A6M33BIO - Biometrie

Biometrické metody založené na rozpoznávání hlasu I

Doc. Ing. Petr Pollák, CSc.

26. listopadu 2012 - 22:15

Obsah přednášky

- **Rozpoznávání řečníka**
  - Úvod
  - Typy úloh automatického rozpoznávání řečníka
- **Základní popis řeči (hlasu)**
  - Anatomie hlasového ústrojí
  - Signálový model vzniku řeči
  - Významné charakteristiky řečového signálu
- **Příznaky používané při rozpoznávání řečníka**
  - Obecné požadavky
  - Základní frekvence řeči
  - Formanty
- **Expertní verifikace řečníka**
  - Lingvisticko-fonetické metody
  - Spektrografické metody

I. část

Rozpoznávání řečníka

Aplikační oblasti úlohy rozpoznávání mluvčího

- ověření totožnosti mluvčího z hlediska bezpečnosti
  - kriminalistická a soudní praxe - forenzní aplikace (dosud subjektivní fonetická a lingvistická analýza)
  - identifikace pro přístup k zabezpečeným systémům (bankovní účty, vstupy do chráněných objektů)
  - motivace pro použití
    - náhrada složitěji realizovatelných systémů
    - přirozenost hlasové komunikace

- *identifikace mluvčího s největší podobností hlasu*
  - *komplexní rozpoznávače řeči (LVCSR - diktovací systémy, transkripční systémy pro přepis rozhlasových/TV zpravodajství)*
    - *modely pro konkrétního mluvčího*
    - *skupinové modely*
    - *modely závislé na pohlaví mluvčího*

Úlohy rozpoznávání mluvčího

Používaná řešení :

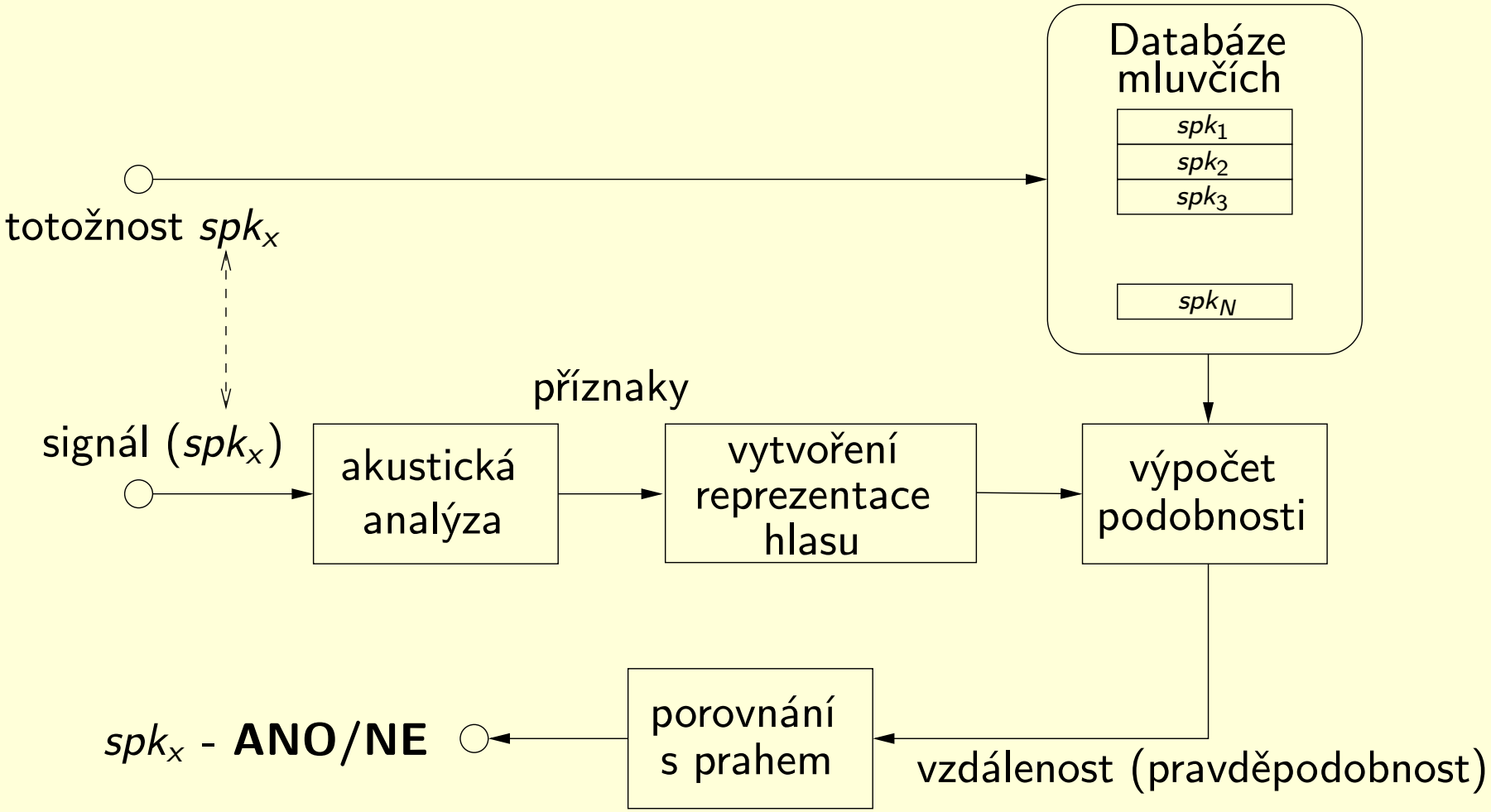
- expertní rozhodování (fonetici, lingvisté)
- automatizované vyhodnocování

Základní úlohy automatického rozpoznávání mluvčího :

1. **Verifikace mluvčího**
   - ověření předpokládané totožnosti mluvčího

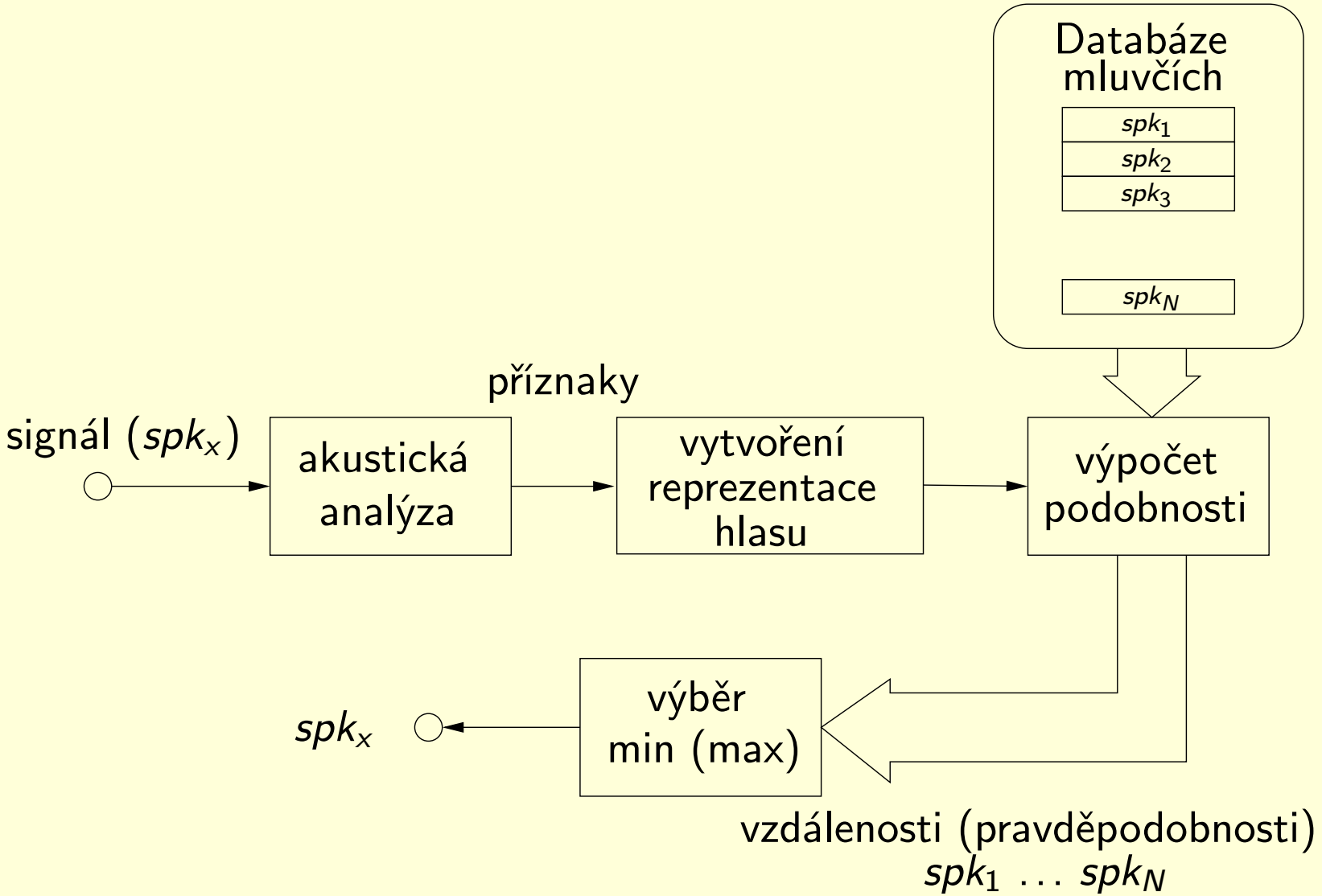
2. **Identifikace mluvčího**
   - **Identifikace v uzavřené množině**
     rozpoznání neznámého mluvčího z dané množiny mluvčích
   - **Identifikace v otevřené množině**
     rozpoznání neznámého mluvčího z neomezené množiny mluvčích $\longrightarrow$ identifikace & verifikace

Verifikace mluvčího



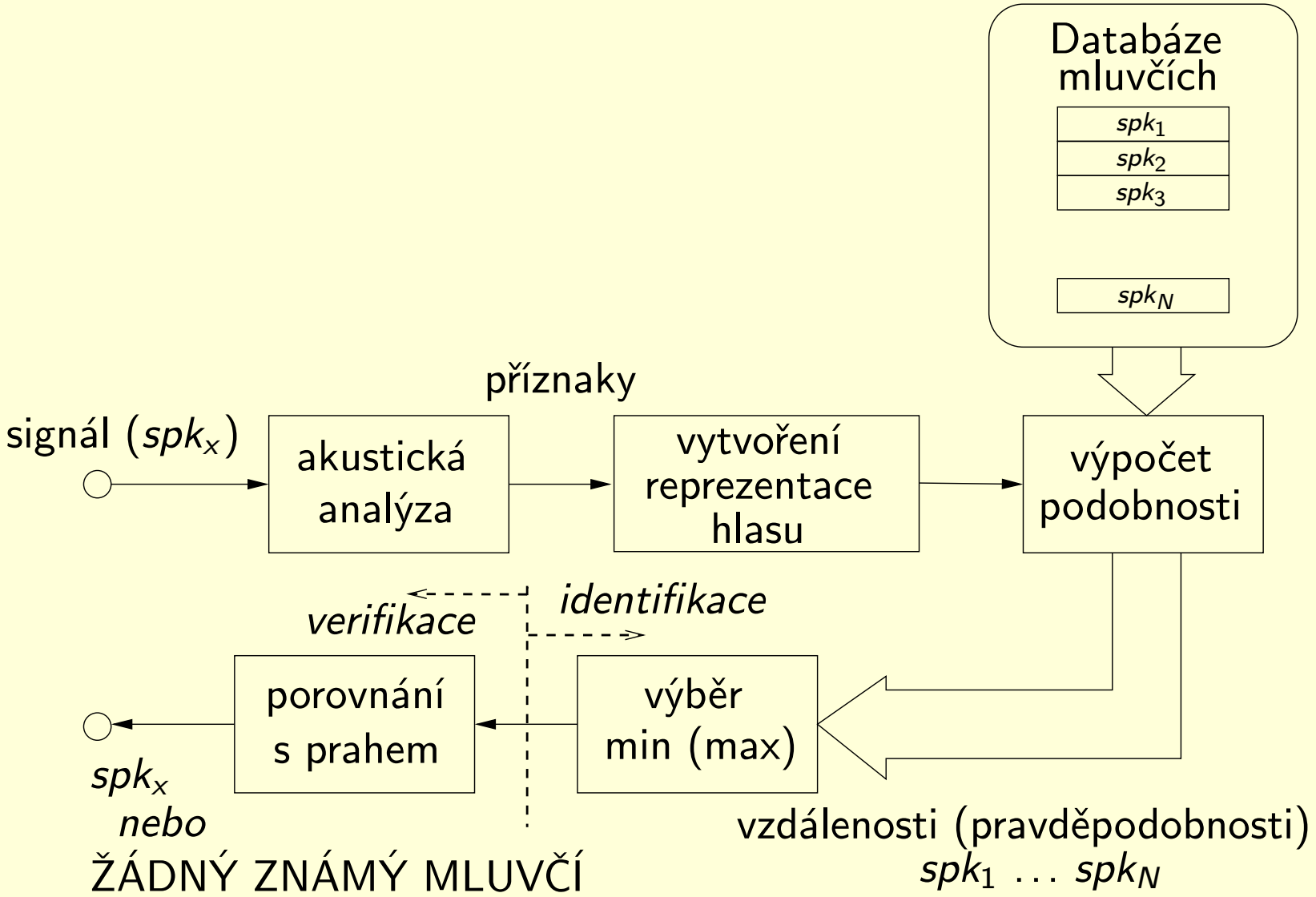
- ověření předpokládané totožnosti mluvčího
- **VÝSLEDEK** = **přijetí / odmítnutí** předpokl. totožnosti

