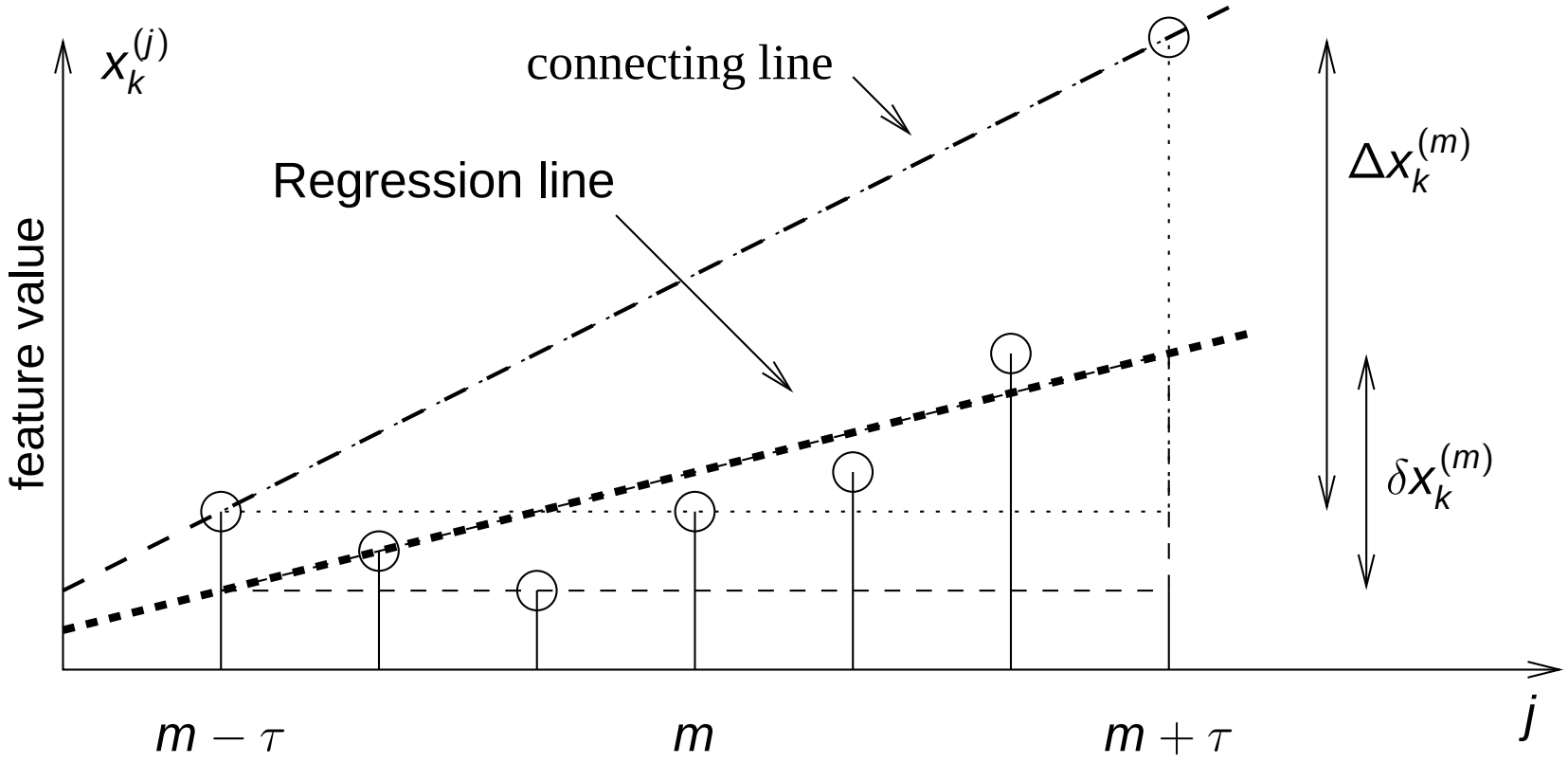
- model sklonu příznaku $\boldsymbol{x}_k^{(m)}$ v čase m
- $\boldsymbol{x}_k^{(m)}$ je přitom diskrétní vzorkovací hodnota příznaku $\tilde{x}_k(t)$ (časově kontinuální)

diskrétní aproximace sklonu pro průběh příznaku v intervalu $m - \tau \leq j \leq m + \tau$ pomocí

- prvních diferencí
  $\Delta x_k^{(m)} = x_k^{(m+\tau)} - x_k^{(m-\tau)}$ pro všechna k, m
- sklonu regresní přímky
  $$\delta x_k^{(m)} = \frac{\sum_{j=-\tau}^{\tau} j \cdot x_k^{(m+j)}}{\sum_{j=-\tau}^{\tau} j^2}$$ pro všechna k, m

Derivace (mel-)cepstrálních parametrů

Metody regrese a tvoření diferencí mají různé výsledky:



Derivace vyššího řádu

- často: derivace druhého řádu
- jednoduchá možnost: aproximace derivací vyššího řádu opakovanou aplikací lineární derivace, např. druhé diference
  $\Delta^2 x_k^{(m)} = \Delta x_k^{(m+\tau)} - \Delta x_k^{(m-\tau)}$ pro všechna k, m
- anebo: r-tá derivace z koeficientů **regresních polynomů** r-tého řádu

$$\delta^r x_k^{(m)} = \frac{\sum_{j=-\tau}^{\tau} \rho_r(j, 2\tau+1) \cdot x_k^{(m+j)}}{\sum_{j=-\tau}^{\tau} \rho_r^2(j, 2\tau+1)}$$ pro všechna k, m

ρ_r pocházejí z ortogonálního systému polynomů, např.:

$$\begin{aligned} \rho_0(t, \alpha) &= 1 \\ \rho_1(t, \alpha) &= t \\ \rho_2(t, \alpha) &= t^2 - \frac{1}{12}(\alpha^2 - 1) \end{aligned}$$

„Standardní“ příznakový vektor pro rozpoznávání řeči

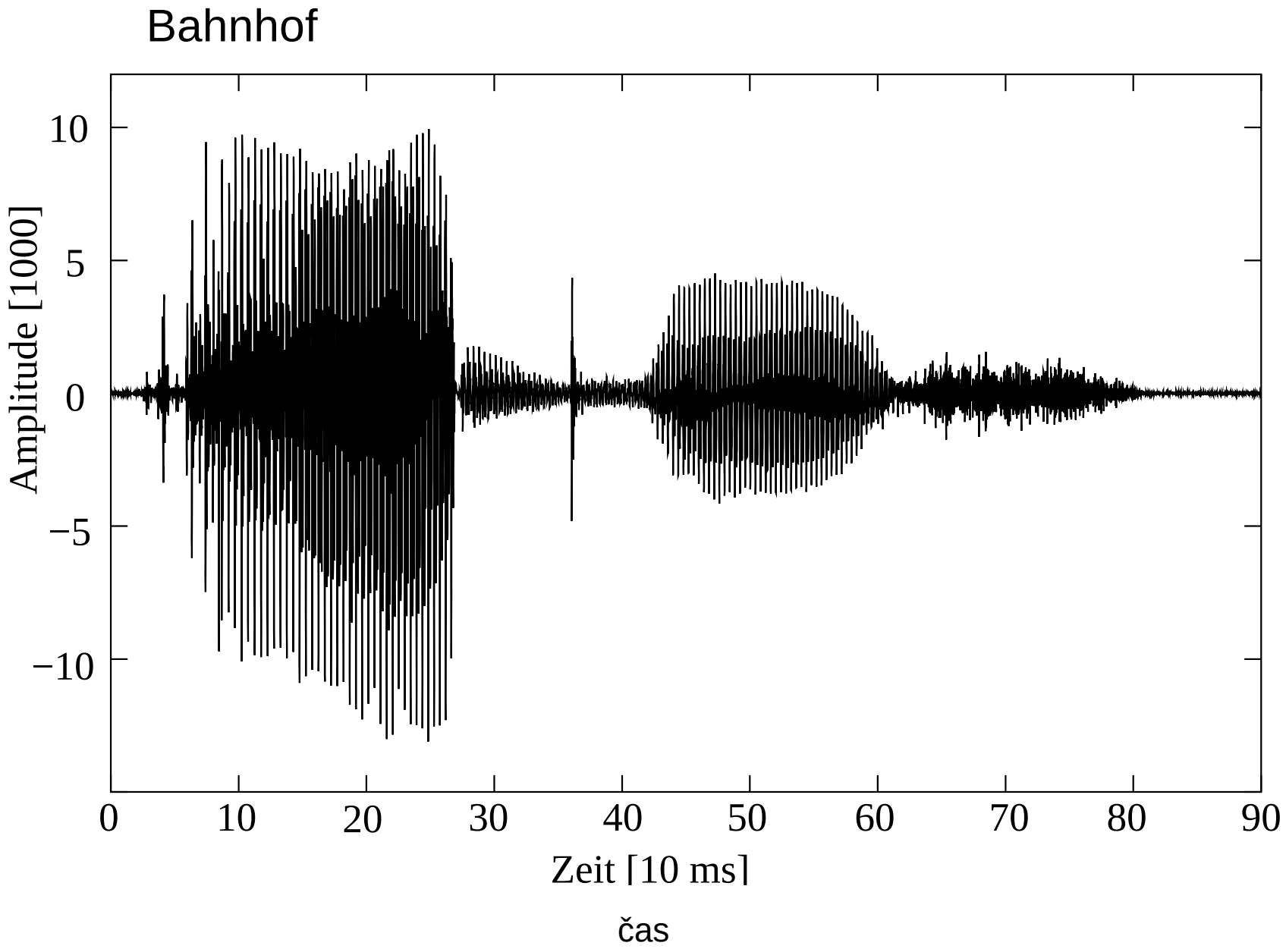
- krátkodobá energie
- 12 mel-cepstrálních koeficientů; první (c_0) je nahrazen krátkodobou energií
- 1. derivace těch 12 koefizientů $\rightarrow$ 12 koefizientů 1. řádu
- 2. derivace těch 12 koefizientů $\rightarrow$ 12 koefizientů 2. řádu

$\rightarrow$ 36 koeficientů

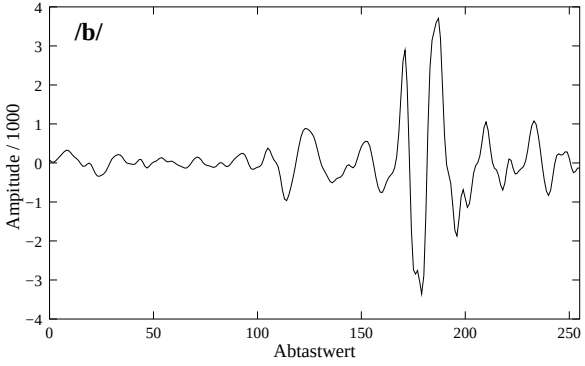
Existuje hodně variací tohoto schématu, např. jen 24 koeficientů (žádná 2. derivace) nebo krátkodobá energie a c_0 jsou užívány současně.

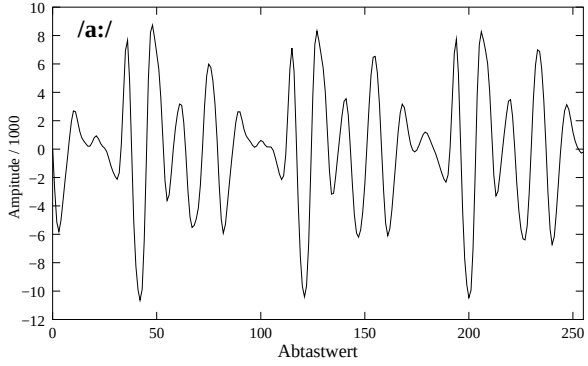
- často redukce dimenze pomocí problémově závislého rozvoje do řady, přitom dekorelace, redukce dimenze, ale malá ztrata informací

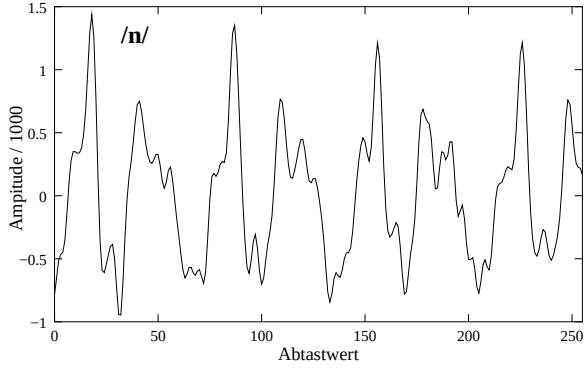
„Bahnhof“ („nádraží“)