Identifikace mluvčího (v uzavřené množině)



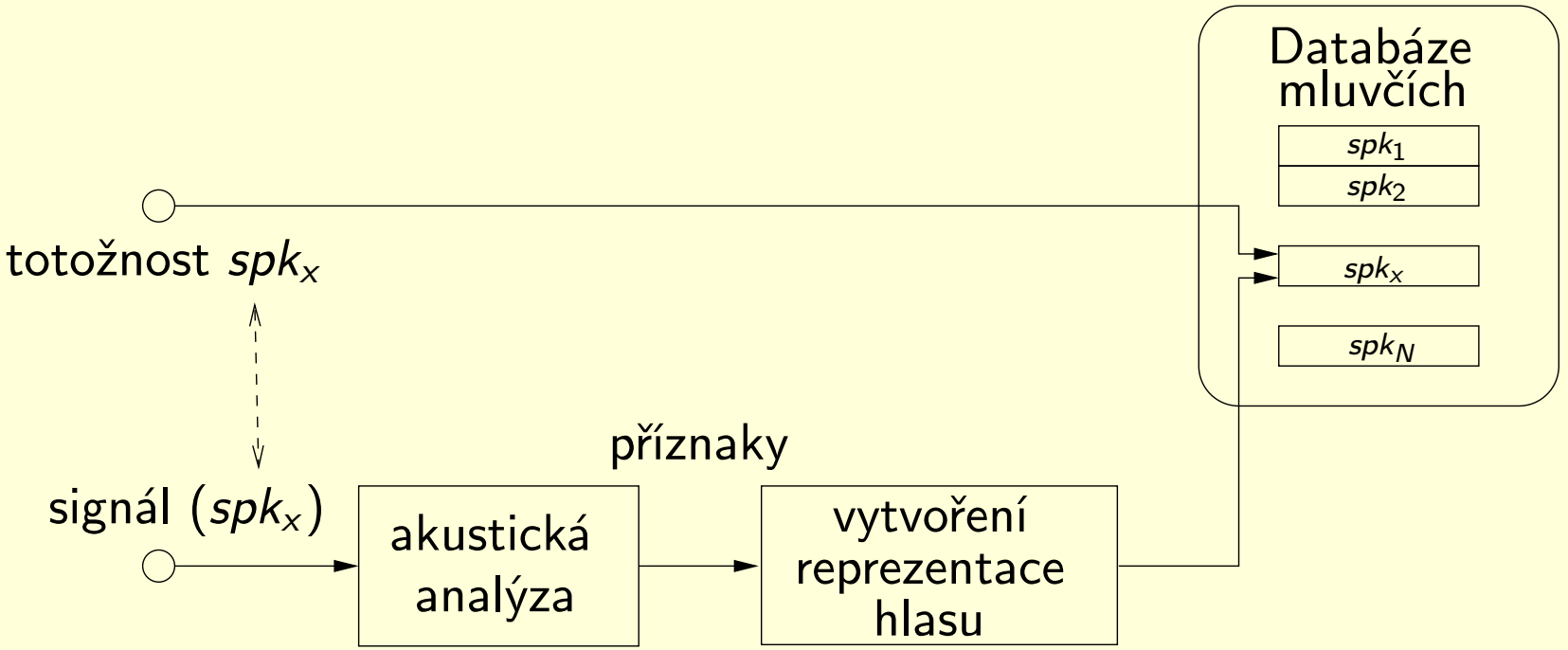
- rozpoznání neznámého mluvčího (největší podobnost hlasu)
- **VÝSLEDEK = ID mluvčího / skupiny**

Identifikace mluvčího (v otevřené množině)



- rozpoznání neznámého mluvčího (největší podobnost hlasu)
- **VÝSLEDEK** = **ID mluvčího / skupiny** nebo **ZAMÍTNUTÍ**

Vytvoření modelů referenčních mluvčích



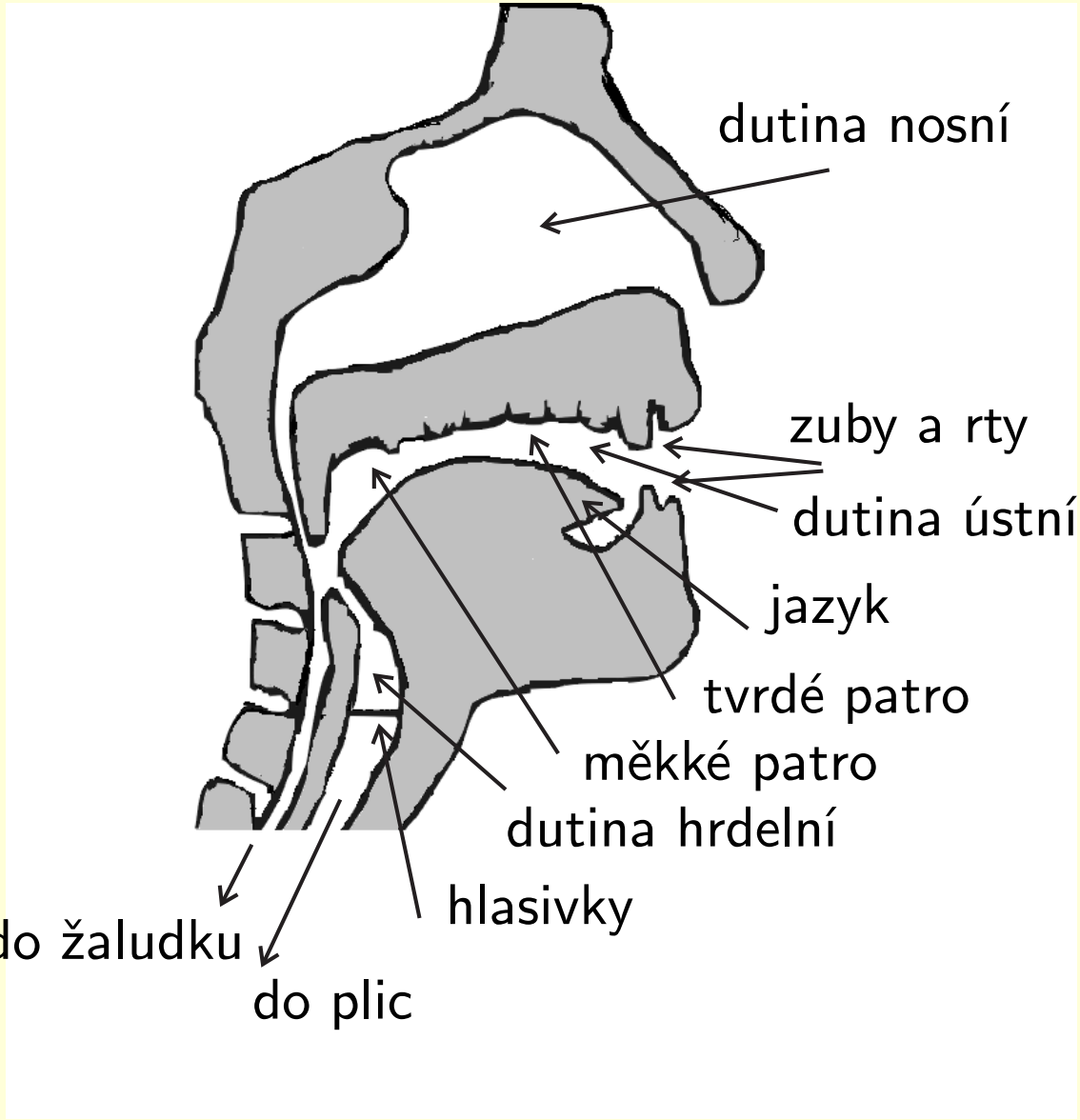
Reprezentace mluvčích v referenční databázi:

- referenční promluvy (DTW)
- statistické modely rozložení příznaků (GMM)
- kódové knihy příznaků (VQ)

II. část

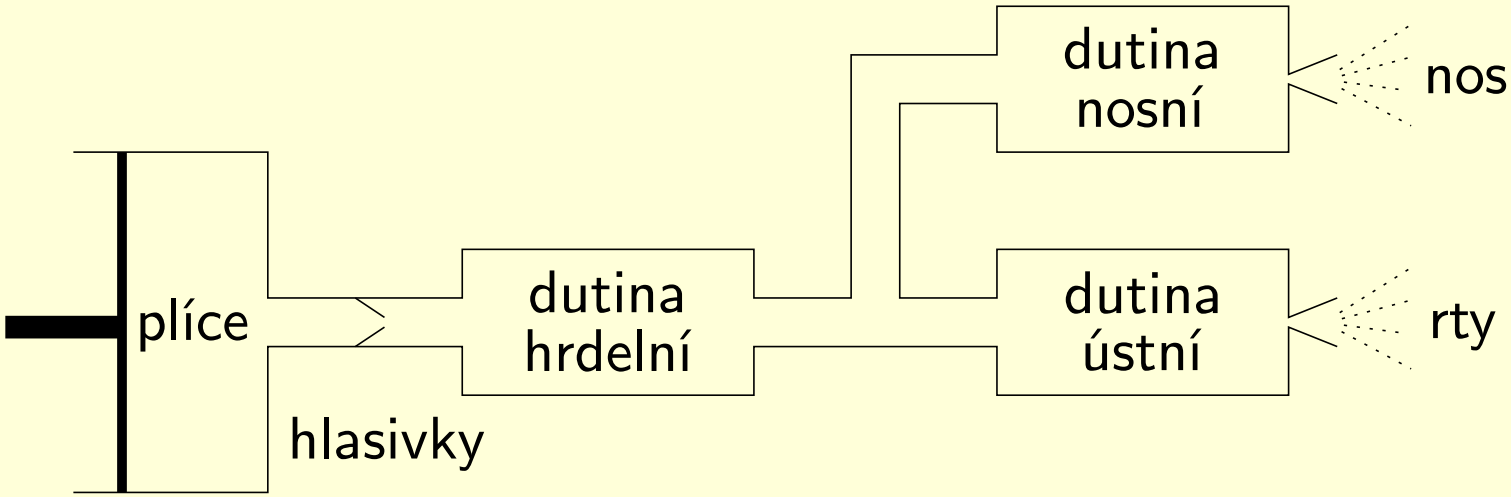
Model vzniku řeči

Artikulační orgány hlasového ústrojí člověka



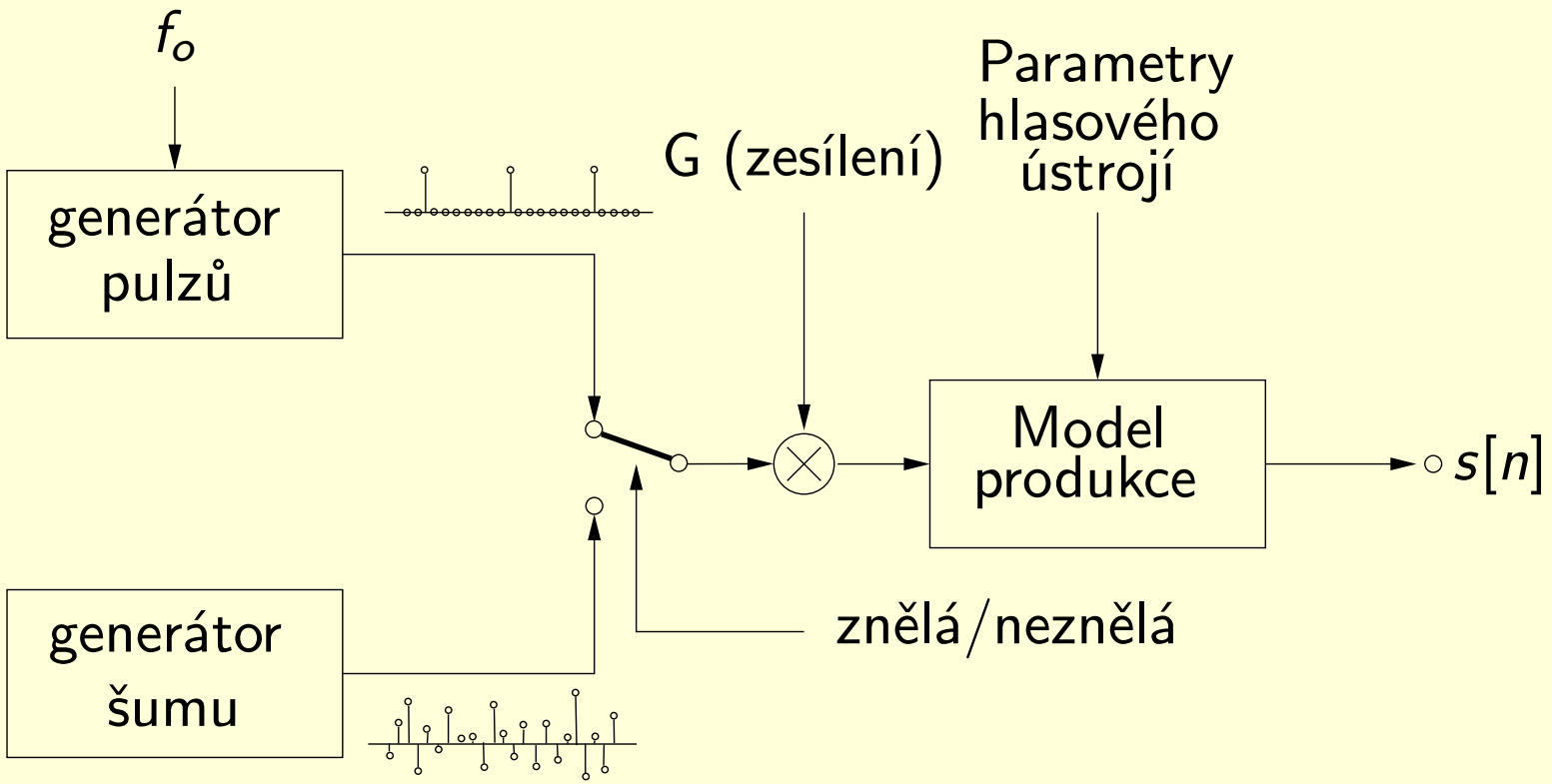
Model vzniku řeči

Model hlasového ústrojí člověka



Model vzniku řeči

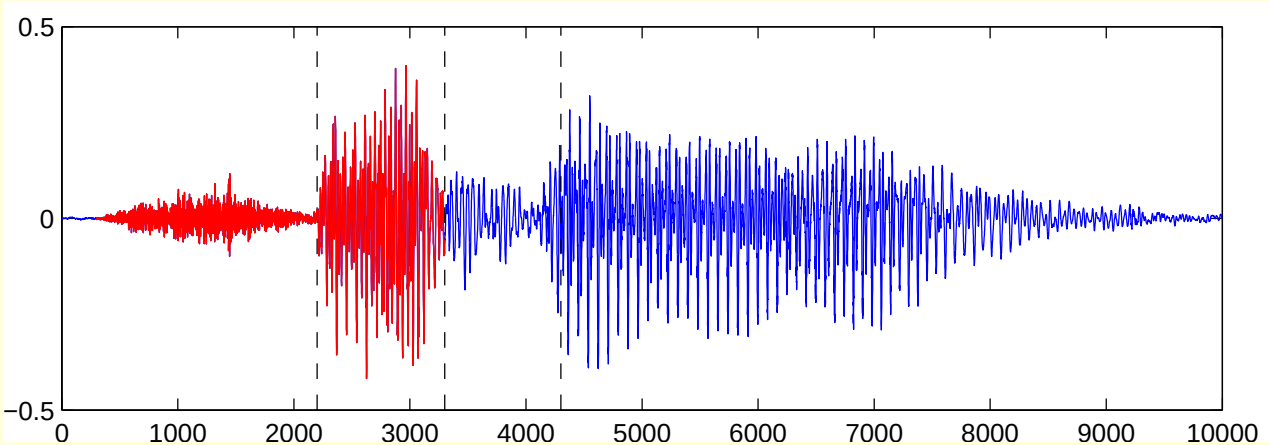
Model generování řečového signálu



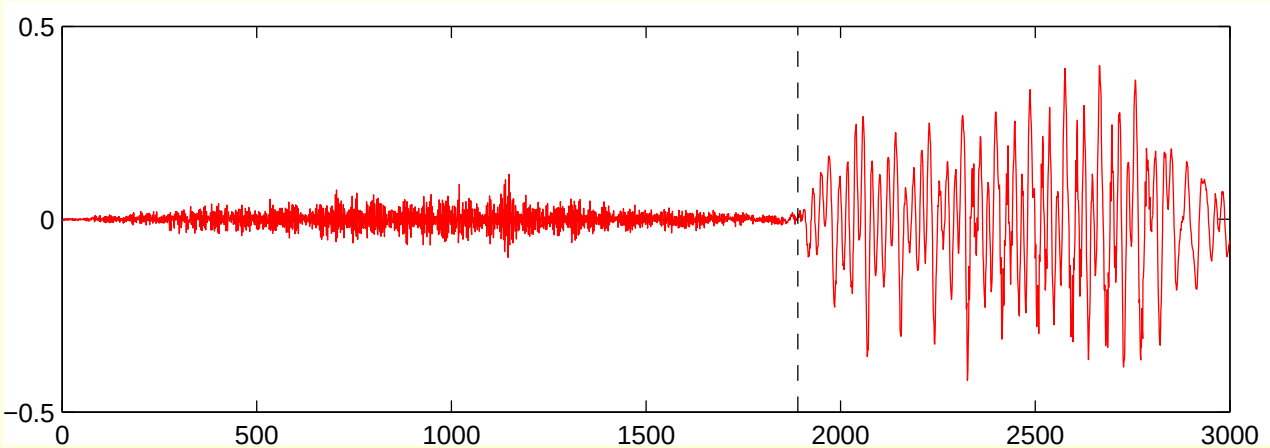
Model produkce řeči - AR model - nejjednodušší model
- snadná identifikace parametrů AR modelu pomocí LPC analýzy
- souvislost s rezonátory hlasového ústrojí

Řečové hlásky v časové oblasti

Slovo **“šedý”**



Slabika **“še”**

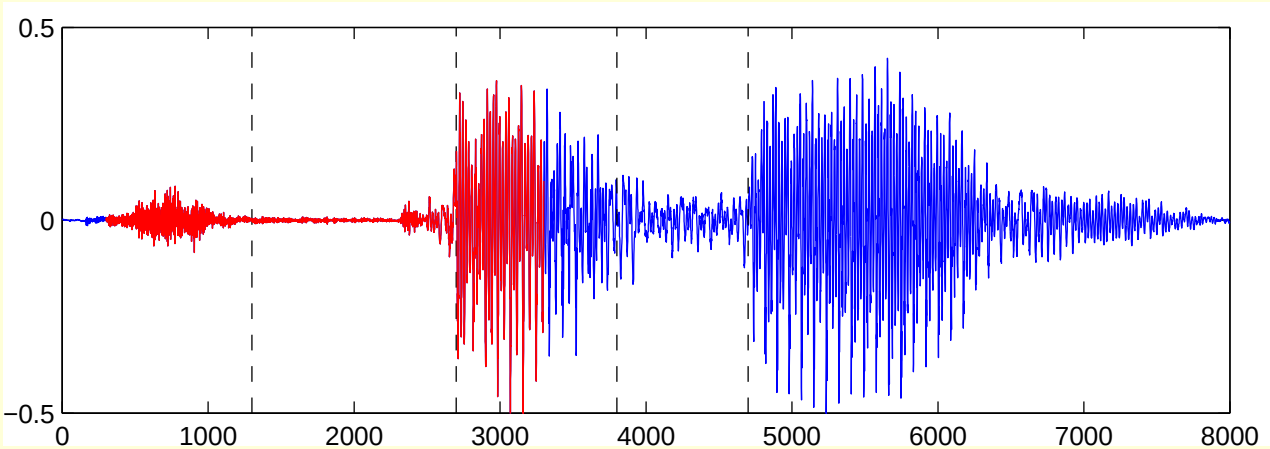


Hláska **“š”** ... neznělá, šumový charakter

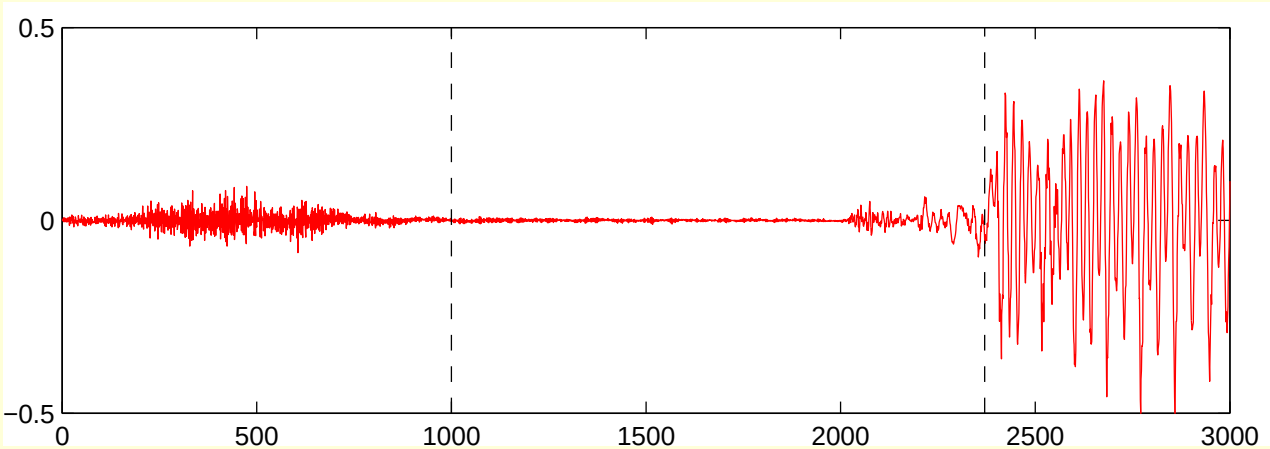
Hláska **“e”** ... znělá, periodický charakter (harmonická struktura)