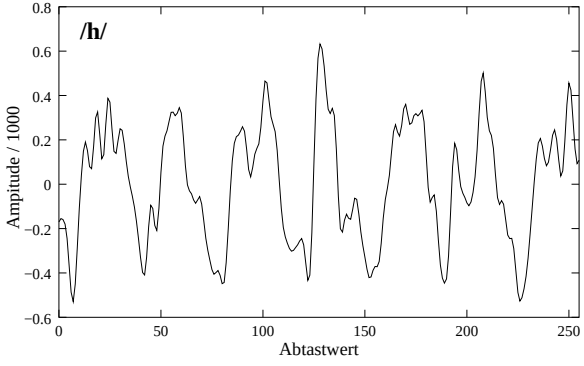


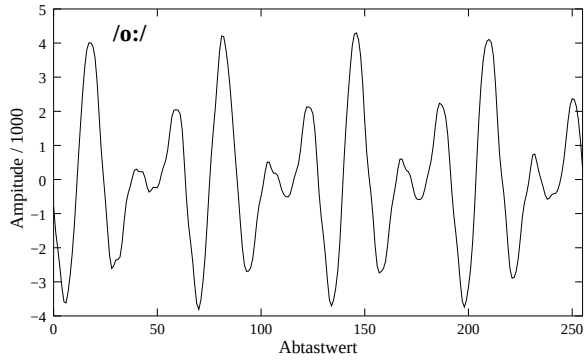
Hlásky ve slově „Bahnhof“

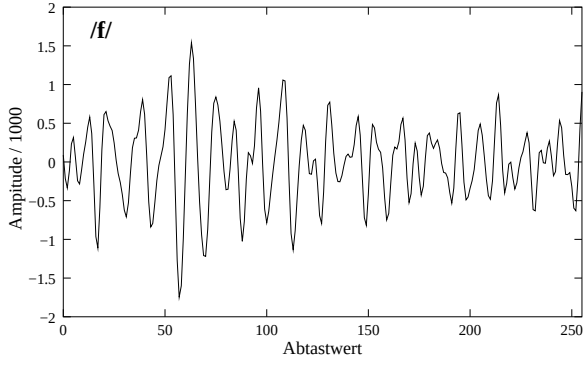






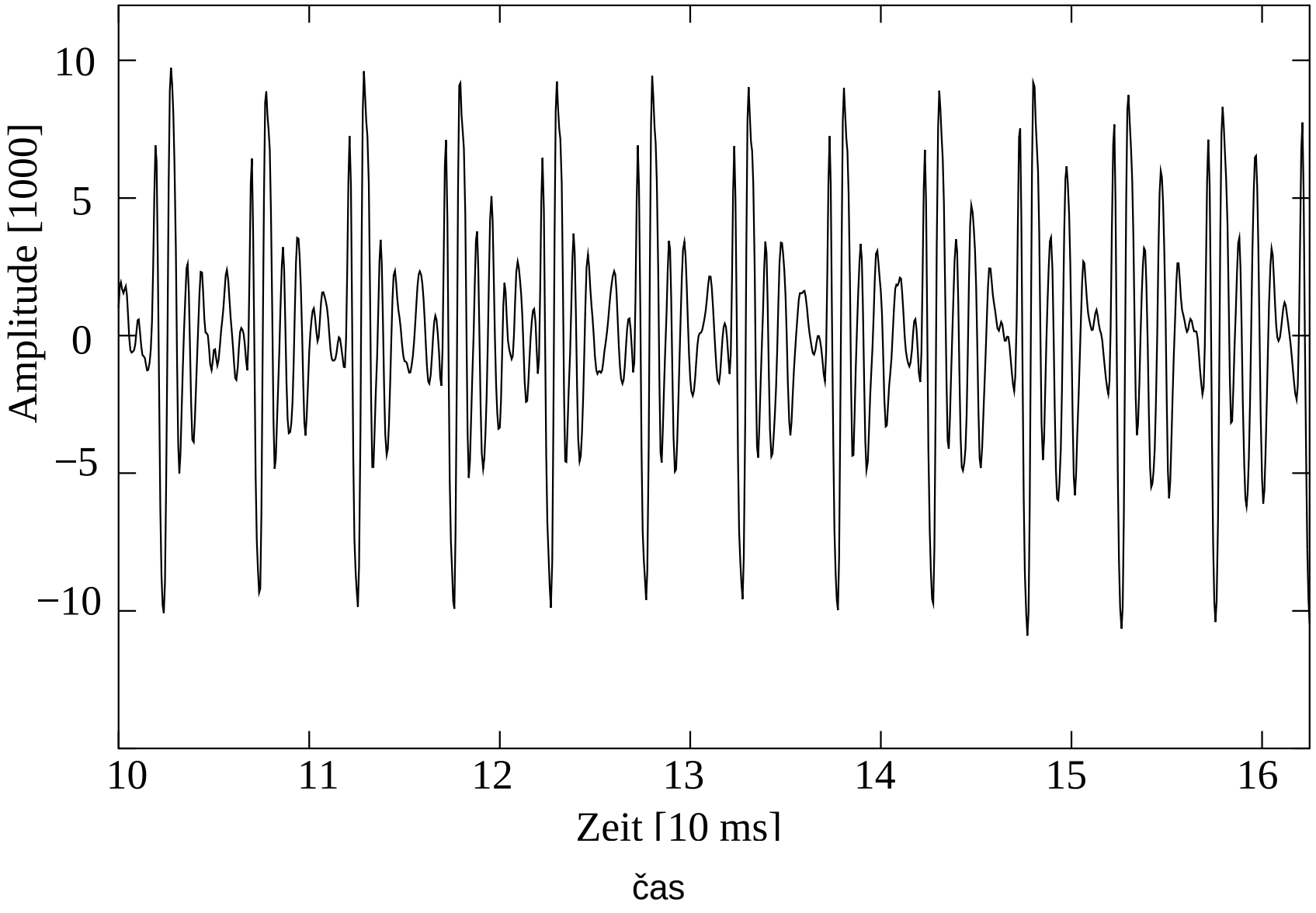




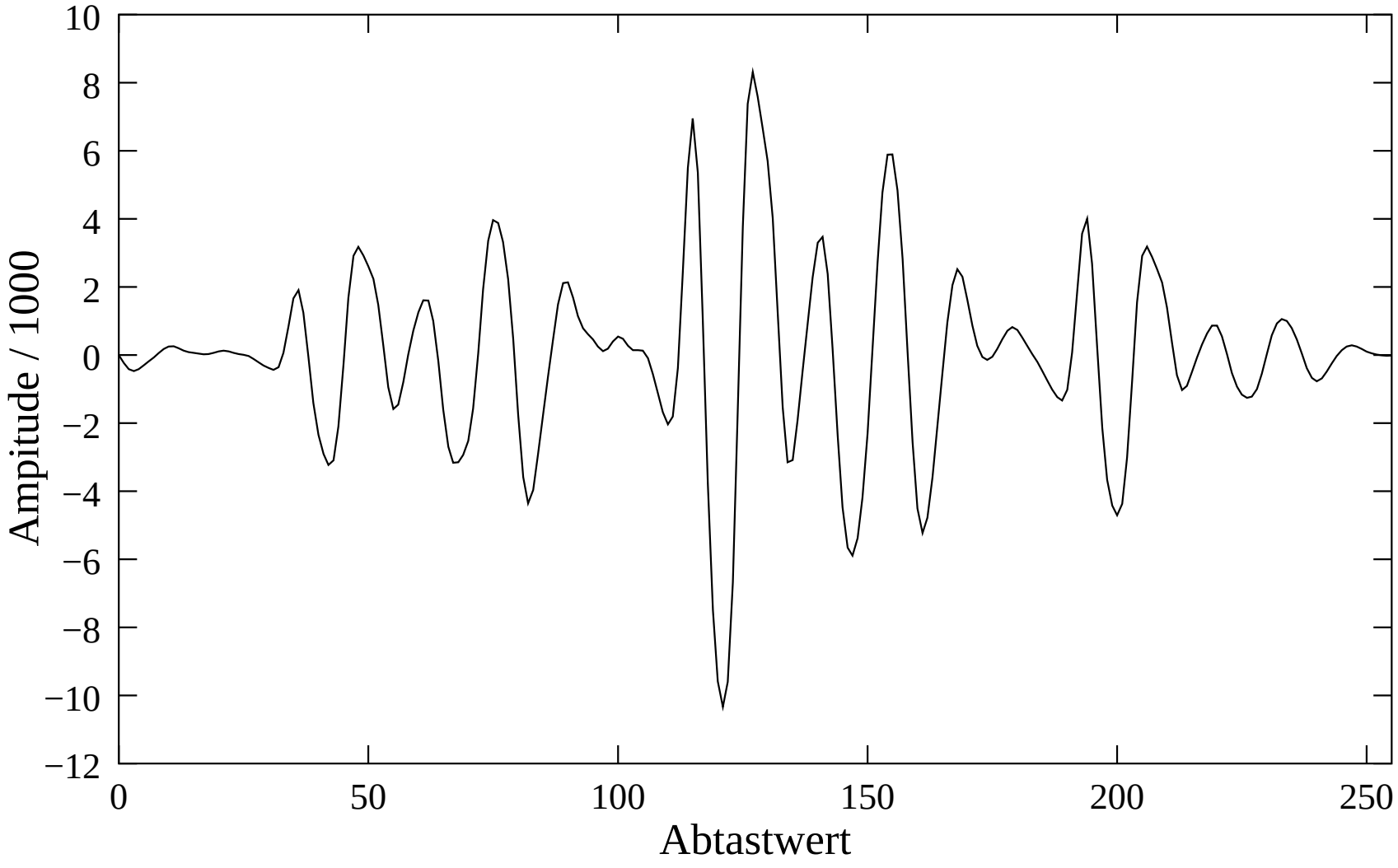


(různé rozsahy hodnot na amplitudových osách)

/a:/ z „Bahnhof“ (64 ms)

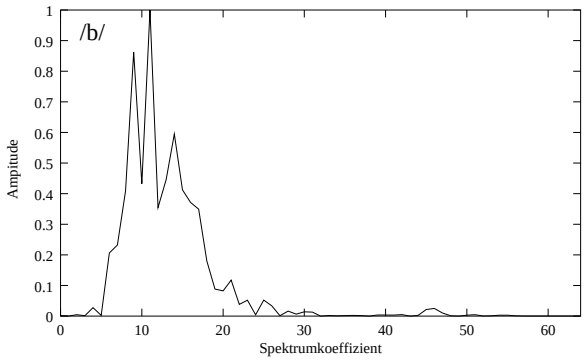


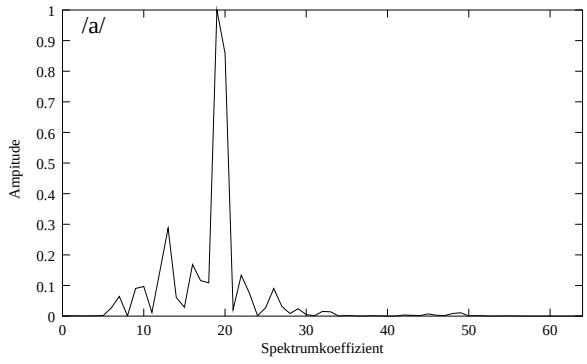
/a:/ z „Bahnhof“ (16 ms Hammingovo okno)

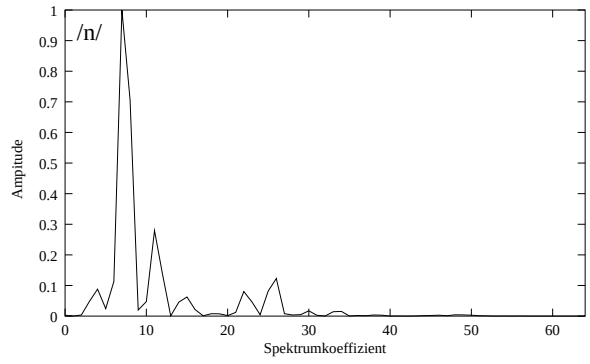


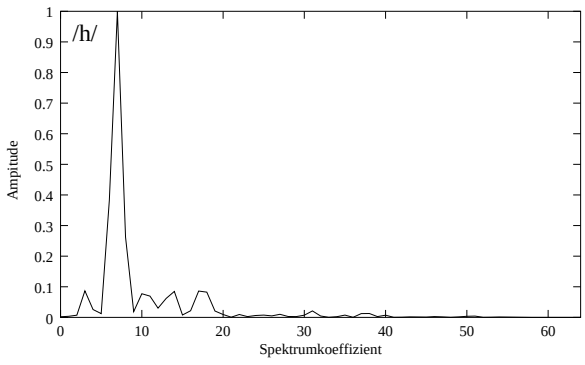
vzorkovací hodnota

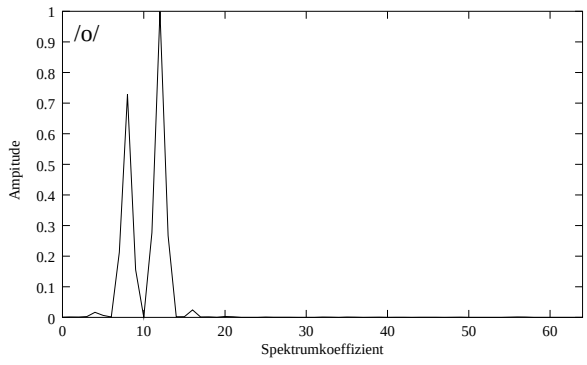
Spektra hlásek ve slově „Bahnhof“ (lineární)