Řečové hlásky v časové oblasti

Slovo **“čtyři”**



Slabika **“čty”**

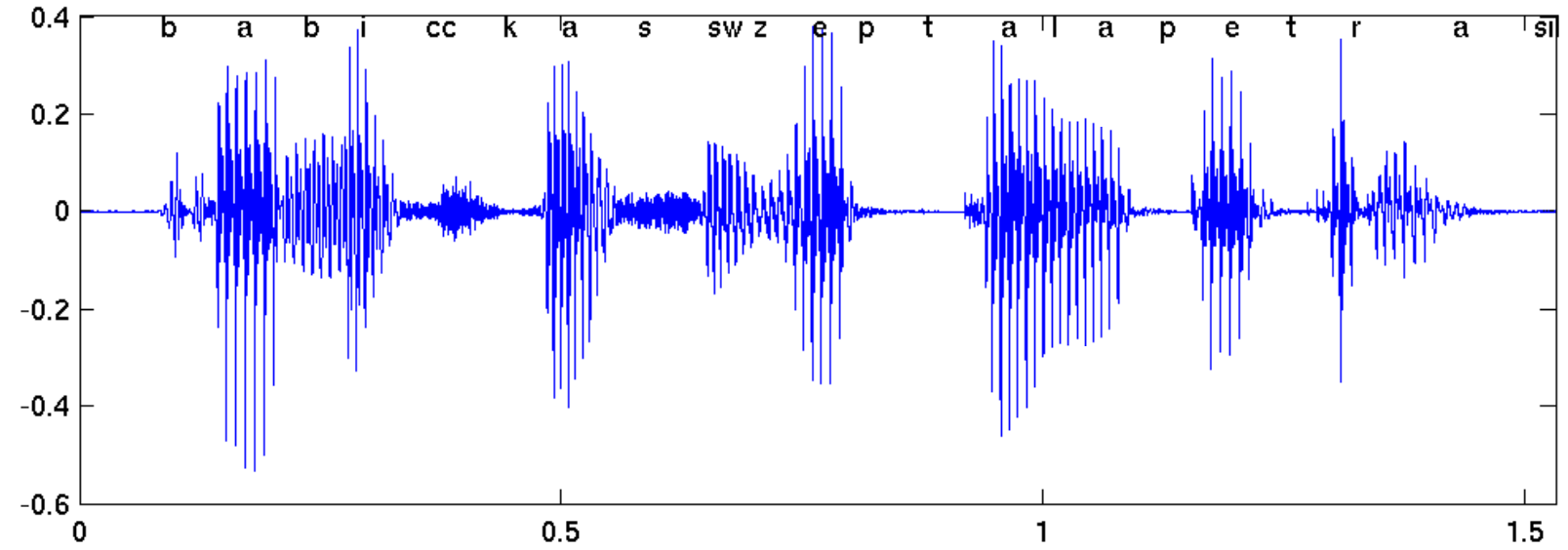


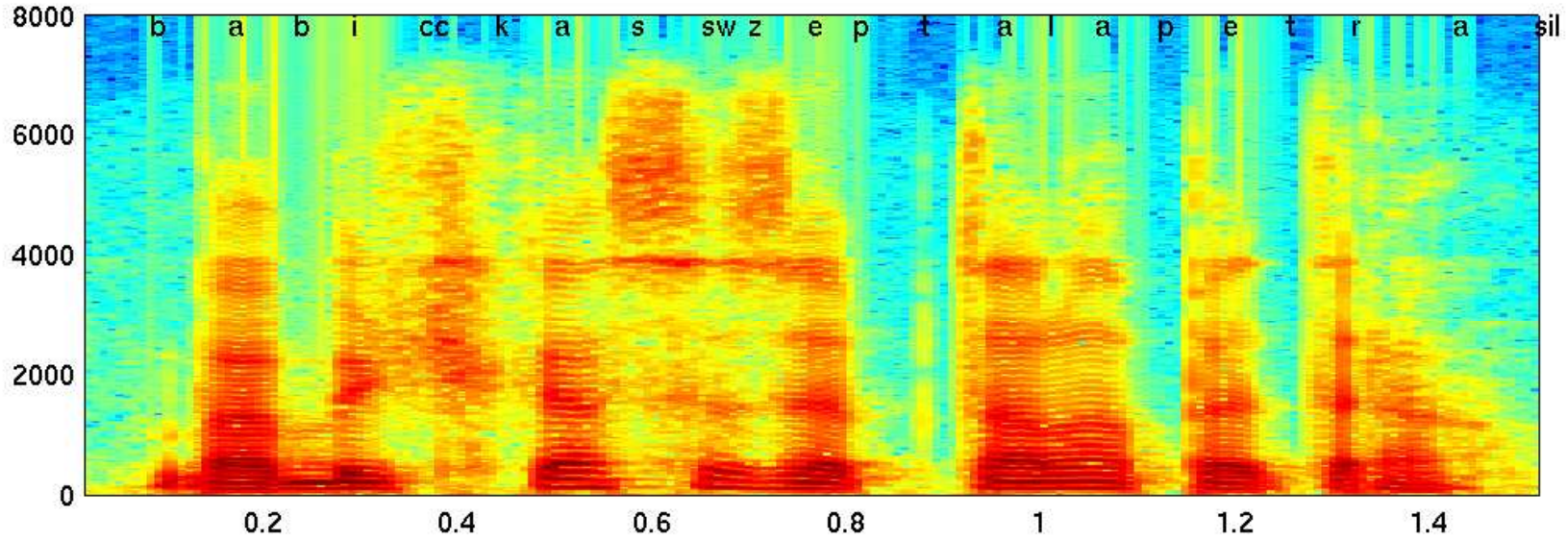
Hláska **“č”** ... neznělá, šumový charakter

Hláska **“t”** ... plozivní, okluze (závěr) + exploze

Hláska **“y”** ... znělá, periodický charakter (harmonická struktura)

Časová a spektrální reprezentace promluvy





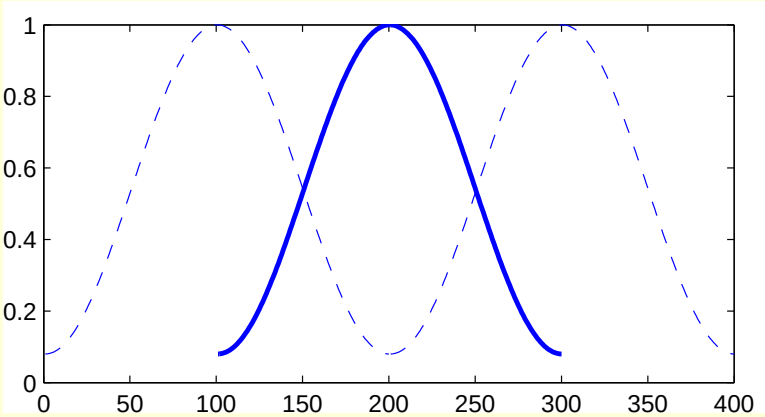
Spektrální analýza řeči na bázi DFT

Specifické vlastnosti:

- **řeč je obecně nestacionární signál** $\Rightarrow$ nutná segmentace a sledování vývoje krátkodobého spektra (spektrogram)
- **řeč je kvazistacionární**
  (tj. stacionární v krátkém časovém intervalu - cca 10-100 ms)
  $\Rightarrow$ 20-30 ms - typická délka krátkodobého segmentu
- **DFT spektrum je ovlivněno prosakováním**
  $\Rightarrow$ nutné váhování vhodným oknem (Hammingovo)
  $\Rightarrow$ nutná segmentace s překryvem (obvykle 50%)

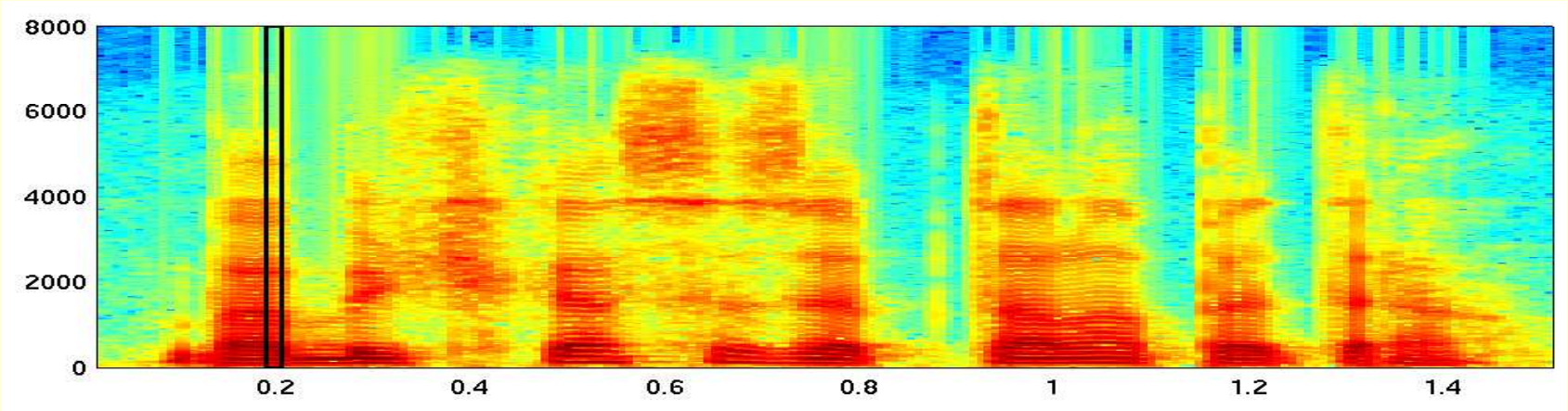
$$w[n] = 0,54 - 0,46 \cos \frac{2\pi n}{N}$$

$$\text{pro } 0 \leq n \leq N-1 \,.$$



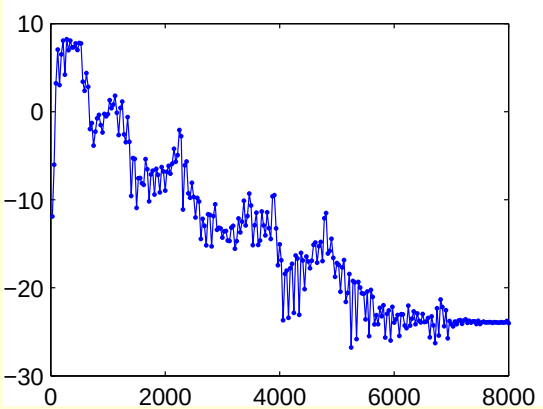
Přehled možností spektrální reprezentace promluvy

Spektrogram celé promluvy



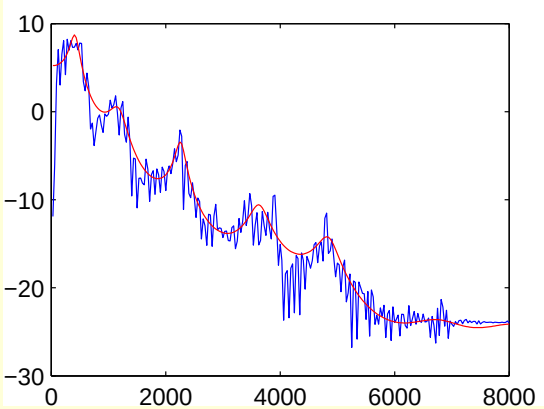
Spektrální reprezentace vybraného segmentu

DFT spektrum:



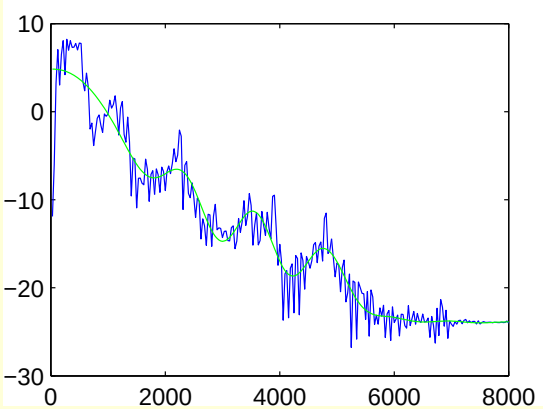
256 vzorků spektra
(amplitudové sp.)

LPC reprezentace:



16 koeficientů a_k
(autoregresní koef.)

Kepstrální koeficienty:



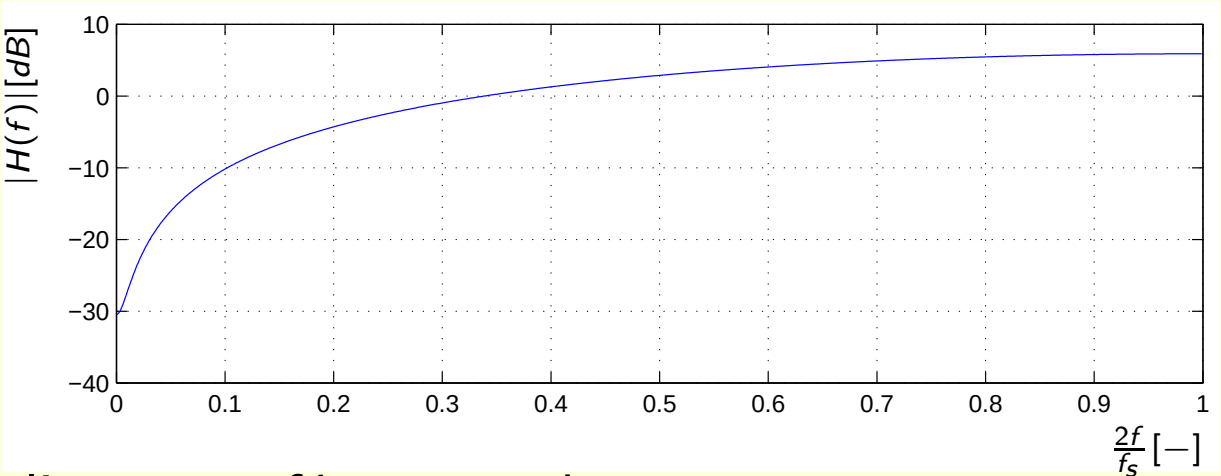
16 koeficientů c_n
(reálné kepstrum)

Sklon amplitudového spektra - vyšší kmitočty - nižší energie

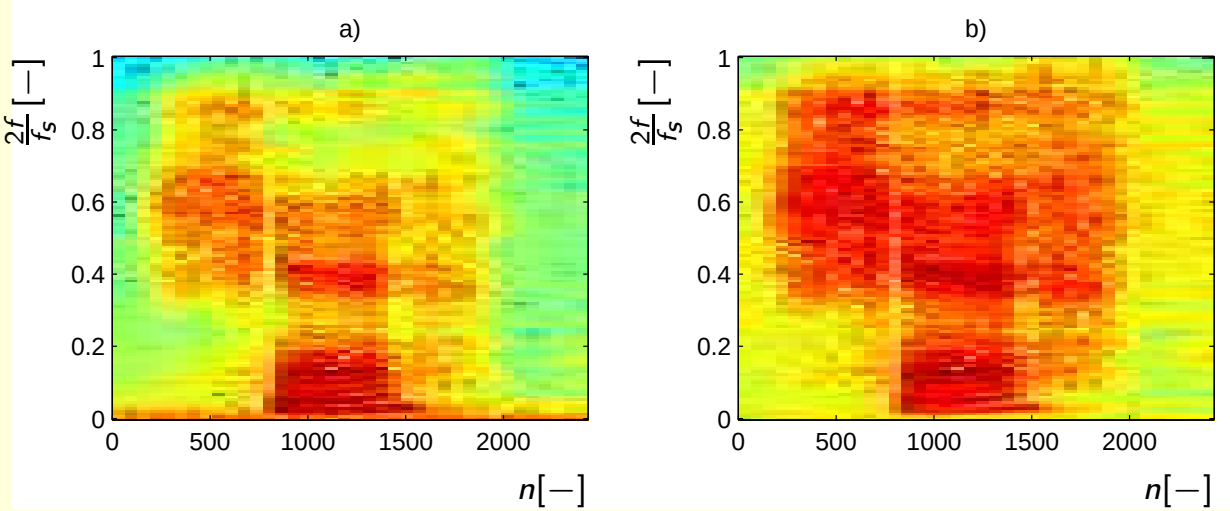
Preemfáze signálu

$$s'[n] = s[n] - m \cdot s[n-1], \qquad m = 0,97$$

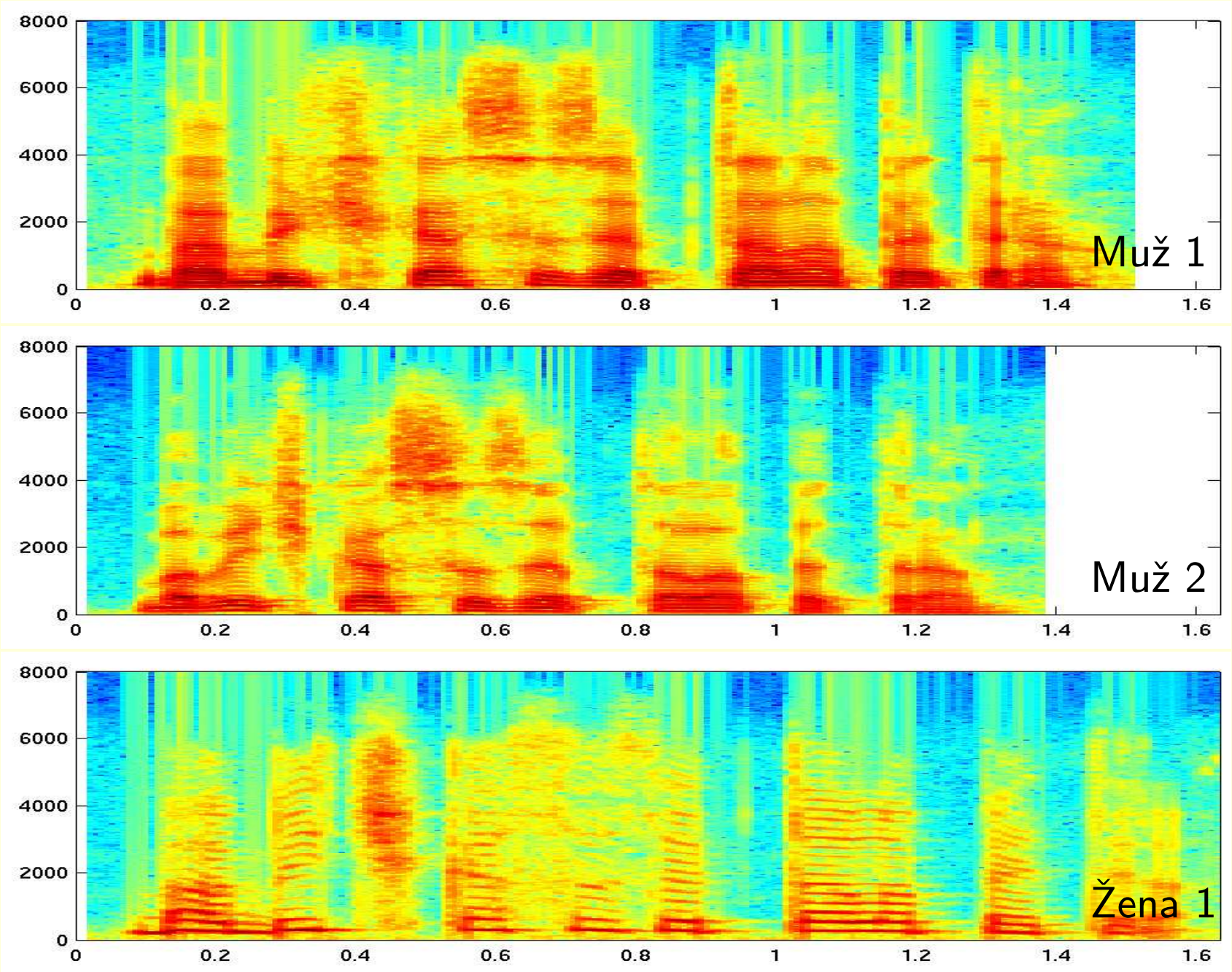
Frekvenční charakteristika preemfázového filtru



Ilustrace vlivu preemfáze ve spektrogramu
(kompenzace sklonu amlitudového spektra)

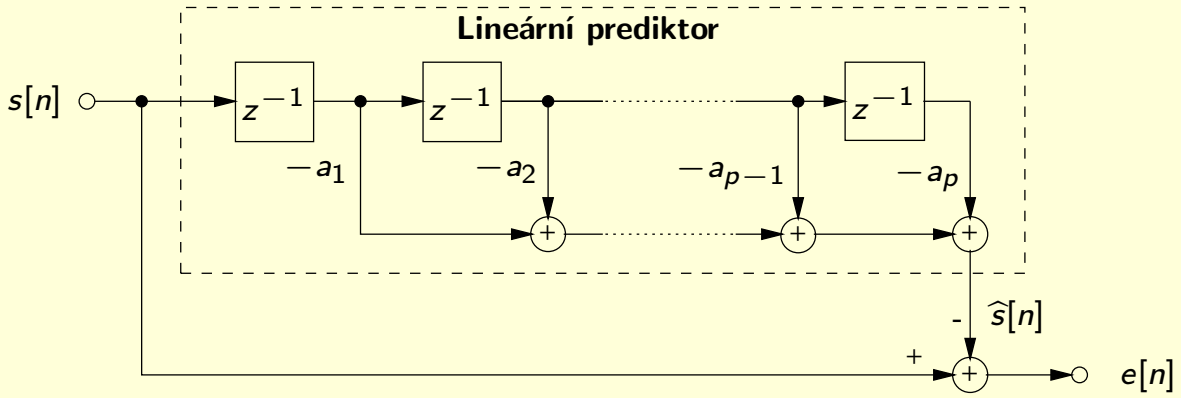


Variabilita stejné promluvy pro různé mluvčí



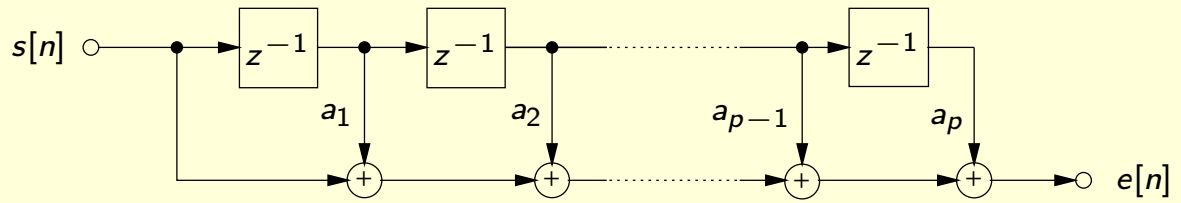
Lineární prediktivní analýza

Lineární predikce : $\widehat{s}[n] = -\sum_{k=1}^{p} a_k s[n-k]$.



Chybový signál (míra kvality prediktoru)

$$e[n] = s[n] - \widehat{s}[n] = s[n] + \sum_{k=1}^{p} a_k s[n-k] = \sum_{k=0}^{p} a_k s[n-k] \ .$$



Princip LPC analýzy

IDEA: přesnější predikce $\longrightarrow$ nižší úroveň chybového signálu

Kritérium - výkon chybového signálu

$$J = E\left\{e^2[n]\right\}$$

Hledání koeficientů a_k $\equiv$Minimalizace chyby predikce
$\equiv$ hledání minima J, i.e.

$$\frac{\partial J}{\partial a_k} = 0\ , \quad \text{for} \quad k\ = 1, 2, ..., p \quad \Rightarrow \quad p \text{ lineárních rovnic}$$

Řešení a metody výpočtu (pro různé definice J):

- **autokorelační metoda** - nejčastěji používaný přístup
- Levinson-Durbinův algoritmus (rychlý výpočet autokor.met.)
- Burgův algoritmus - vychází z křížové struktury filtru

Autokorelační metoda, Yuleovy-Walkerovy rovnice

$$\begin{bmatrix} R[0] & R[1] & R[2] & \dots & R[p-1] \\ R[1] & R[0] & R[1] & & R[p-2] \\ R[2] & R[1] & R[0] & \ddots & R[p-3] \\ \vdots & & \ddots & \ddots & \vdots \\ R[p-1] & R[p-2] & R[p-3] & \dots & R[0] \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R[1] \\ R[2] \\ \vdots \\ \vdots \\ R[p] \end{bmatrix}$$

$R[k]$ .... autokorelační koeficienty analyzovaného signálu

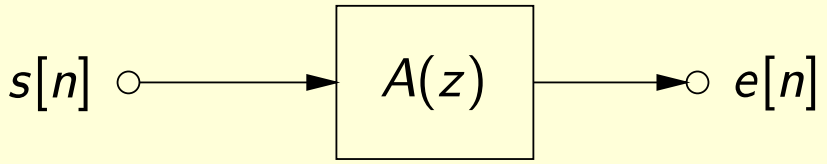
VÝSLEDEK:

a_k .... autoregresní koeficienty (AR model signálu)

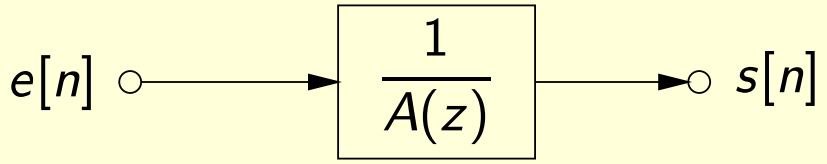
$P_p = R[0] + \sum_{k=1}^{p} a_k R[k]$ .... výkon chybového signálu

AR model signálu

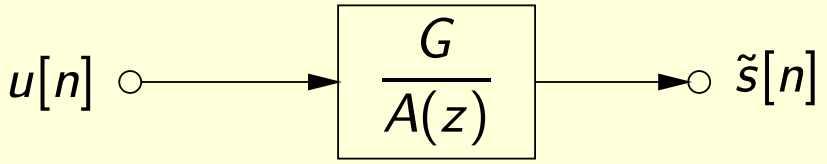
Dekorelační (analyzující) filtr : $A(z) = \sum_{k=0}^{p} a_k z^{-k}$



Syntéza se skutečným chybovým signálem (ideální případ)



Syntéza s umělým signálem s jednotkovým výkonem (AR model)
- G závisí na úrovni analyzovaného signálu ($G = \sqrt{P_p}$)



Spektrální vlastnosti AR modelu

Obecný popis AR modelu v Z-oblasti

$$\tilde{S}(z) = H(z) \cdot U(z)$$

Popis AR modelu ve frekvenční oblasti

$$S_{\tilde{s}}(e^{j\Theta}) = |H(e^{j\Theta})|^2 \cdot S_u(e^{j\Theta})$$

Vlastnosti a důsledky: - $S_u(e^{j\Theta})$ je ploché
$\rightarrow$ tvar $S_{\tilde{s}}(e^{j\Theta})$ je kompletně zahrnut v AR modelu

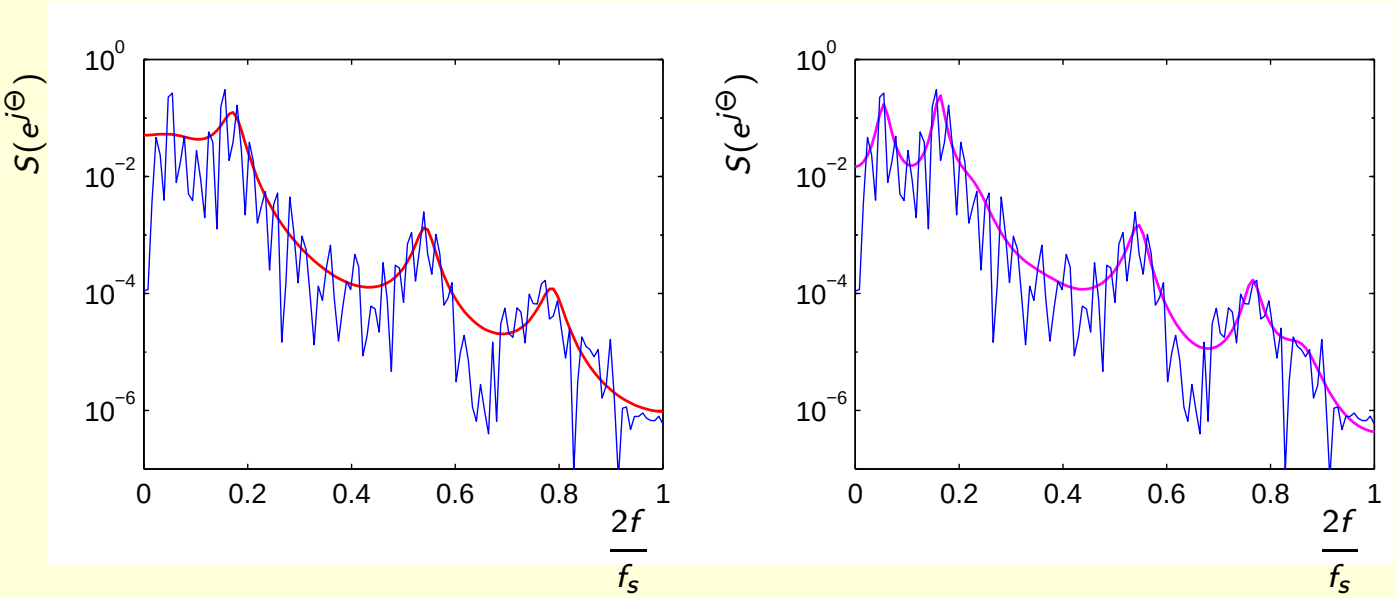
$$\Downarrow$$

LPC spektrum (pokud $S_u(e^{j\Theta}) = 1$)

$$S_{\tilde{s}}(e^{j\Theta}) = |H(e^{j\Theta})|^2 = \frac{G^2}{|A(e^{j\Theta})|^2}$$