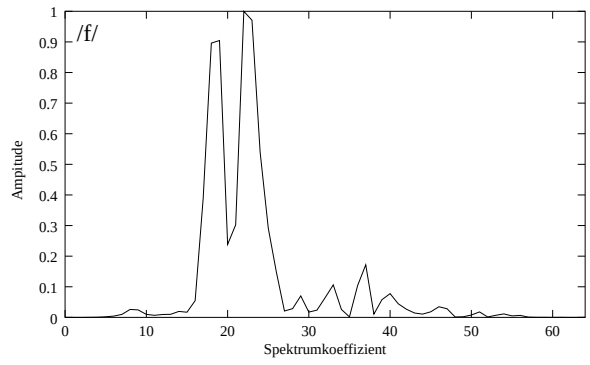




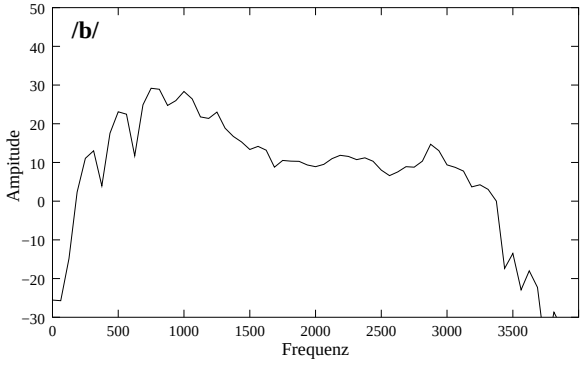


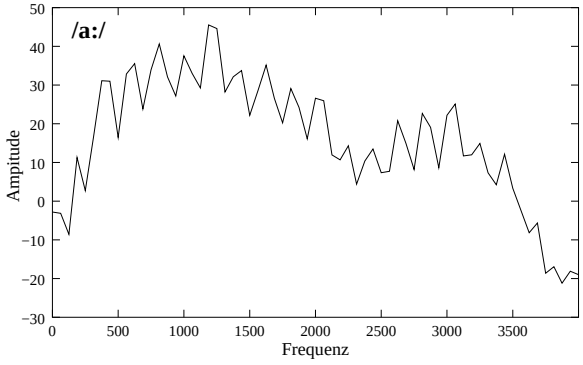


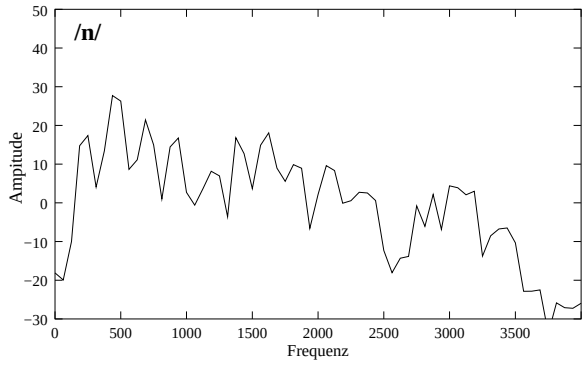


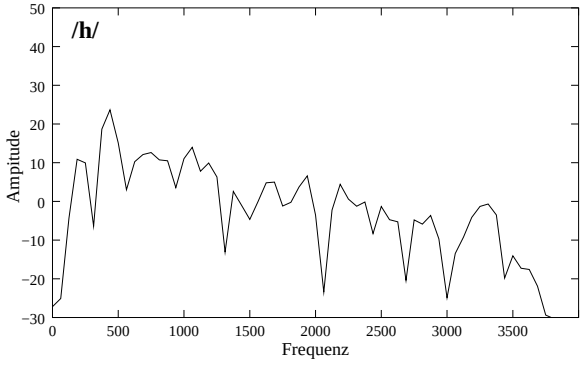


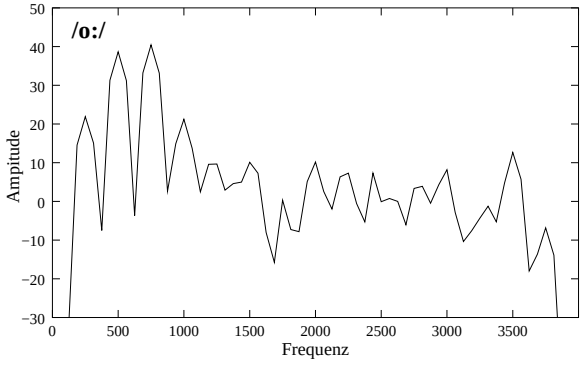
Spektra hlásek ve slově „Bahnhof“ (logaritmovaná)

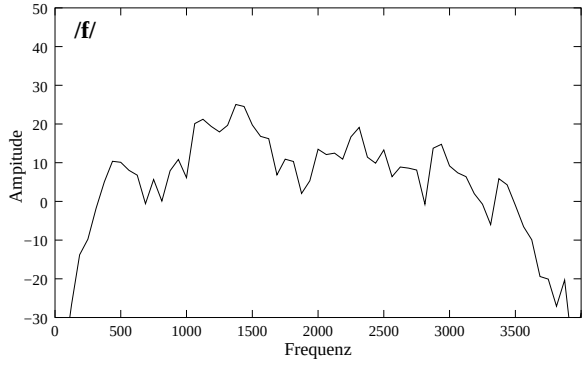




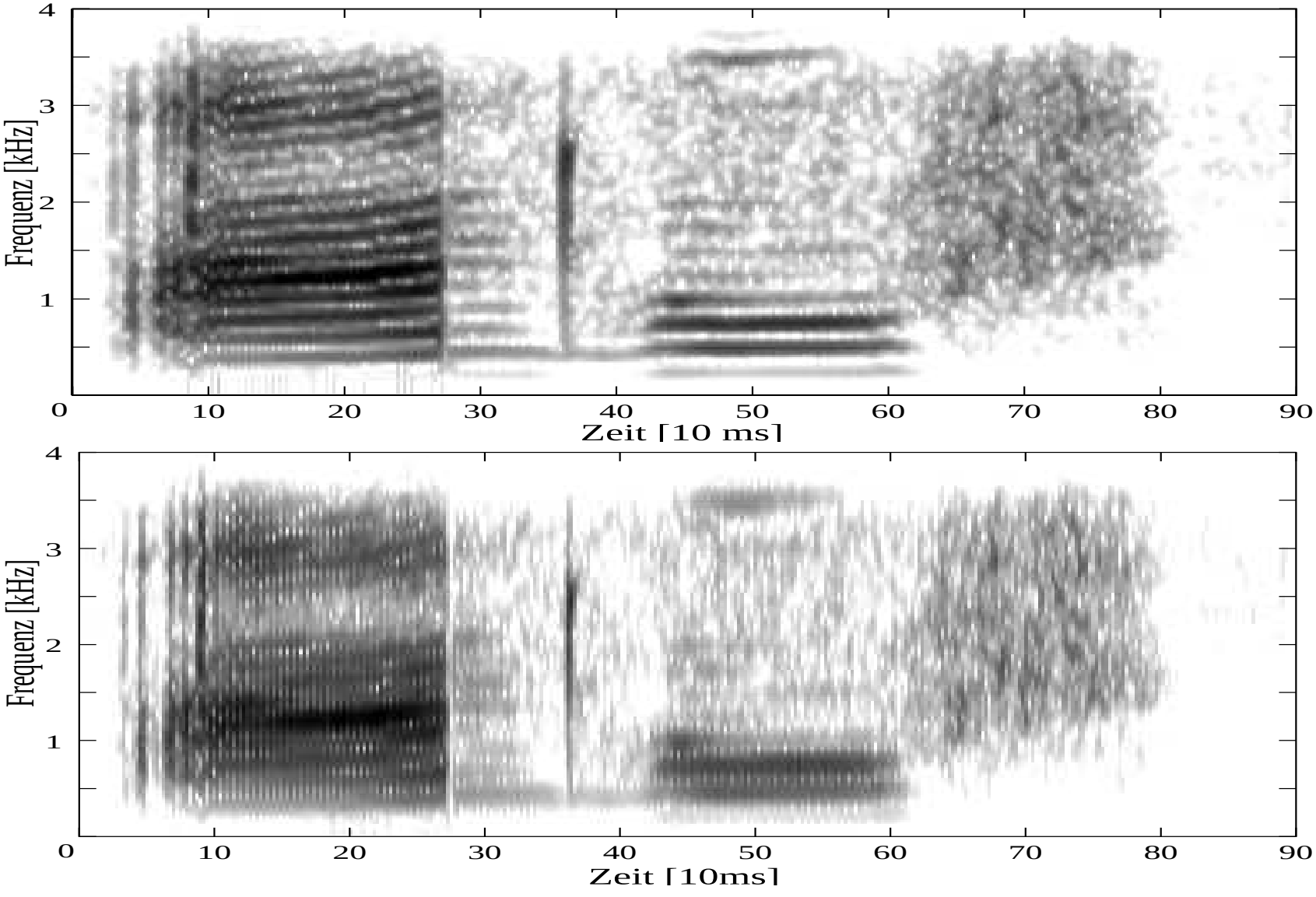




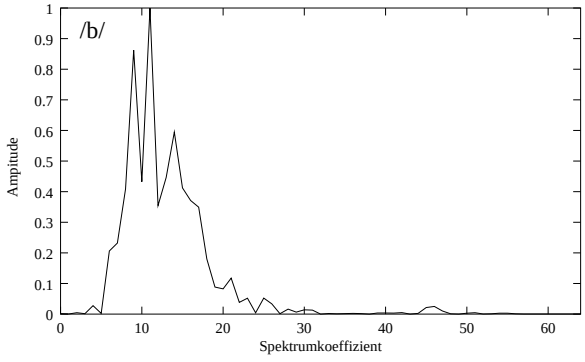


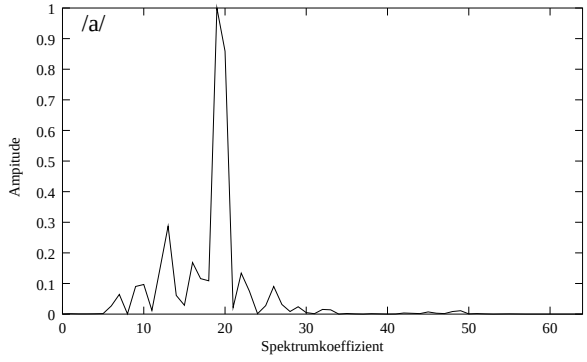


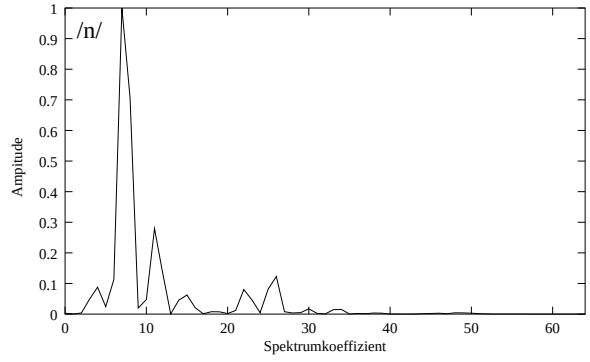
Spektrogramy slova „Bahnhof“

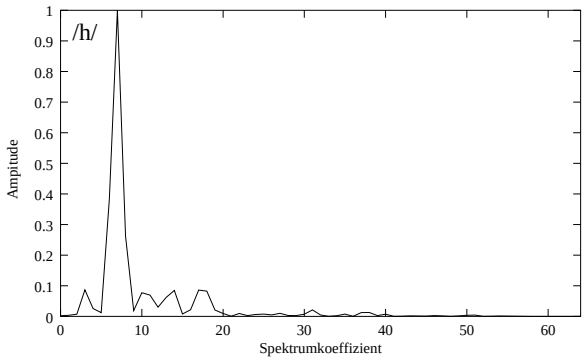


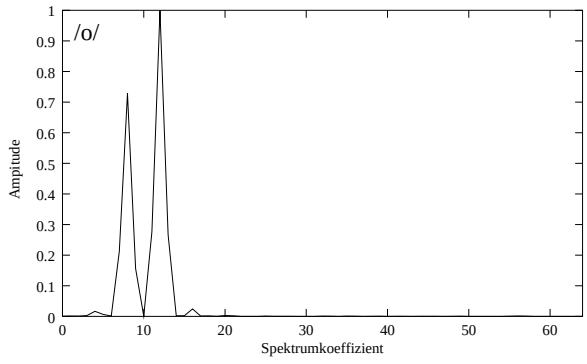
Spektra hlásek ve slově „Bahnhof“ (lineární)