Srovnání LPC a DFT spektra

$$S_{\tilde{s}}(e^{j\Theta}) = |H(e^{j\Theta})|^2 \quad \approx \quad \frac{|S[k]|^2}{N}$$



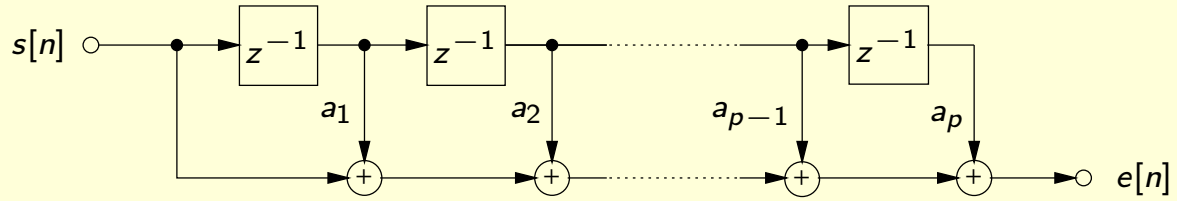
Vyhlazené odhady spektrální výkonové hustoty pomocí LPC:
a) vyhlazený odhad PSD pomocí LPC, $p = 10$,
b) vyhlazený odhad PSD pomocí LPC, $p = 16$.

Vlastnosti LPC spektra

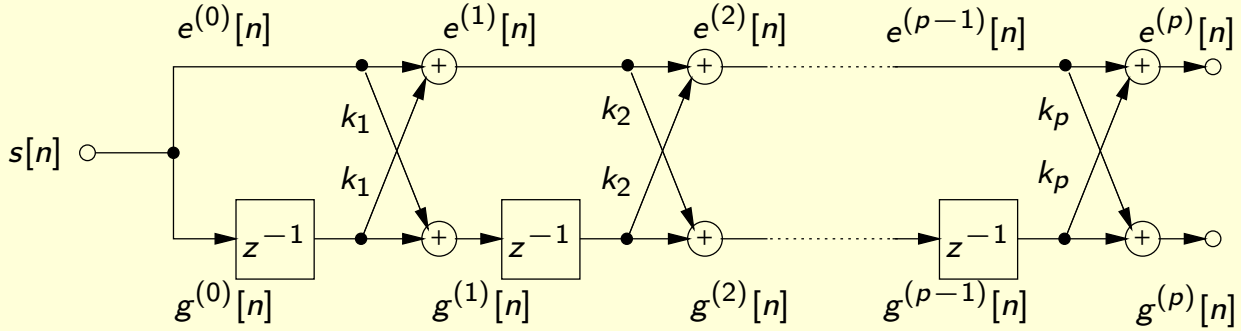
- AR model = "all-pole" filtr
  $\longrightarrow$ modeluje pouze špičky ve spektru
  $\longrightarrow$ rezonátory v dutinách vokálního traktu
- reálný pól modeluje špičku v 0 nebo $f_s/2$
- obecná špička je modelována dvojicí komplexně združených pólů
- vyšší řád AR modelu znamená více špiček v LPC spektru
  $\longrightarrow$ typické hodnoty pro řeč:
  - $p = 12$ pro $f_s = 8$ kHz
  - $p = 16$ pro $f_s = 16$ kHz

Křížová struktura AR modelu

Trasnverzální struktura analyzujícího FIR filtru:



Křížová struktura analyzujícího FIR filtru:



k_k ..... koeficienty odrazu, přepočet k_k vs. a_k - Levinsonova rekurze

Inicializace:

$$a_1^{(1)} = k_1$$

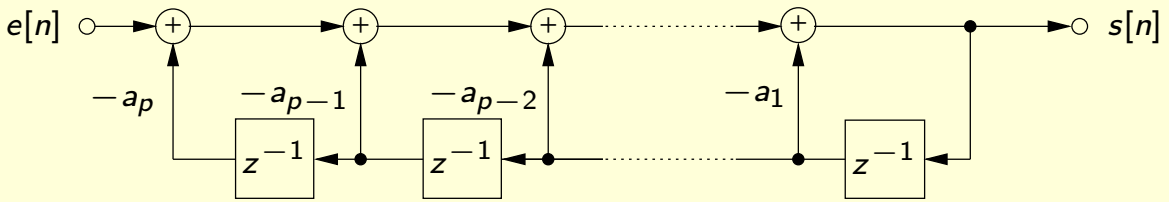
Výpočet pro $m = 2, 3, \ldots, p$**:**

$$a_m^{(m)} = k_m$$

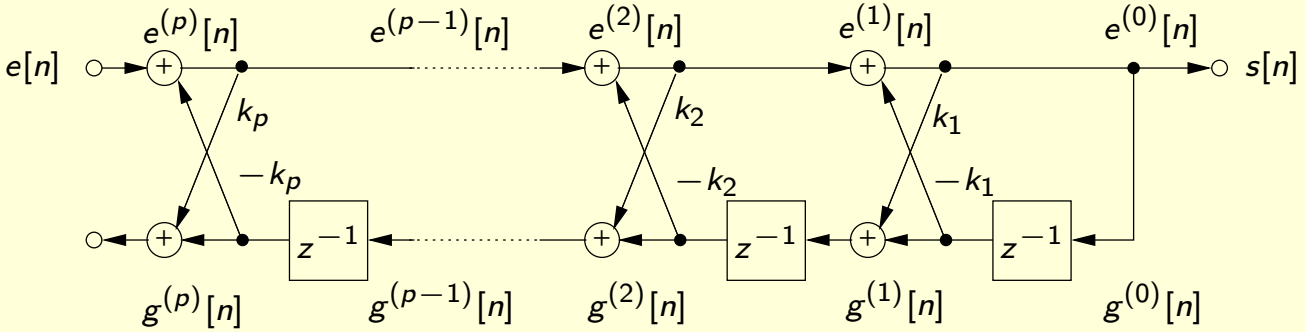
$$a_j^{(m)} = a_j^{(m-1)} + k_m a_{m-j}^{(m-1)}, \quad j = 1, 2, \ldots, m-1$$

Křížová struktura AR modelu

Trasnverzální struktura syntetizujícího all-pole IIR filtru:

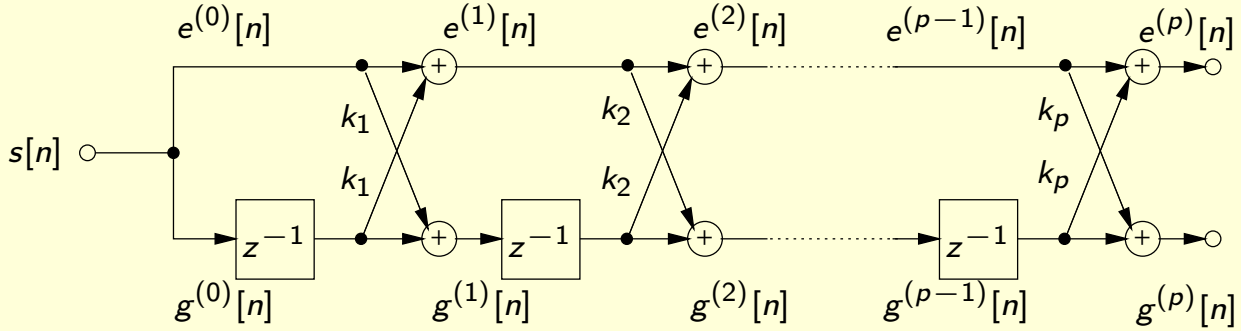


Křížová struktura syntetizujícího all-pole IIR filtru:



Burgův algoritmus

Křížová struktura analyzujícího FIR filtru:



Minimalizační kritérium (pro každou sekci křížové struktury):

$$J = \frac{1}{2} \sum_{n=0}^{N-1} \left[\left(e^{(m)}[n] \right)^2 + \left(g^{(m)}[n] \right)^2 \right] \qquad \text{pro } m = 1,\ 2,\ \ldots,\ p\ .$$

Výpočet m-tého koeficientu odrazu:

$$k_m = - \frac{2 \cdot \sum_{n=m}^{N-1} \left(e^{(m-1)}[n] \cdot g^{(m-1)}[n-1] \right)}{\sum_{n=m}^{N-1} \left(e^{(m-1)}[n] \right)^2 + \sum_{n=m}^{N-1} \left(g^{(m-1)}[n-1] \right)^2}$$

Autoregresní koeficienty a_k - výpočet Levinsonovou rekurzí (L.-D. alg.)

III. část

Řečové příznaky pro rozpoznávání řečníka

Řečové příznaky v úlohách rozpoznávání mluvčího

Obecné požadavky pro příznaky resp. systémy identifikace

- **vysoká variabilita pro různé mluvčí**
- **nízká variabilita pro jednoho mluvčího**
  (možné vlivy - aktuální stav, nálada, stres, hluk, styl promluvy)

- snadný a efektivní výpočet
- odolnost vůči šumu a zkreslení (výše zmiňované jevy)
- odolnost proti imitaci hlasu

$\Downarrow$

Vnitřní příznaky - související s hlasovým ústrojím člověka

$\Downarrow$

- přesnější a komplexnější rozhodování v případě verifikace
  $\longrightarrow$ formanty, f_o, speciální příznaky (často textově závislé)
- *obecné příznaky (používané zejména v komplexním systému rozpoznávání řeči s identifikací řečníka)* $\longrightarrow$ *kepstra*

Řečové příznaky v úlohách rozpoznávání mluvčího

Používané příznaky

- základní frekvence (charakteristika hlasu)
- formantové kmitočty (souvislost s délkou vokálního traktu)

- kepstrální příznaky (MFCC, PLPC) - obecně používané (možnost vyhlazení variability mezi mluvčími)
- LPC kepstrální příznaky (variabilita mezi mluvčími, malá robustnost vůči šumu)
- parametry AR modelu (menší robustnost, Itakurova vzdálenost)

- kombinované vektory příznaků pro komplexnější a sofistikovanější rozhodování

Základní tón řeči

- základní frekvence f_o
- pro znělé hlásky s harmonickou strukturou
- souvisí s kmitáním hlasivek

- hodnota f_o je ovlivněna vlastnostmi hlasivek (pružnost, hmotnost, délka)

$\downarrow$

- různá výška hlasu, různá intonace ve větě u jednotlivých mluvčích

Odhad základního tónu řeči

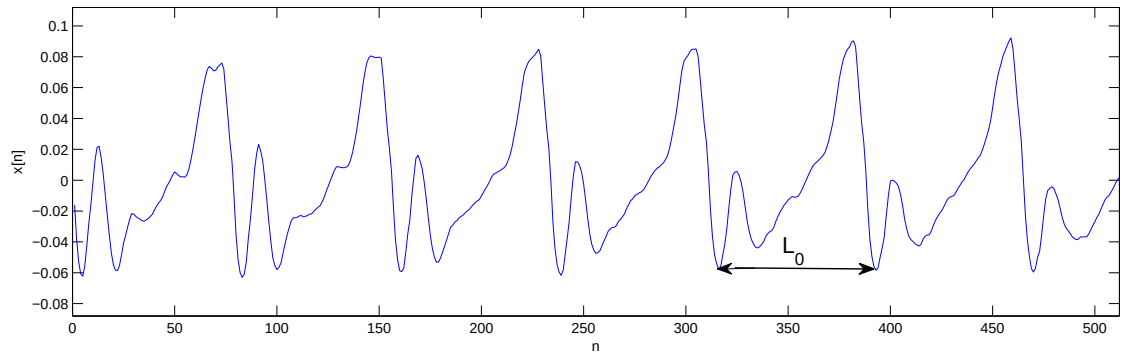
f_o ..... základní tón (frekvence) řeči ..... $f_o = \frac{1}{T_o}$

T_o (L_o) ..... základní perioda (v sekundách vs. ve vzorcích)