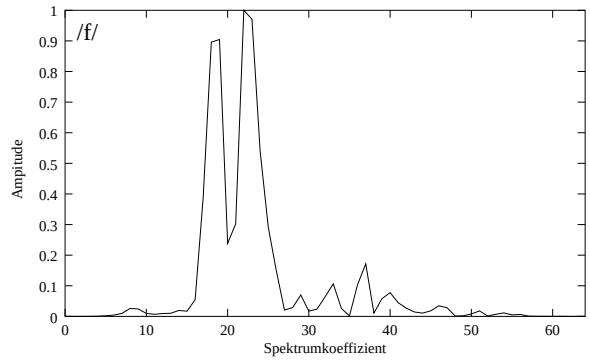




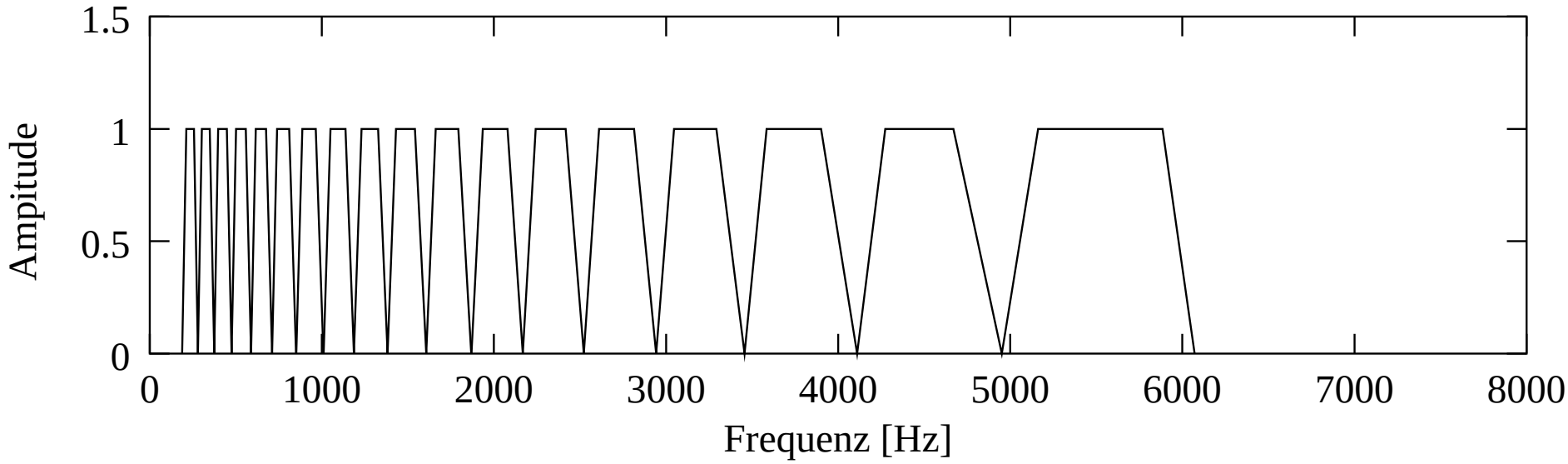




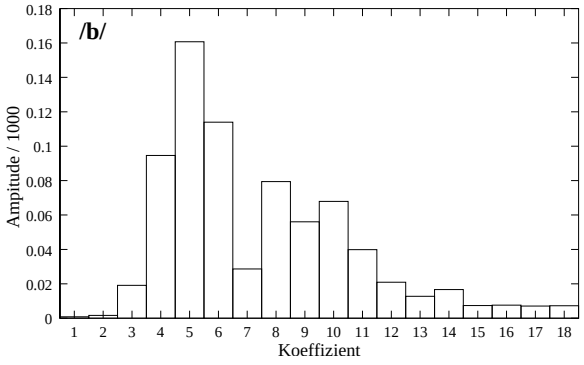


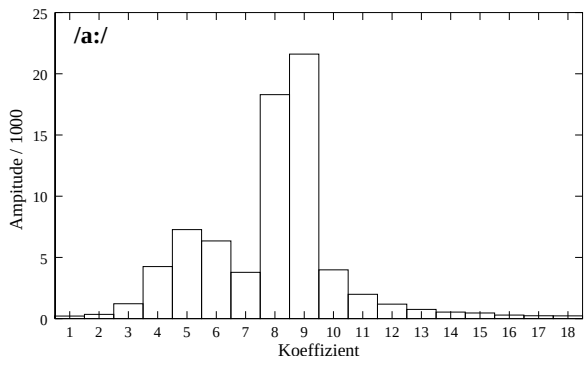


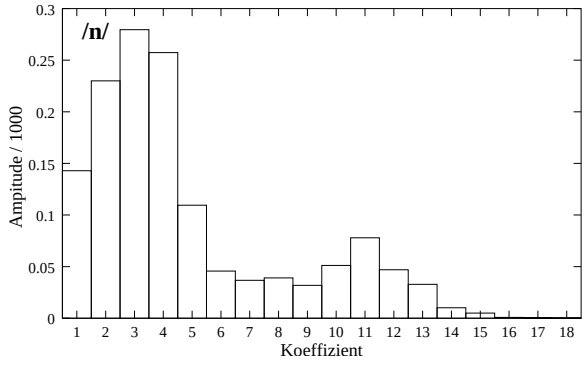
Mel-banka filtrů s 18 lichoběžníkovými filtry

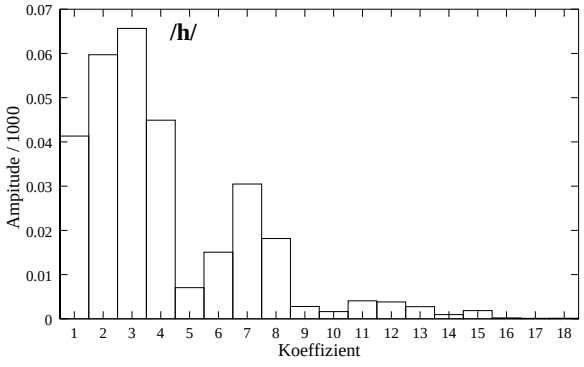


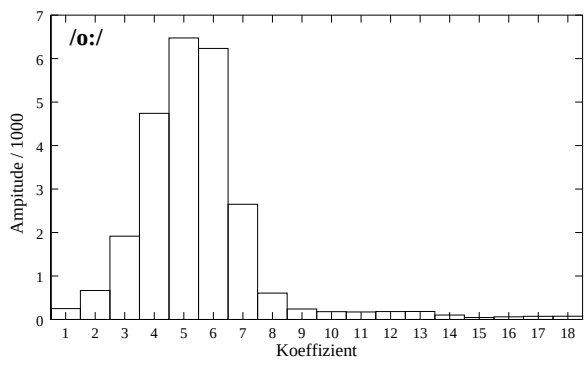
lineární Mel-spektra hlásek ve slově „Bahnhof“

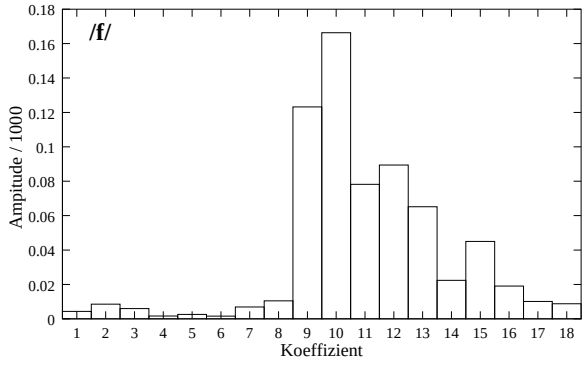






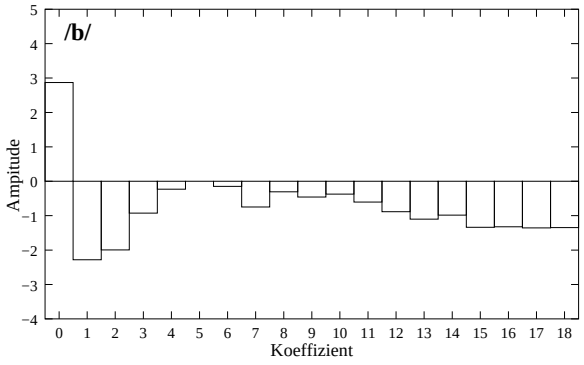


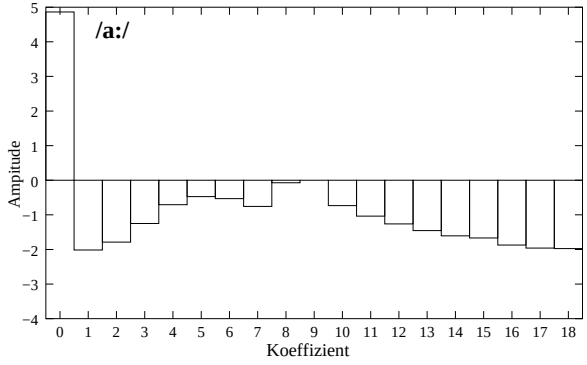


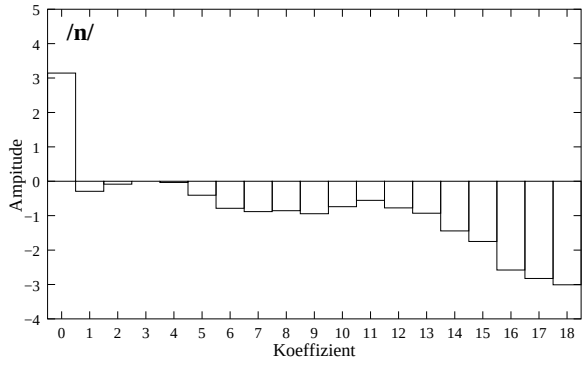


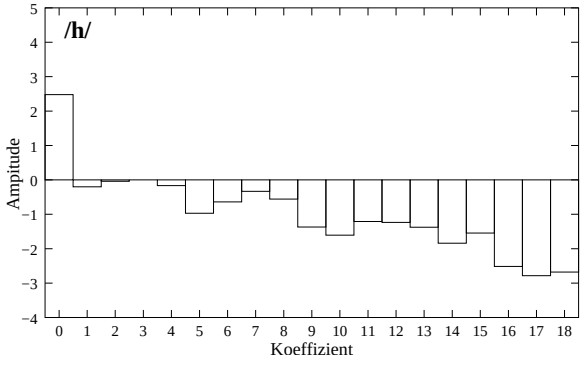
(různé rozsahy hodnot na amplitudových osách)

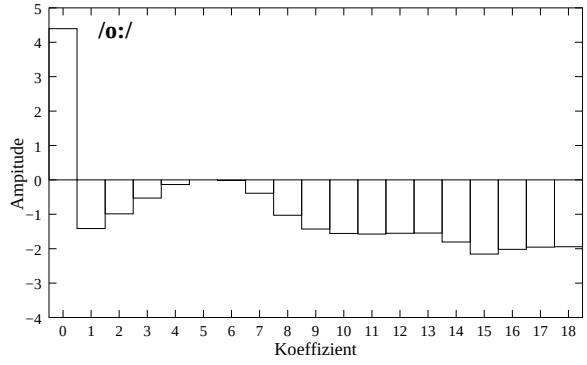
Logaritmovaná mel-spektra hlásek ve slově „Bahnhof“

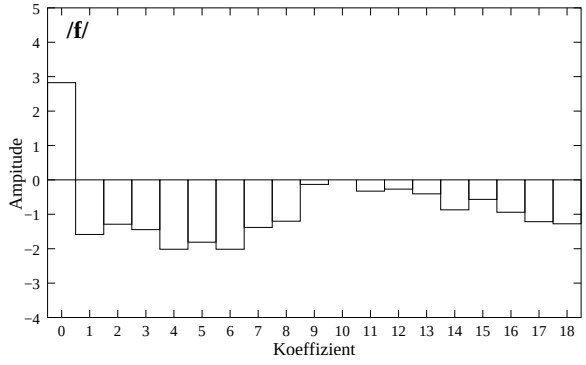






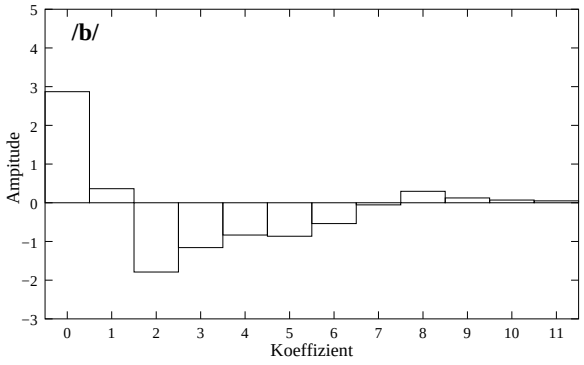


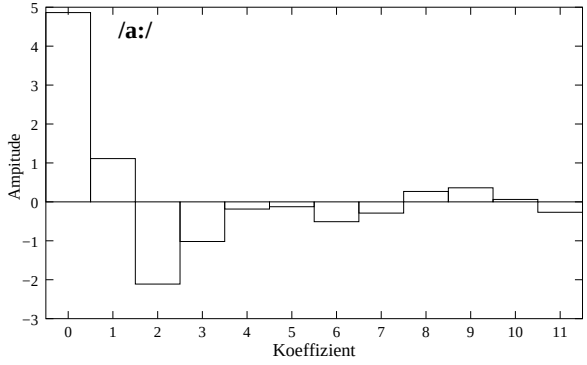


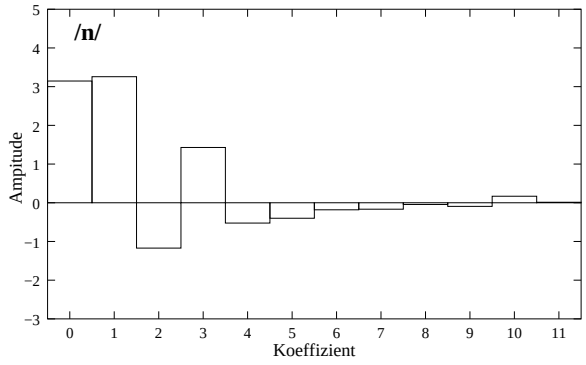


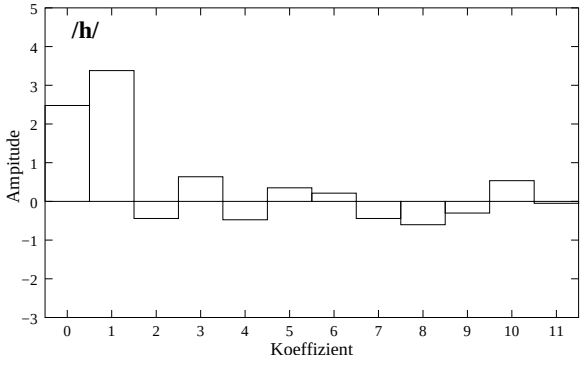
(příznak 0 je energie)

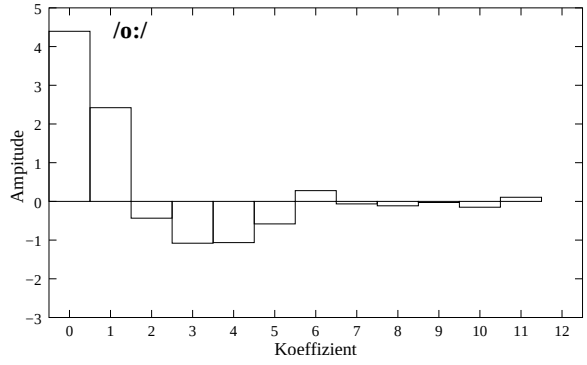
MFCC hlásek ve slově „Bahnhof“

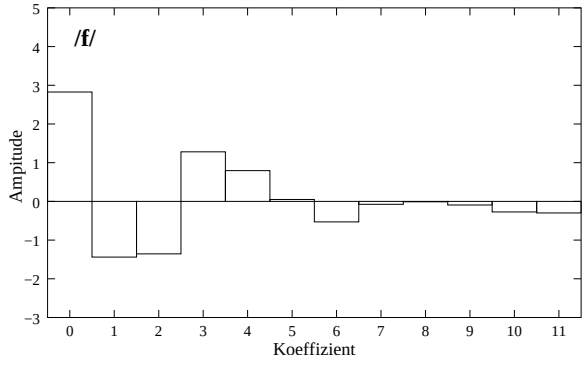






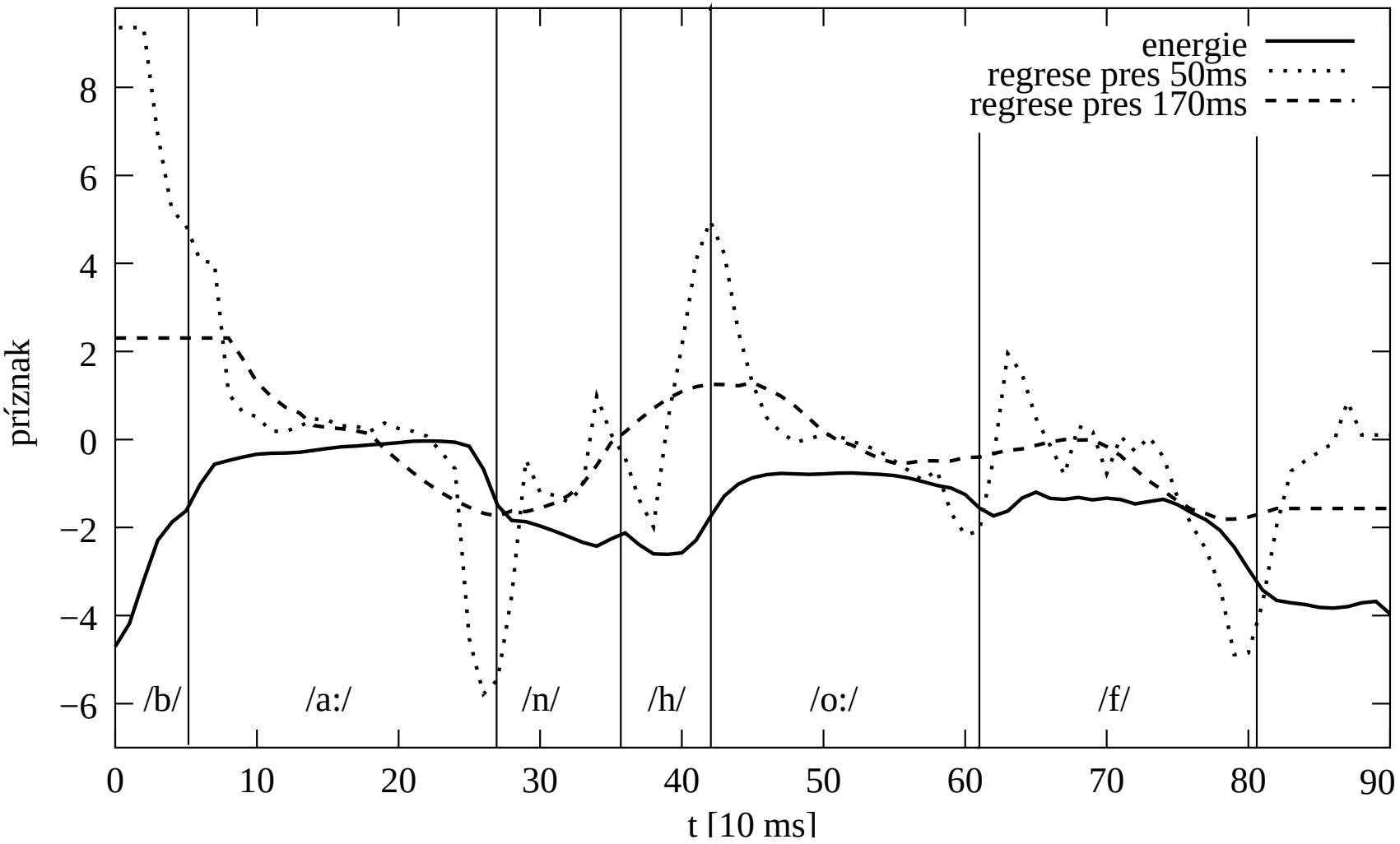




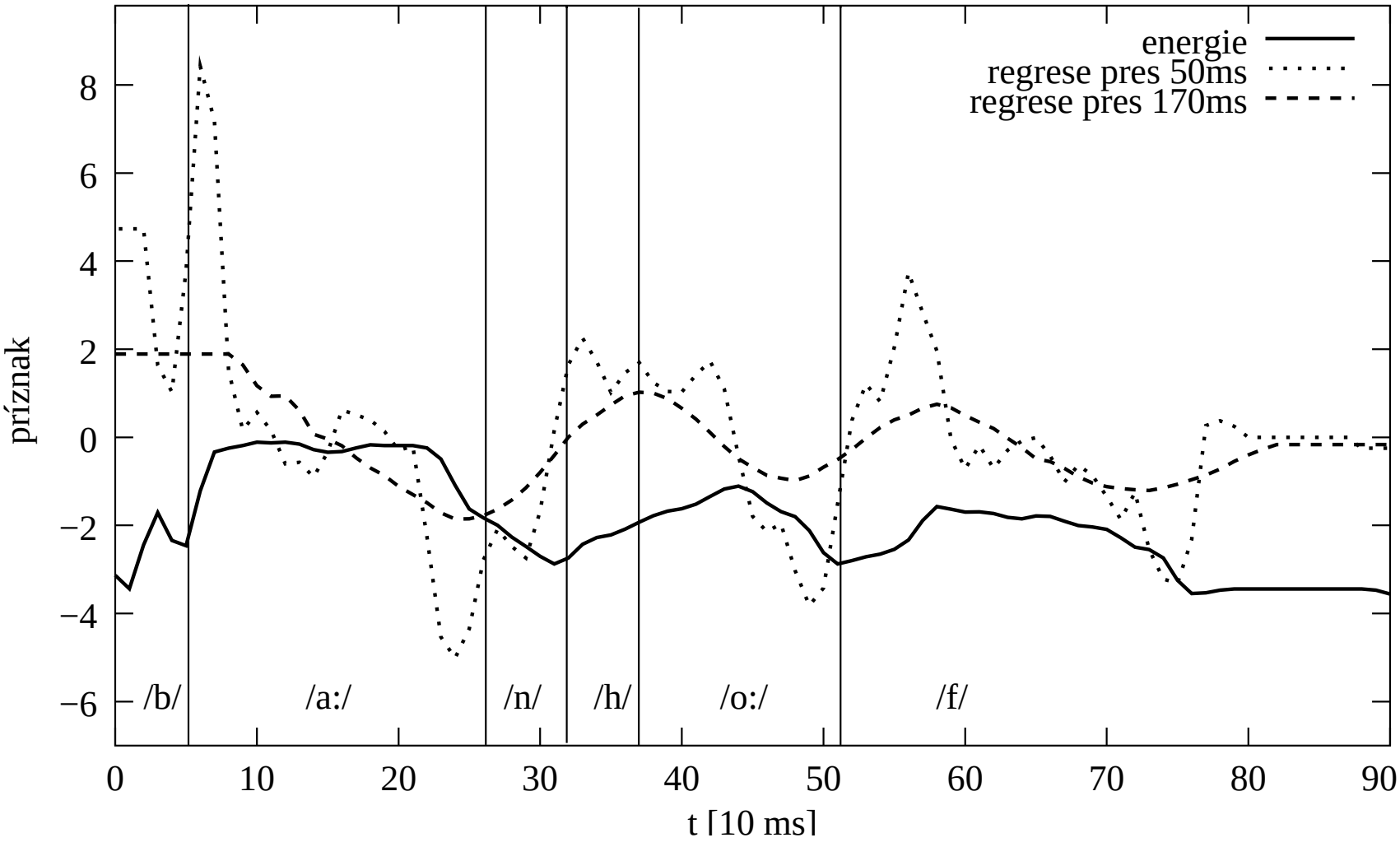


(příznak 0 je energie)

Celková energie a derivace slova „Bahnhof“ (I)

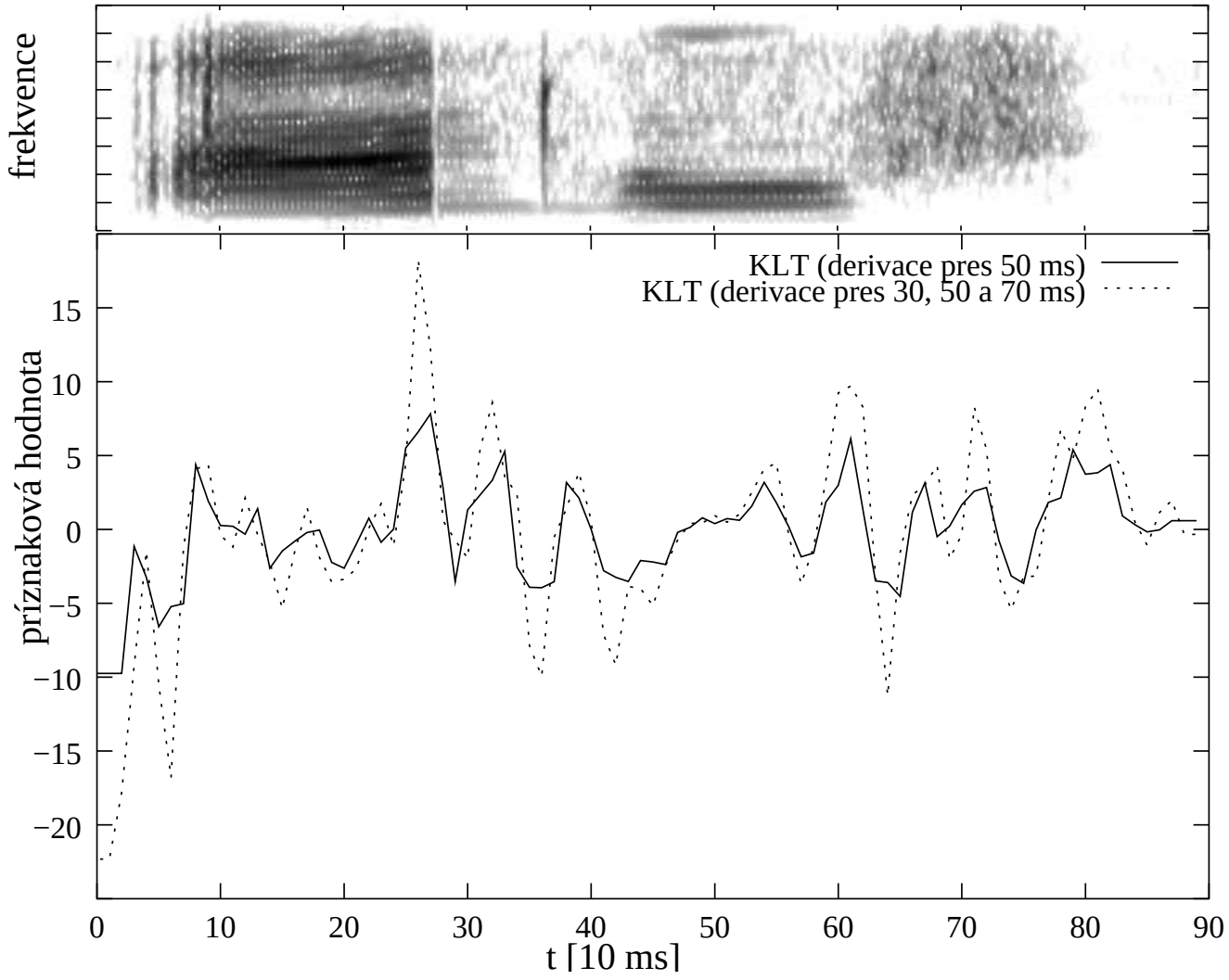


Celková energie a derivace slova „Bahnhof“ (II)

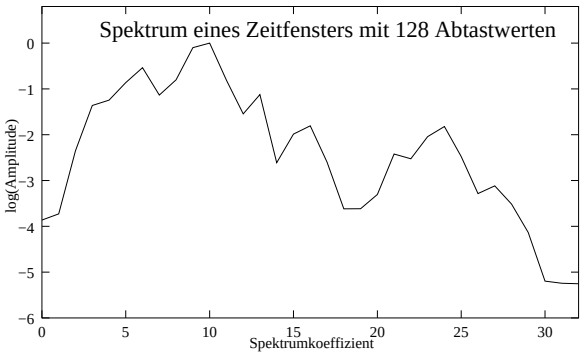


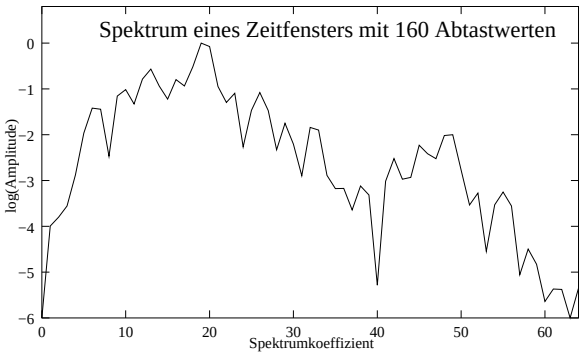
První dynamický příznak po KLT derivací

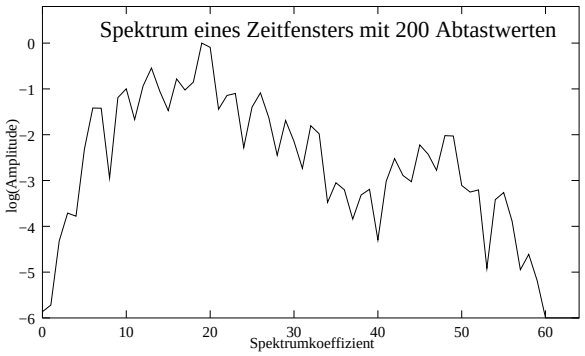
KLT: Karhunen-Loèveho transformace ≈ Principal Component analysis (PCA)

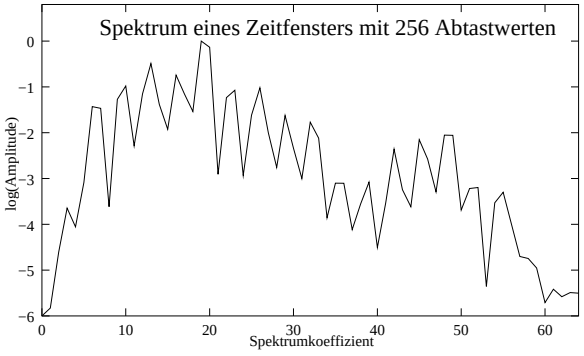


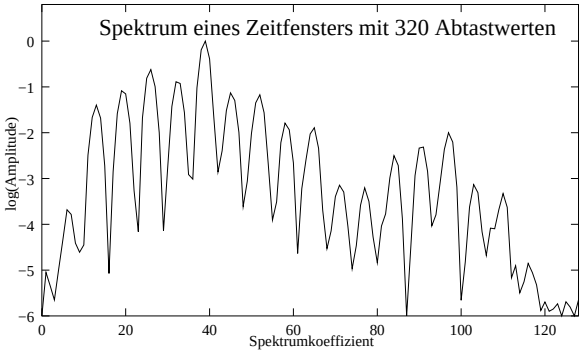
Spektra při různých časových rozlišeních (logaritmov.)