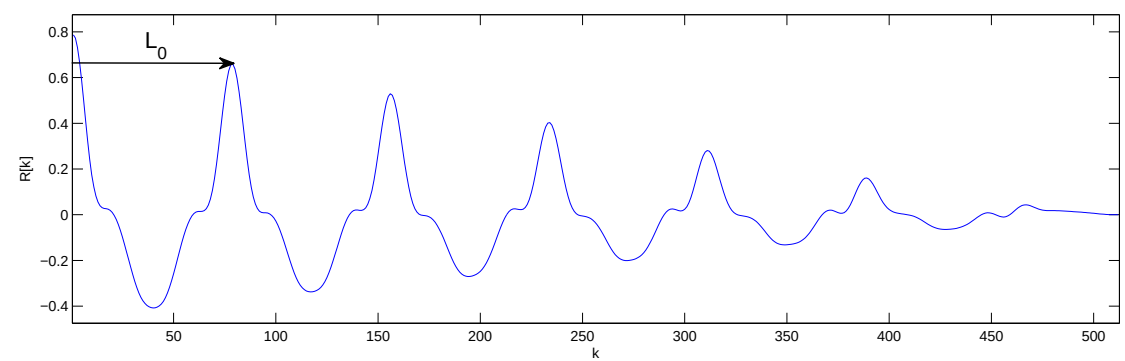
Nejčastější metoda odhadu - na bázi autokorelační funkce

(hledání postranního maxima autokorelační funkce)

segment signálu

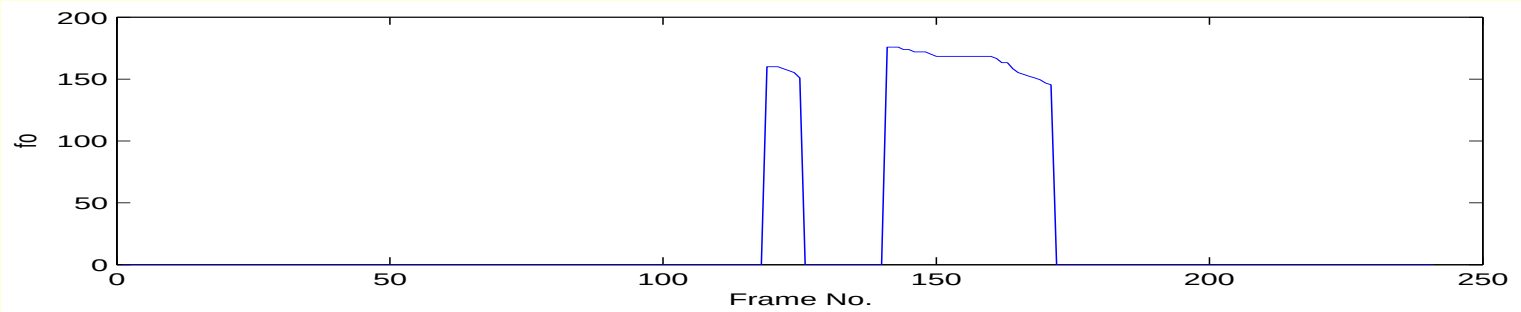


odhad autokorelační funkce

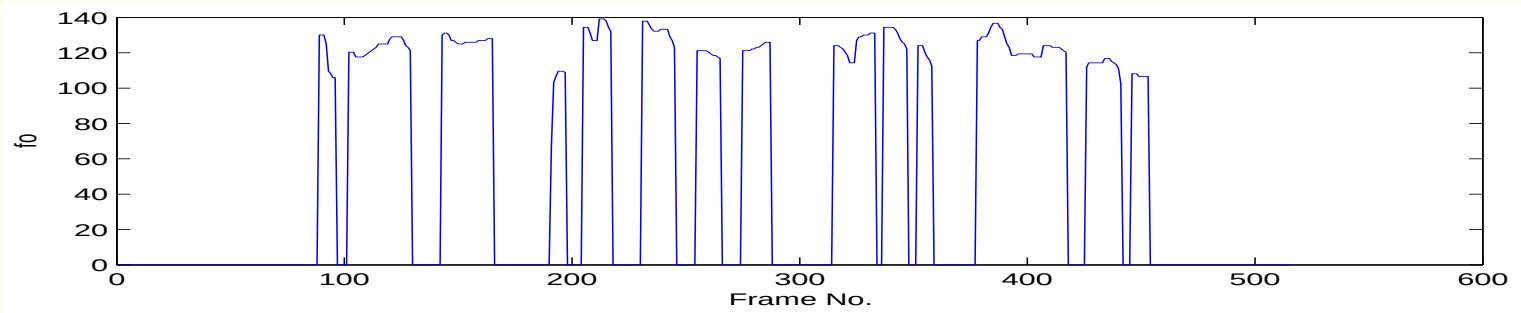


Průběh základního tónu v promluvě

Krátká promluva - slovo



Delší promluva - věta



Průběh f_o v promluvě $\longrightarrow$ získaná (naučená) charakteristika

Průměrná hodnota f_o

Formanty - definice

- **Formant (formantové frekvence)**
  $\longrightarrow$ centrální kmitočty kmitočty rezonátorů vokálního traktu

- významné špičky ve **VYHLAZENÉM krátkodobém spektru**
- významné formanty F1 - F4 v pásmu do 4 kHz
- F5 - méně významný (obtížně odhadnutelný formant)
- !! Nezaměňovat se základním tónem řeči F0 (f_0)
  (f_0 není detekovatelné ve vyhlazeném spektru)

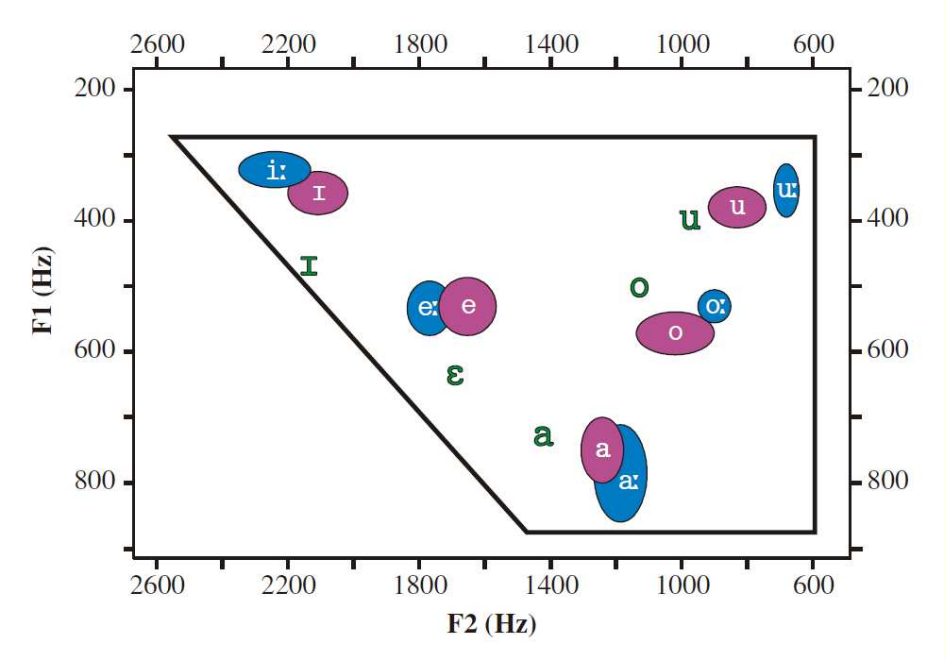
$\downarrow$

Souvislost s fyziologií vokálního traktu = vhodný vnitřní příznak
(délka vok. traktu je nepřímo úměrná formantovým frekvencím)

$$VTL = \frac{(2i-1)\cdot c}{4F_i}$$

Významné formanty samohlásek

	I	E	A	O	U
F1	300 - 500	480 - 700	700 - 1100	500 - 700	300 - 500
F2	2000 - 2800	1560 - 2100	1100 - 1500	850 - 1200	600 - 1000
F3	2600 - 3500	2500 - 3000	2500 - 3000	2500 - 3000	2400 - 2900



	přední	střední	zadní
vysoké	i		u
středové	e		o
nízké		a	

Pro rozlišení mluvčích - vzdálenost sousedních formantů

Formanty - Odhad na bázi LPC

- špičky LPC spektra - rezonátory = formanty
- F_i - formantová frekvence (centrální kmitočet rezonátoru)
- B_i - šířka pásma formantu
- špičky LPC spektra - určené **póly přenosové funkce** p_i

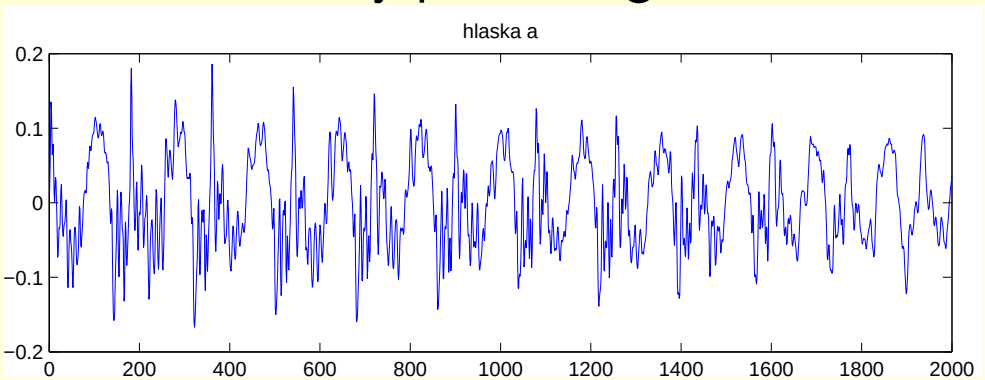
$$F_i = f_s \cdot \arg\, p_i / 2\pi$$

$$B_i = -f_s \cdot \ln |p_i| / 2\pi$$

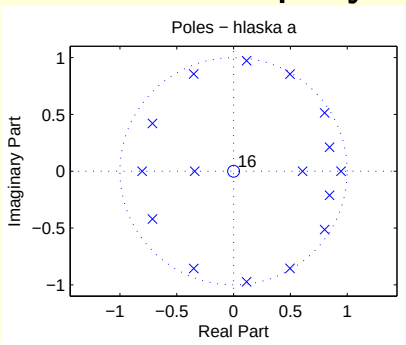
- Problémy:
  - obecně menší robustnost LPC analýzy (závislost na datech)
  - určení vhodného řádu (vliv přítomnosti šumu)
  - seřazení vypočítaných pólů (sledování stejného formantu)
  - vyřazení nadbytečného pólu (méně významné špičky)

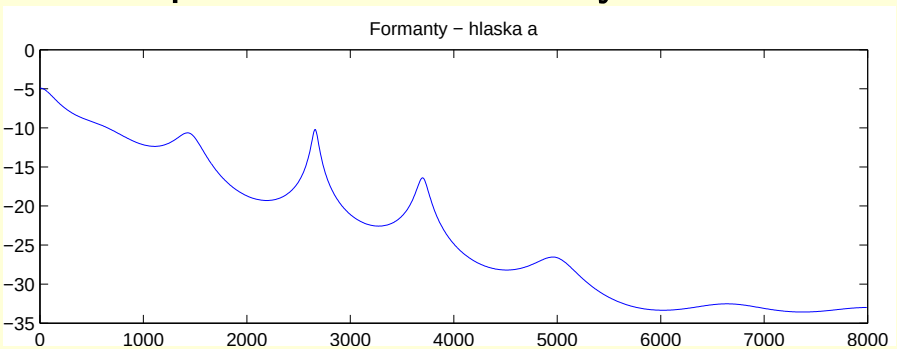
Odhad formantů na bázi LPC - příklad

Časový průběh signálu

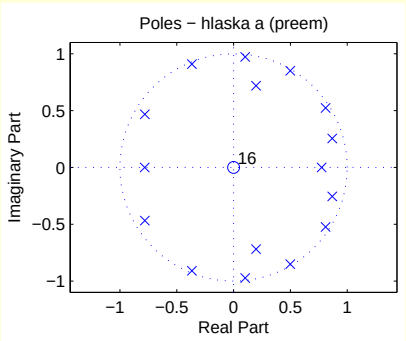


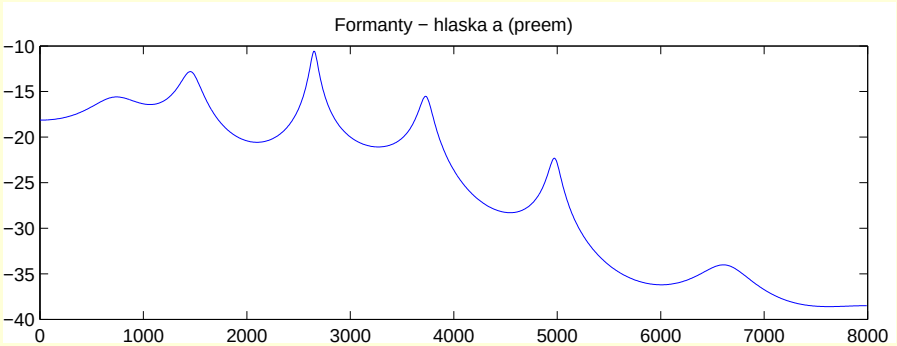
póly & LPC spektrum s formanty





póly & LPC spektrum s formanty (preemfáze)





Speciální příznaky pro rozpoznávání mluvčího

- F2 v “n”
- F3 v “u”
- F2 v “i”
- délka trvání “k”

- ..... *obecnější formulace* .....
- hodnota formantu ve vybrané hlásce
- šířka pásma vybraného formantu ve vybrané hlásce
- směrnice poklesu formantu ve vybrané hlásce
- Průběh F0 ve vybrané větě (slově)
- průměrná hodnota F0 ve větě (slově)
- ..... apod. .....