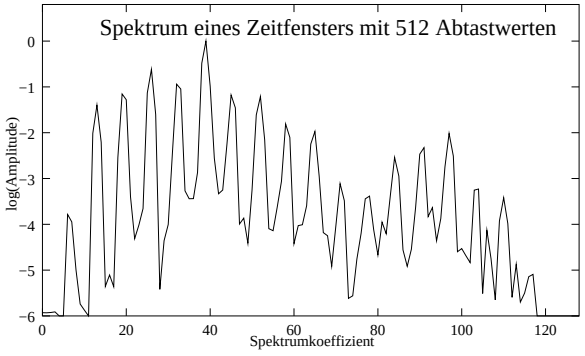












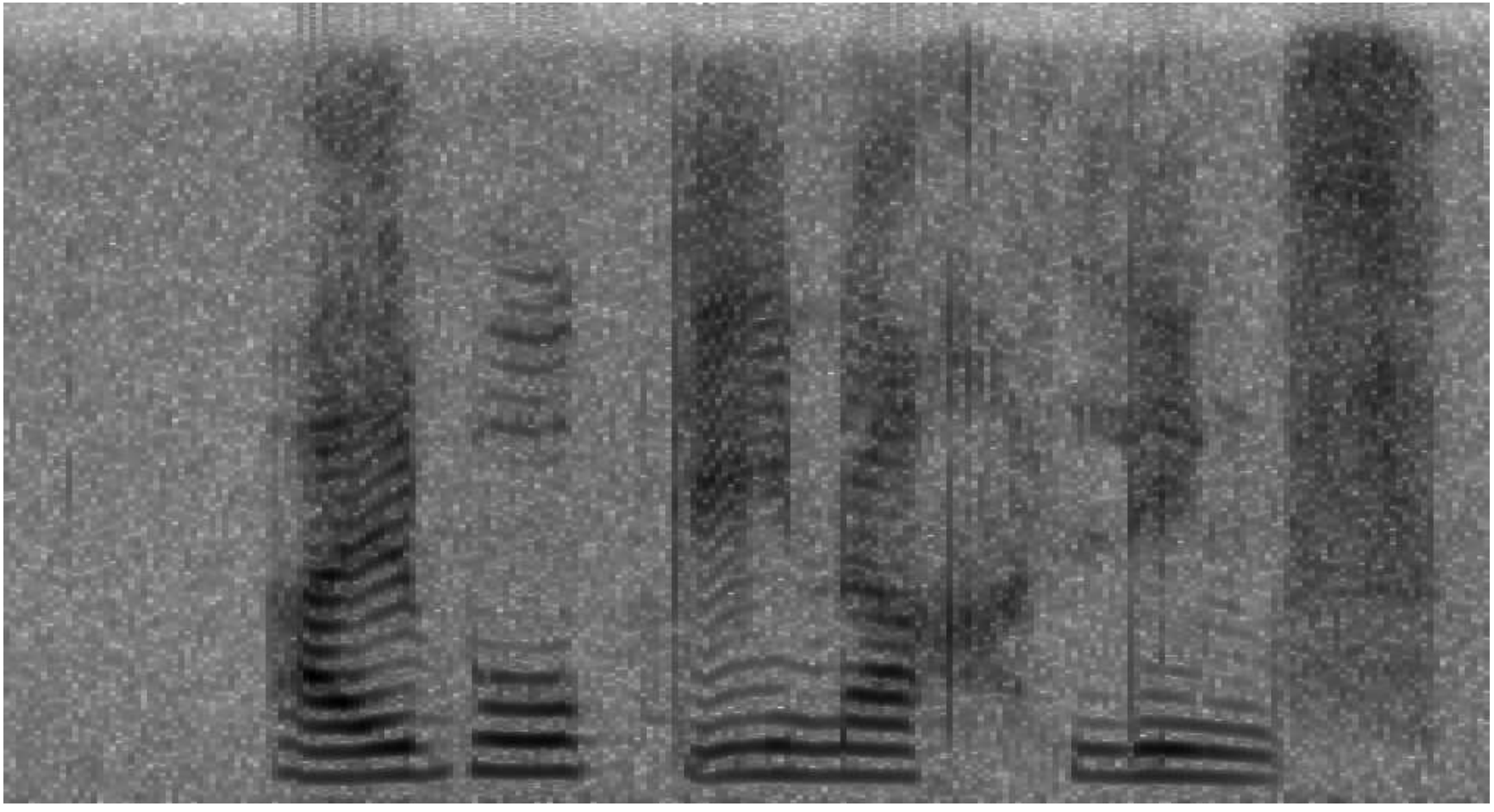
hláska /a:/ ve slově „Bahnhof“ („nádraží“);
okna s 128, 160, 200, 256, 320, 512 vzorkovacími hodnotami

Co slyšíme z MFCC?

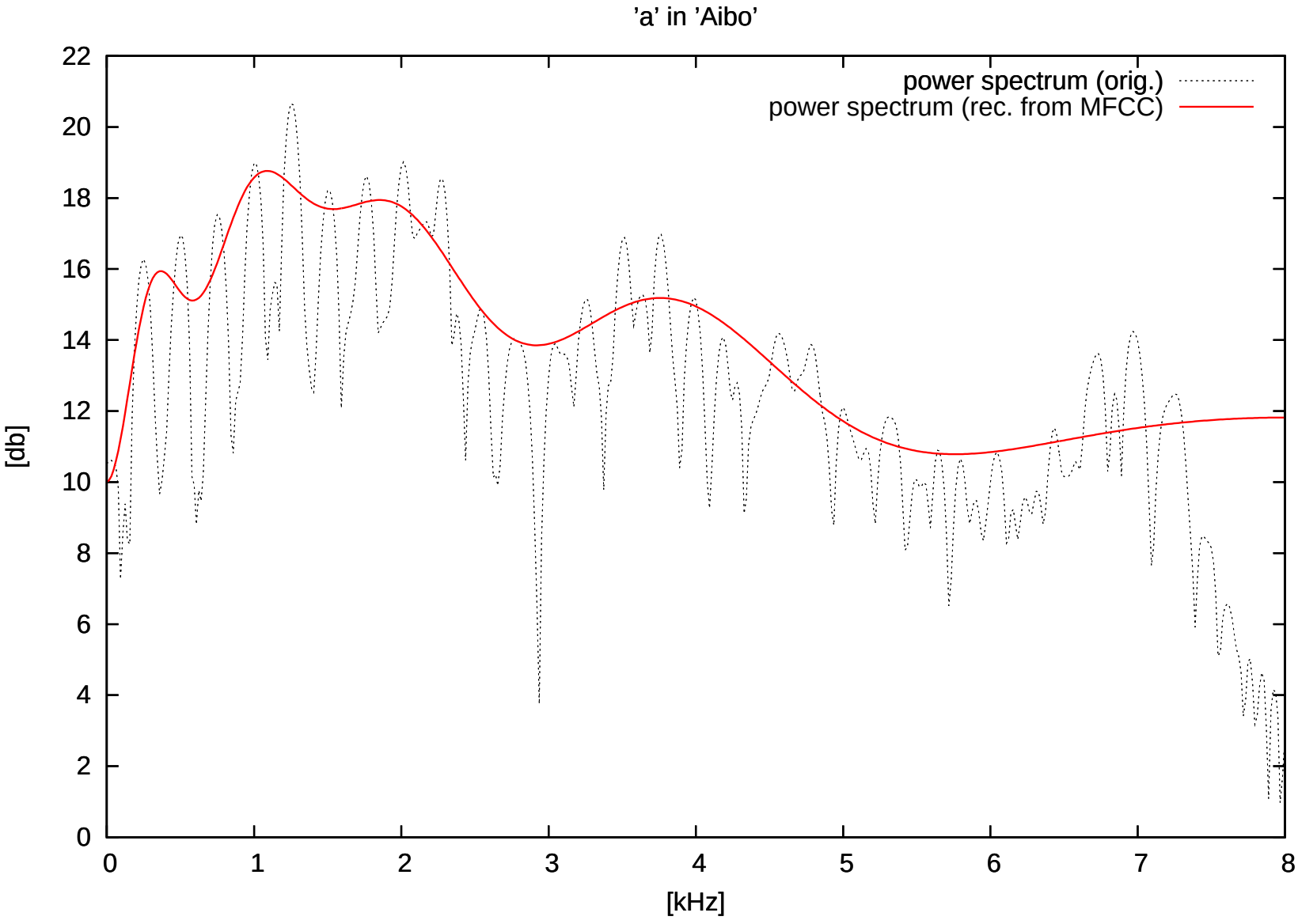
- můžeme rekonstruovat spektrum z MFCC
- ze spektra můžeme rekonstruovat signál
- příklad: *Aibo, geh' nach links! (Aibo, jdi doleva!)*
- 16 kHz, Hammingovo okno, 20 ms okno analyzování, 10 ms posuv, 1024 FFT
- mel-spektrum s 25 koef., DCT, 12 MFCC
- hlazení mel-pásmy a redukce na 12 koef.

log-spektrum

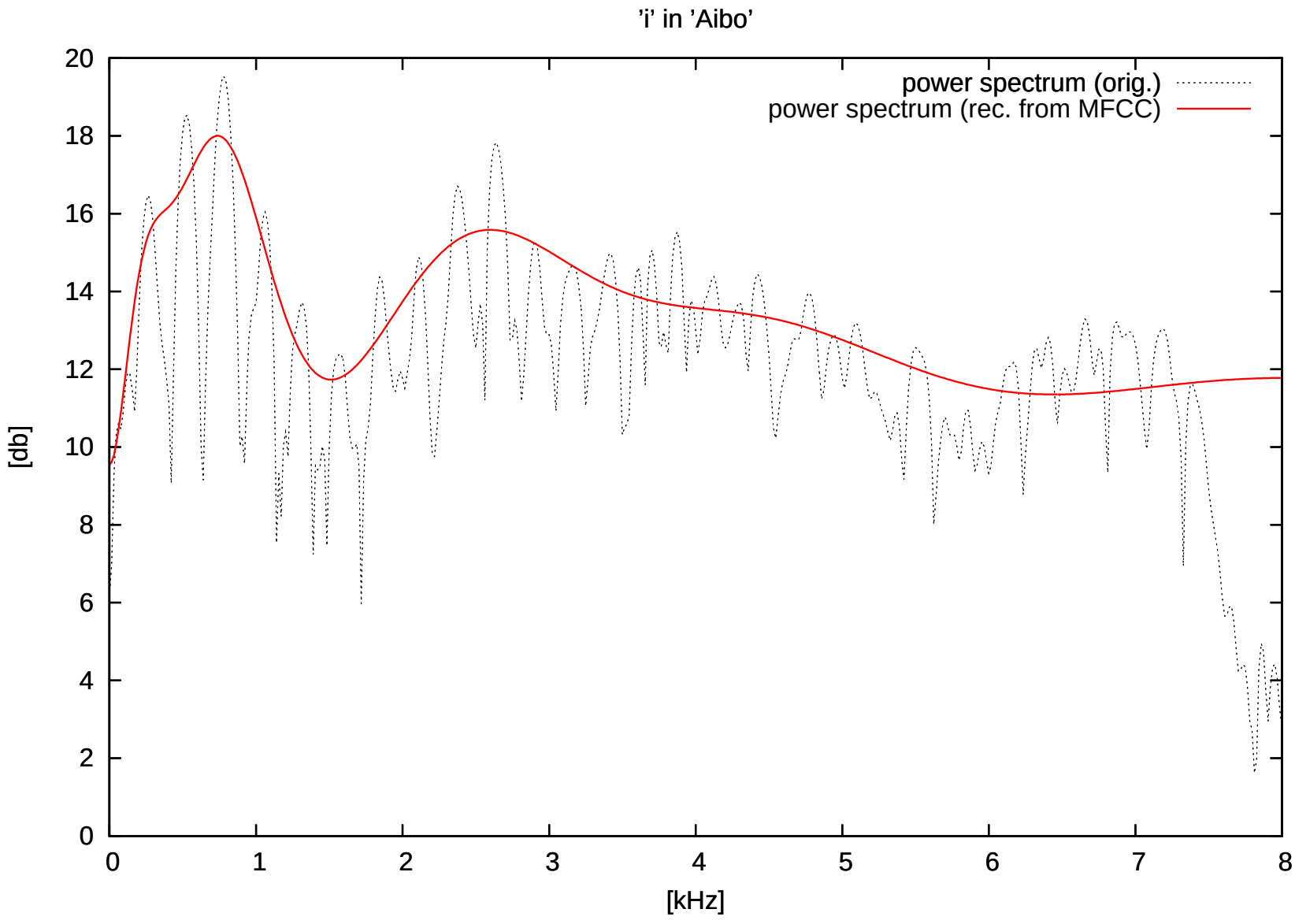
„Aibo, geh’ nach links!“ („Aibo, jdi doleva!“)



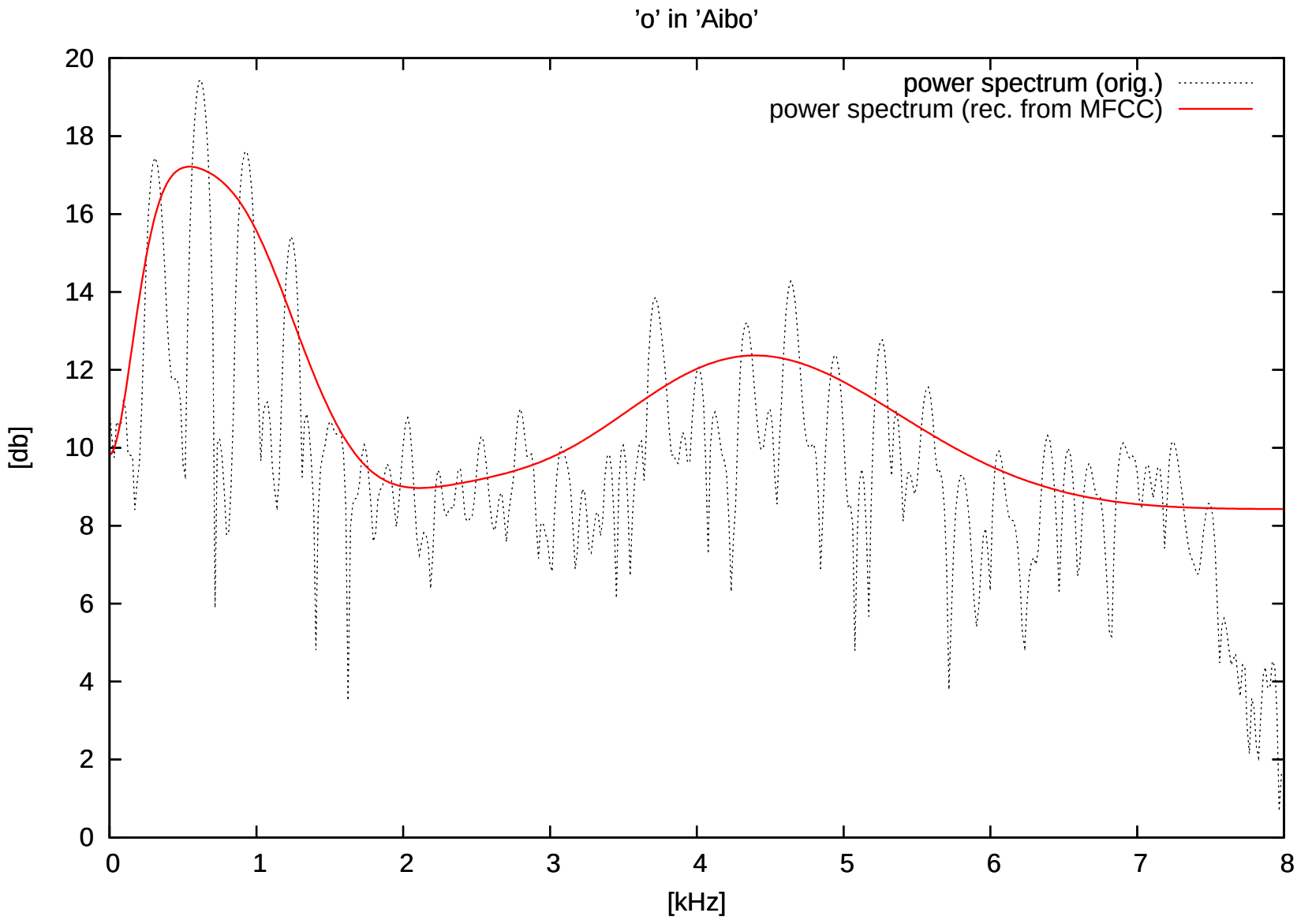
Spektrum rekonstruované z MFCC



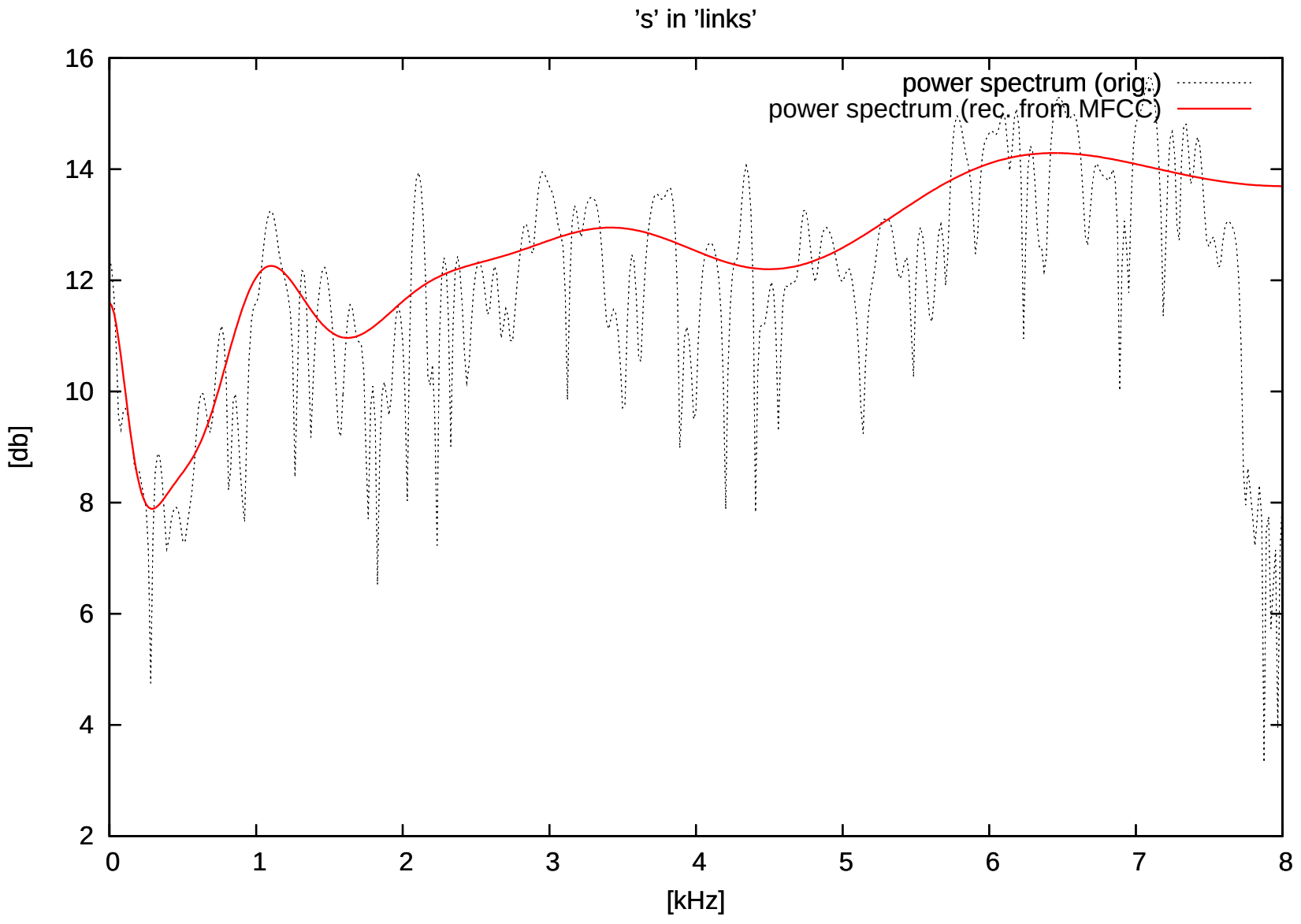
Spektrum rekonstruované z MFCC



Spektrum rekonstruované z MFCC

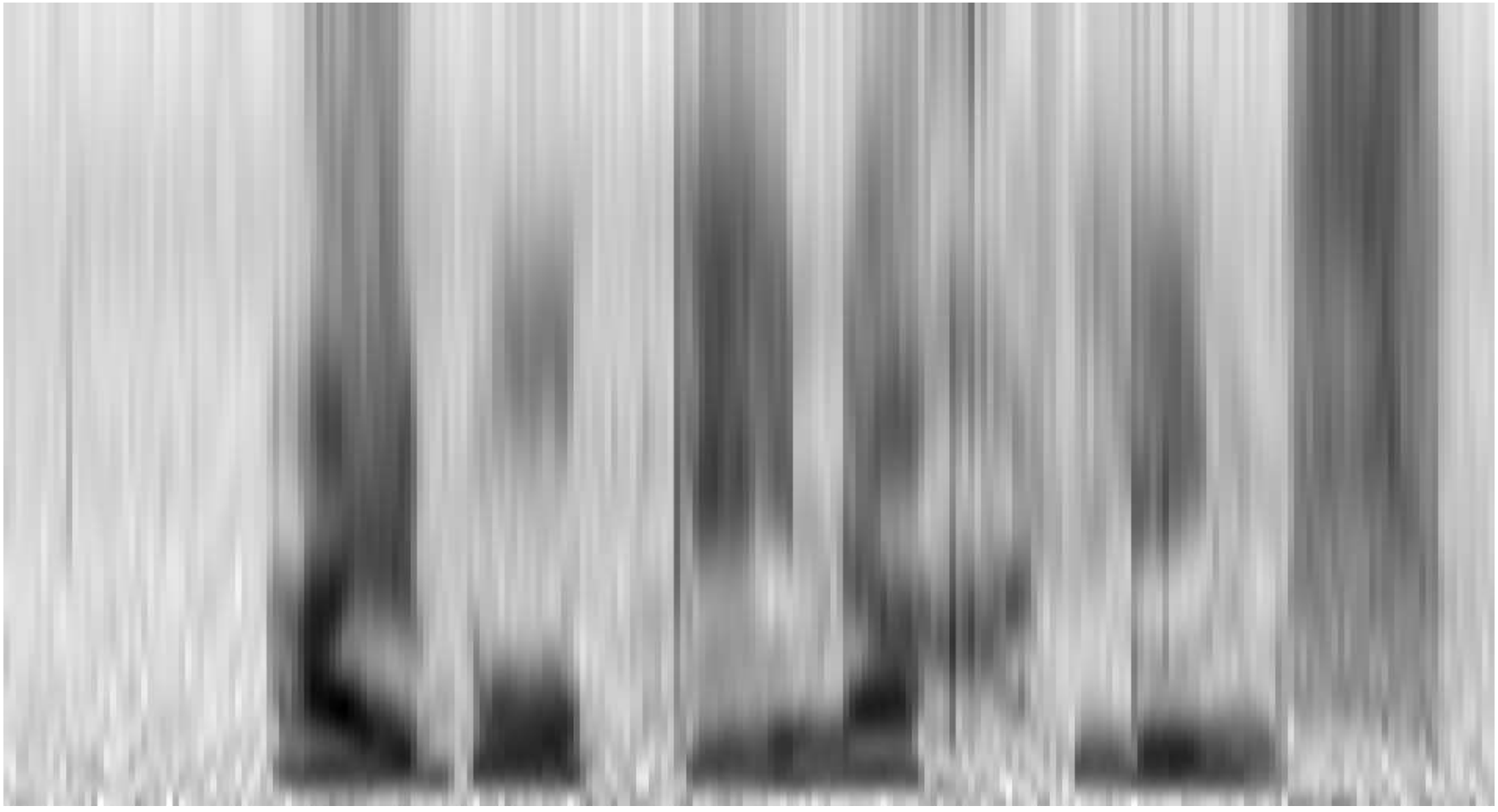


Spektrum rekonstruované z MFCC



Resyntéza z MFCC

„Aibo, geh’ nach links!“ („Aibo, jdi doleva!“)



Lineární prediktivní kódování (LPC)

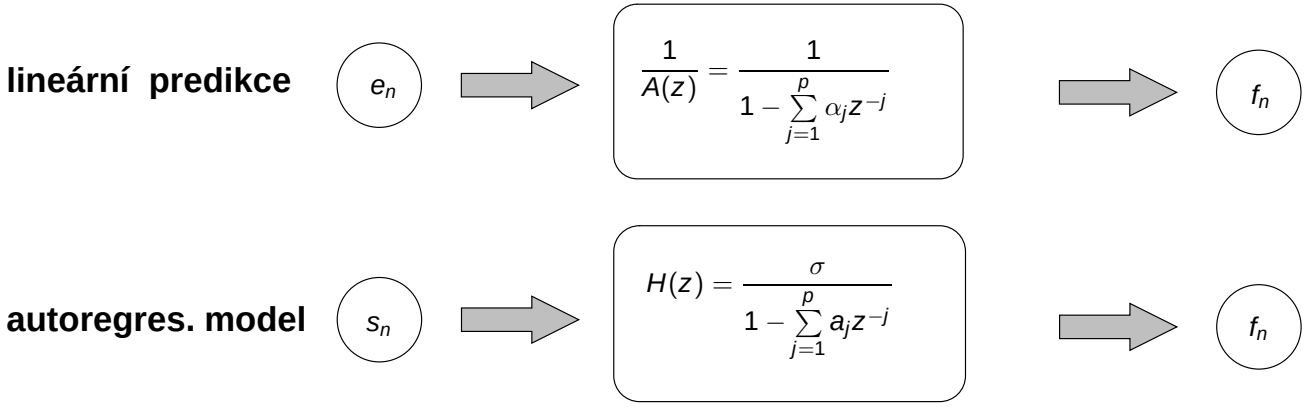
Motivace

- source-filter model řeči: prvotní signál s_n, který je filtrován systémem $H(z)$
- $H(z)$ je při autoregresivním modelu aproximován allpole-systémem: $H(z) \simeq \sigma / A(z)$ kde $A(z) = 1 - \sum_{j=1}^{p} a_j z^{-j}$
- pro z-transformaci $F(z)$ vzorkovacích hodnot f_n řečového signálu platí $F(z) = S(z) \cdot H(z) = S(z) \cdot \frac{\sigma}{A(z)}$
- sdružování s_n a σ k $e_n \rightarrow$ aproximace řečového signálu allpole-filtrem: $F(z) = E(z) \cdot \frac{1}{A(z)}$
  kde $E(z)$ je z-obraz pro $e_n = \sigma s_n$

Lineární prediktivní kódování (LPC)