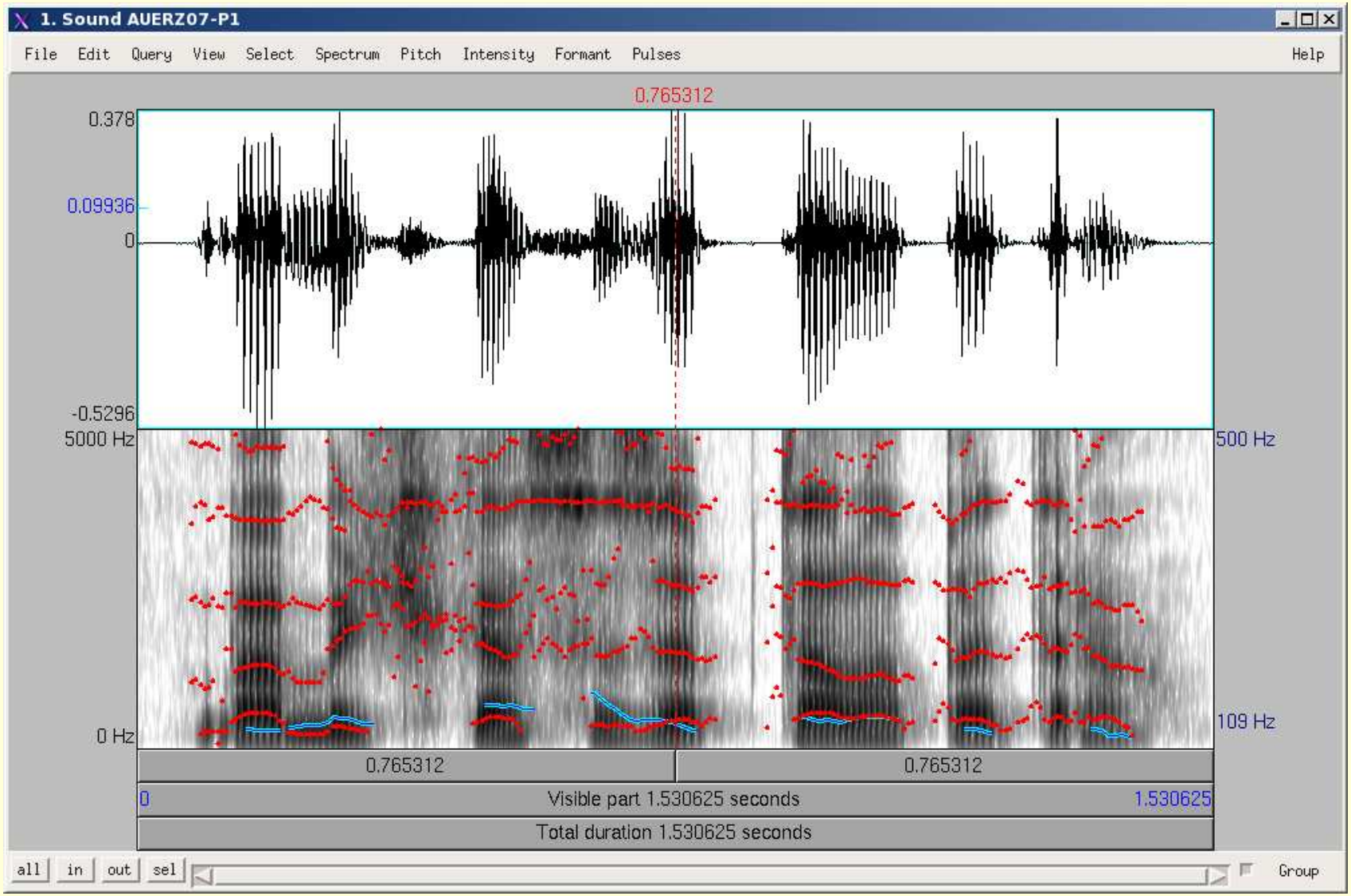
Expertní metody při rozpoznávání mluvčího

- **Forenzní lingvistika a fonetika**
  - sledování osobitých rysů projevu řečníka
  - zaměření na artikulační zvláštnosti
  - typické vedení melodie řeči (intonace)
  - většinou na bázi poslechu

- **Spektrografické metody**
  - Využívají možnost zobrazení diskutovaných hlasových charakteristik (spektrogram, průběhu f_o, trajektorií formantů, apod.)
  - řešeno opět na expertní bázi
  - detaily realizace vybraných hlásek

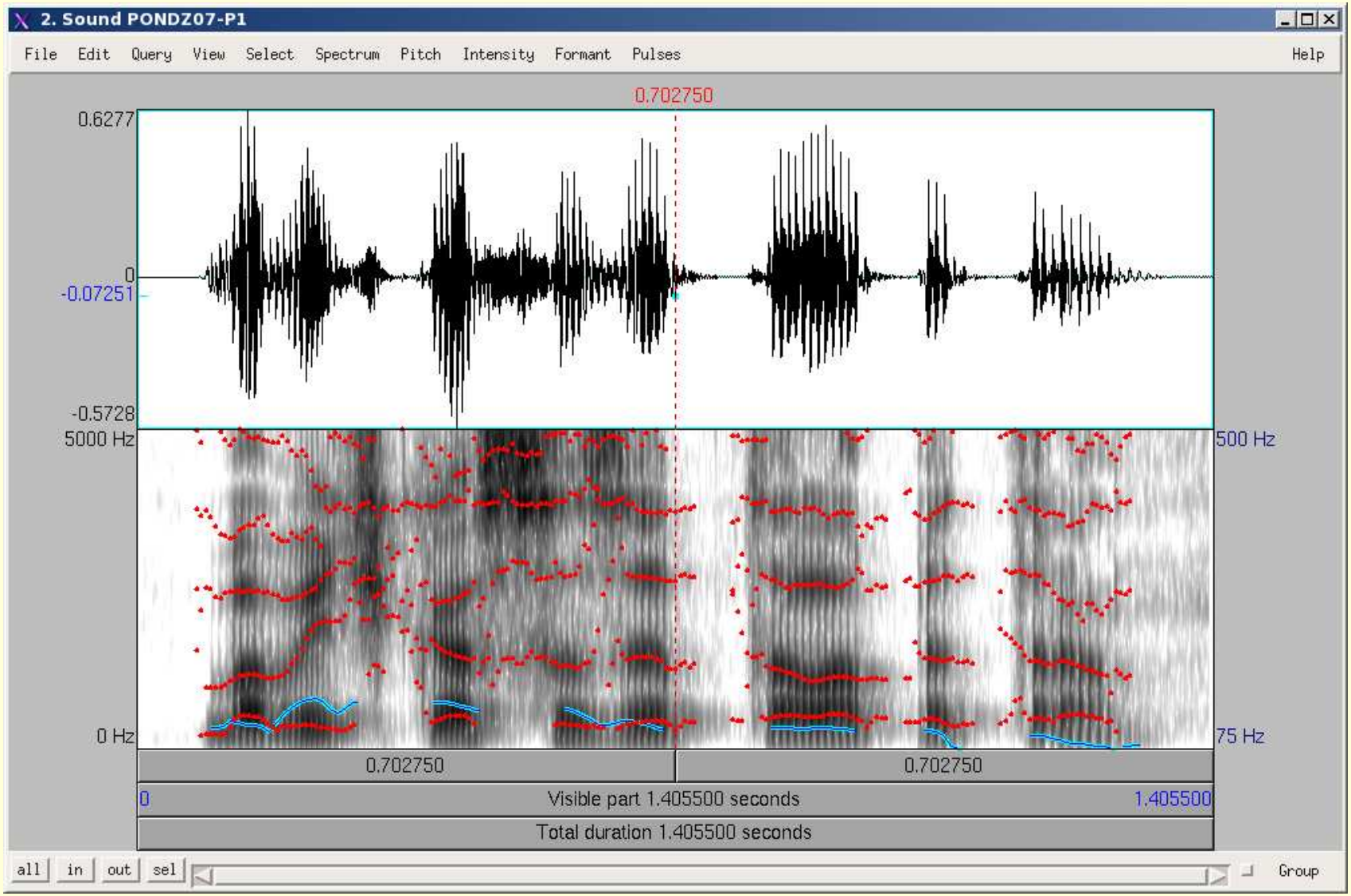
Formanty & základní tón - odhad Praat

Muž 1



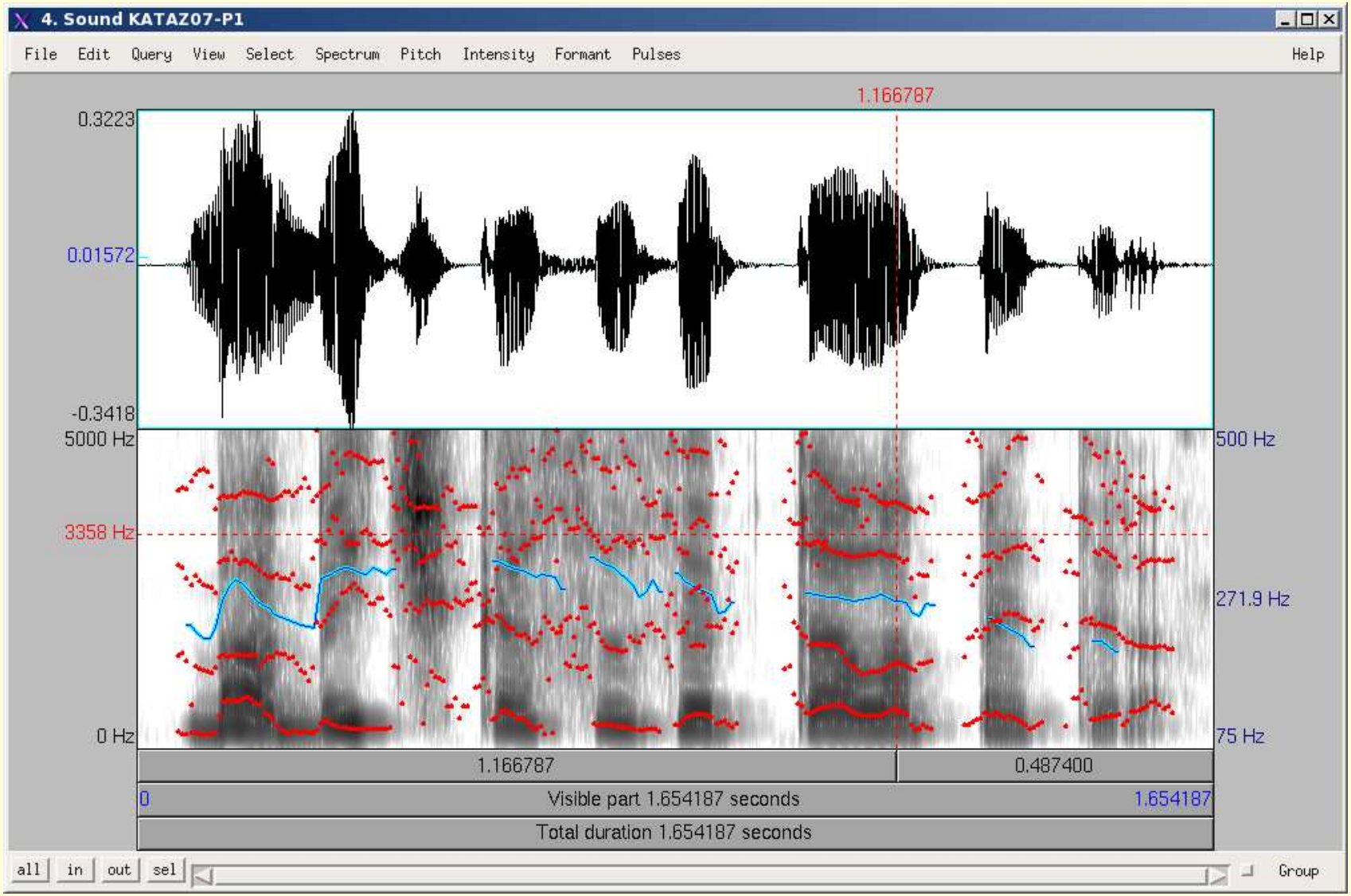
Formanty & základní tón - odhad Praat

Muž 2



Formanty & základní tón - odhad Praat

Žena 1



Děkuji vám za pozornost !