$F(z) = E(z) \cdot \frac{1}{A(z)} \Rightarrow F(z) \cdot A(z) = E(z)$

- Aplikace inverzní z-transformace tvoří
  $e_n = f_n - \sum_{j=1}^{p} a_j f_{n-j}$; $A(z)$ je lineární
- Lineární predikce tvoří předpověď vzorkovací hodnoty f_n prostřednictvím lineární kombinace p předešlých vzorkovacích hodnot f_{n-j}:
  $\hat{f}_n = \sum_{j=1}^{p} \alpha_j f_{n-j}$
- e_n se nazývá chyba předpovědí, protože platí $e_n = f_n - \hat{f}_n$.
- prediktivní koeficienty $\alpha_j =$ modelové parametry a_j:
  $e_n = \sigma s_n$ (chyba = zesilování $\times$ buzení)



Určení prediktivních koeficientů α_j

- minimalizuj akumulovanou kvadratickou chybu
  $\varepsilon = \sum_{n=n_0}^{n_1} e_n^2 = \sum_{n=n_0}^{n_1} \left(\sum_{j=0}^{p} \alpha_j f_{n-j} \right)^2$
- jednodušší notace s $p \times p$-maticí Φ:
  $\varepsilon = \sum_{j=0}^{p} \sum_{k=0}^{p} \alpha_j \phi_{jk} \alpha_k$ kde $\phi_{jk} = \sum_{n=n_0}^{n_1} f_{n-j} f_{n-k}$
- parciální derivace se stanoví na nulu:

$$\partial \varepsilon / \partial \alpha_k = 2 \sum_{j=0}^{p} \alpha_j \phi_{jk}$$

$\Longrightarrow$ lineární soustava rovnic s proměnnými $\alpha_1, \ldots, \alpha_p$
$\sum_{j=1}^{p} \alpha_j \phi_{jk} = -\phi_{0k}$, $k = 1, \ldots, p$

- tyto rovnice se jmenují Yule-Walkerovy rovnice

Určení prediktivních koeficientů α_j

dvě důležité metody řešení Yule-Walkerových rovnic:

- kovarianční metoda
- autokorelační metoda

tady jen autokorelační metoda s Durbinovou rekurzí:

- koeficienty $r_k^{(m)}$ krátkodobé autokorelační funkce
  $r_k^{(m)} = \sum_{n=-\infty}^{\infty} f_n^{(m)} f_{n+k}^{(m)} = \sum_{n=-\infty}^{\infty} f_n w_{m-n} f_{n+k} w_{m-n-k}$
- s pravoúhlým oknem (velikost N) se tyto počítají takto:
  $r_k^{(m)} = \sum_{n=m}^{m+N-k-1} f_n f_{n+k}$
- autokorelační metoda předpokládá, že $f_n = 0 \quad \forall n < n_0, n > n_1$
  $\Rightarrow \phi_{jk}^{(m)} = r_{|j-k|}^{(m)}$
  ϕ je Toeplizova matice (symetricky, všechny hodnoty na diagonále jsou stejné)

Durbinova rekurze

účinný algoritmus pro řešení Yule-Walkerových rovnic, když ϕ je Toeplizova matice

1. inicializuj: $\varepsilon^0 = r_0$, $\alpha_1^0 = 0$
2. proveď iteraci přes $n = 1, \ldots, p$:

$$\begin{aligned} k_n &= \frac{r_n - \sum_{j=1}^{n-1} \alpha_j^{n-1} r_{|n-j|}}{\varepsilon^{n-1}} \\ \alpha_n^n &= k_n \\ \alpha_i^n &= \alpha_i^{n-1} - k_n \cdot \alpha_{n-i}^{n-1} \quad \forall 1 \leq i < n \\ \varepsilon^n &= \varepsilon^{n-1} \cdot (1 - k_n^2) \end{aligned}$$

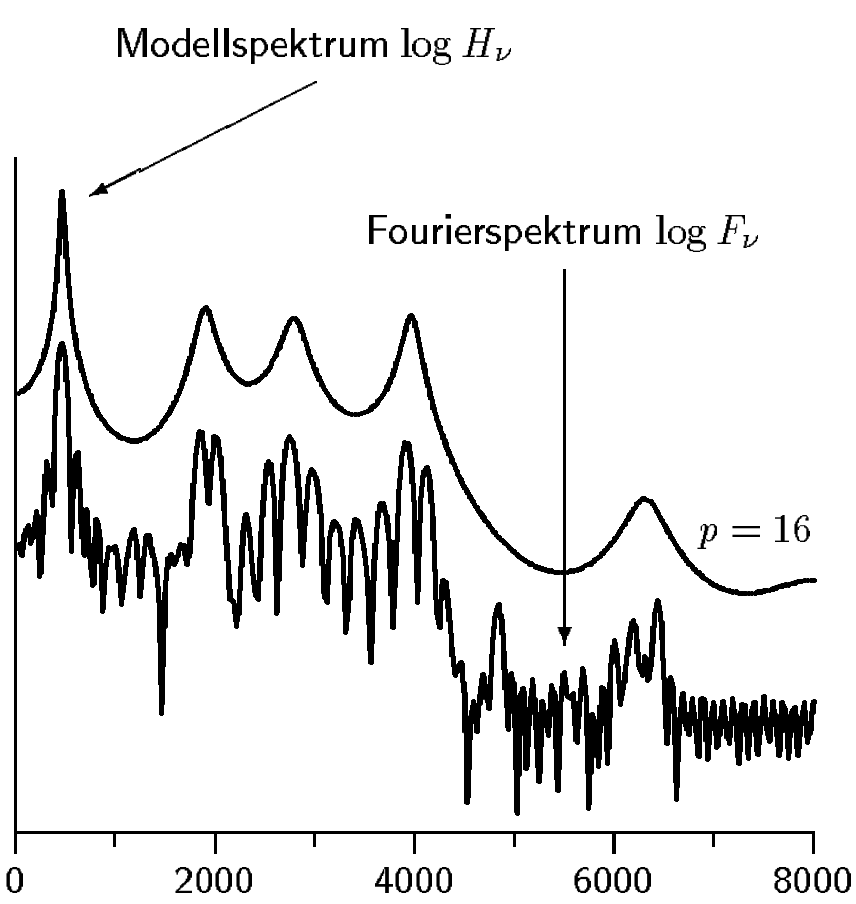
3. řešení: $\alpha_i = \alpha_i^p \quad \forall 1 \leq i \leq p$

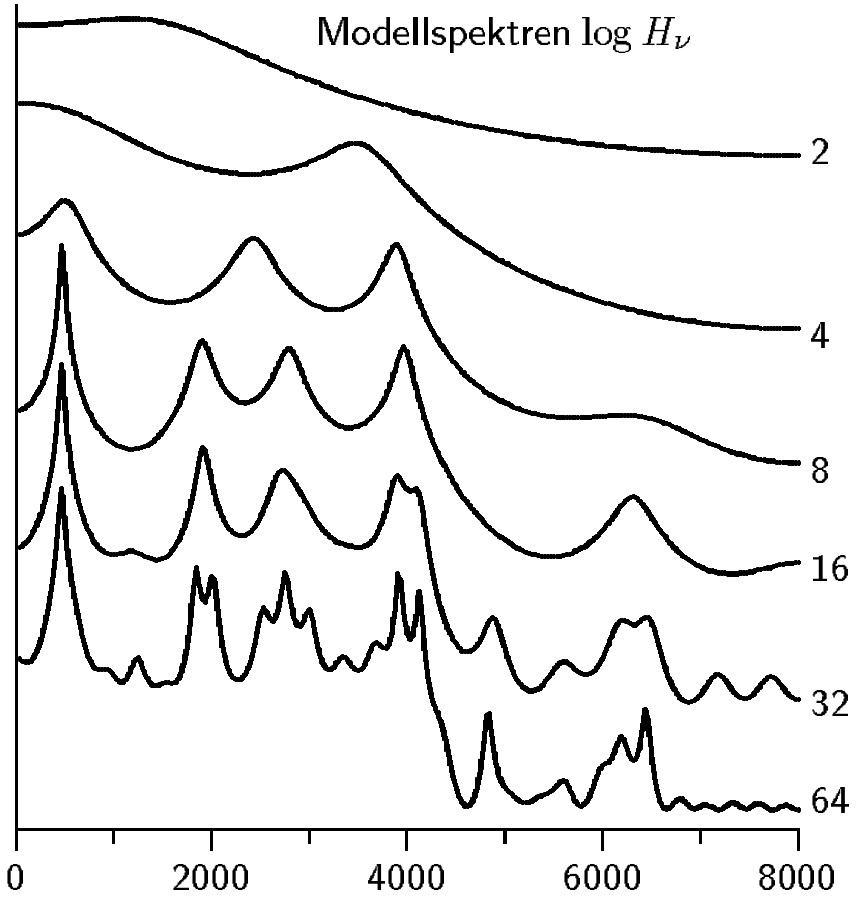
- k_n se nazývají reflexní koeficienty

Modelové spektrum

- Koeficienty lineární predikce slouží k počítání frekvenčního přenosu artikulačních orgánů:
  předpoklad (jako nahoře) $H(z) = \sigma/A(z)$
- Predikční koeficienty $\alpha_1, \ldots, \alpha_p$ jsou považovány za identické s autoregresními parametry $a_1, \ldots, a_p$.
- Modelové spektrum hrtanové záklopky (glottis), hlasového traktu a rtů:
  $H_\nu = H(e^{2\pi i\nu/N}) = \frac{\sigma}{A(e^{2\pi i\nu/N})} = \frac{\sigma}{A_\nu}$
- Hodnoty A_ν od $A(z)$ se mohou určit pomocí DFT z
  $(1, \alpha_1, \ldots, \alpha_p, \underbrace{0, \ldots, 0}_{N-p-1 \text{ nul}})$.
- Zesilovací faktor σ se počítá z
  $\sigma^2 = \sum_{j=0}^{p} \alpha_j r_j$.

DFT-spektrum a modelová spektra různých řádů





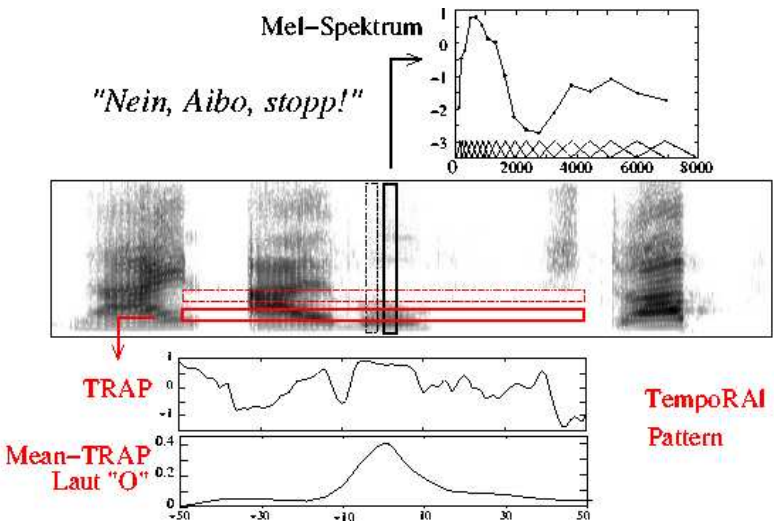
Modelové spektrum

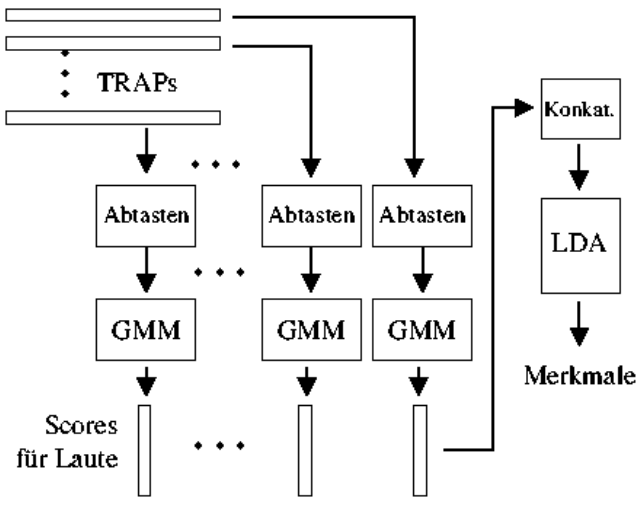
- Modelové spektrum je vyhlazená aproximace DFT-spektra.
- Se stoupajícím řádem p se blíží k DFT-spektru.
- řád příliš nízký $\Rightarrow$ splynutí formantů
- řád příliš vysoký $\Rightarrow$ harmonická struktura je modelovaná
- empirická rovnice výběru p: $p = f_A$ [kHz] $+ 4$

Percepční lineární predikce (PLP)

- kombinace koeficientů LPC s deformací frekvenční osy jako při mel-cepstru
- Podle Wienera a Chinčina autokorelační funkce r_n může být počítána pomocí DFT:
  $FT\{r_n\} = |FT\{f_n\}|^2 \Rightarrow r_n = FT^{-1}\{|FT\{f_n\}|^2\}$
- podobnost s reálním cepstrem!
- PLP: Počítej autokorelaci pro lineární predikci pomocí DFT, ale proveď mezitím Mel-frekvenční deformaci:
  $DFT \rightarrow |\cdot|^2 \rightarrow \text{mel} \rightarrow |\cdot|^{\frac{1}{3}} \rightarrow DFT^{-1} \rightarrow$ Durbinova rekurze
- Logaritmus je přitom jako při root-mel-cepstru nahrazen třetím kořenem.

Klasifikace hlásek pomocí TRAPS (H. Heřmanský)





Laut = hláska, Abtasten = vzorkování, Merkmale = příznaky

	úroveň rozpoznávání v %			komplem. inform.	
	MFCC+Δ	TRAP	všechny	MFCC+Δ	TRAP
AIBO	57,1	57,8	65,9	21,7	23,1

- 12 MFCC, 12 Δ, 12 TRAP příznaků
- TRAPS (“temporal patterns”): 1 sek. kontextu, scores pro 24 třídy hlásek
- Hodně komplementárních informací → 23,1% hlásek, které jsou poznávány pomocí TRAPS, není poznáváno pomocí MFCC.