A6M33BIO - Biometrie

Biometrické metody založené na rozpoznávání hlasu I

Doc. Ing. Petr Pollák, CSc.

26. listopadu 2025 - 0:19

Obsah přednášek

- **Úvod**
  - Aplikace a principy hlasové biometrické verifikace
  - Základní popis vzniku řeči (hlasu)
- **Řečové charakteristiky a příznaky pro identifikaci**
  - Základní frekvence řeči
  - Spektrální charakteristiky
  - LCP analýza, formanty
  - Kepstrum, kepstrální vzdálenost
- **Expertní spektrografické metody analýzy**
- **Reprezentace řečníka a metody srovnávání**
  - DTW
  - VQ, GMM, GMM-UBM
  - i-vektory
  - ANN/DNN (x-vektory)
- **Příklady systémů verifikace**

I. část

Hlasová biometrická identifikace (rozpoznávání řečníka)

Úvod, aplikace, princip tvorby řeči, variabilita hlasu

Aplikační oblasti úlohy rozpoznávání řečníka

SRE - Speaker REcognition

1) přesná identifikace/verifikace totožnosti mluvčího

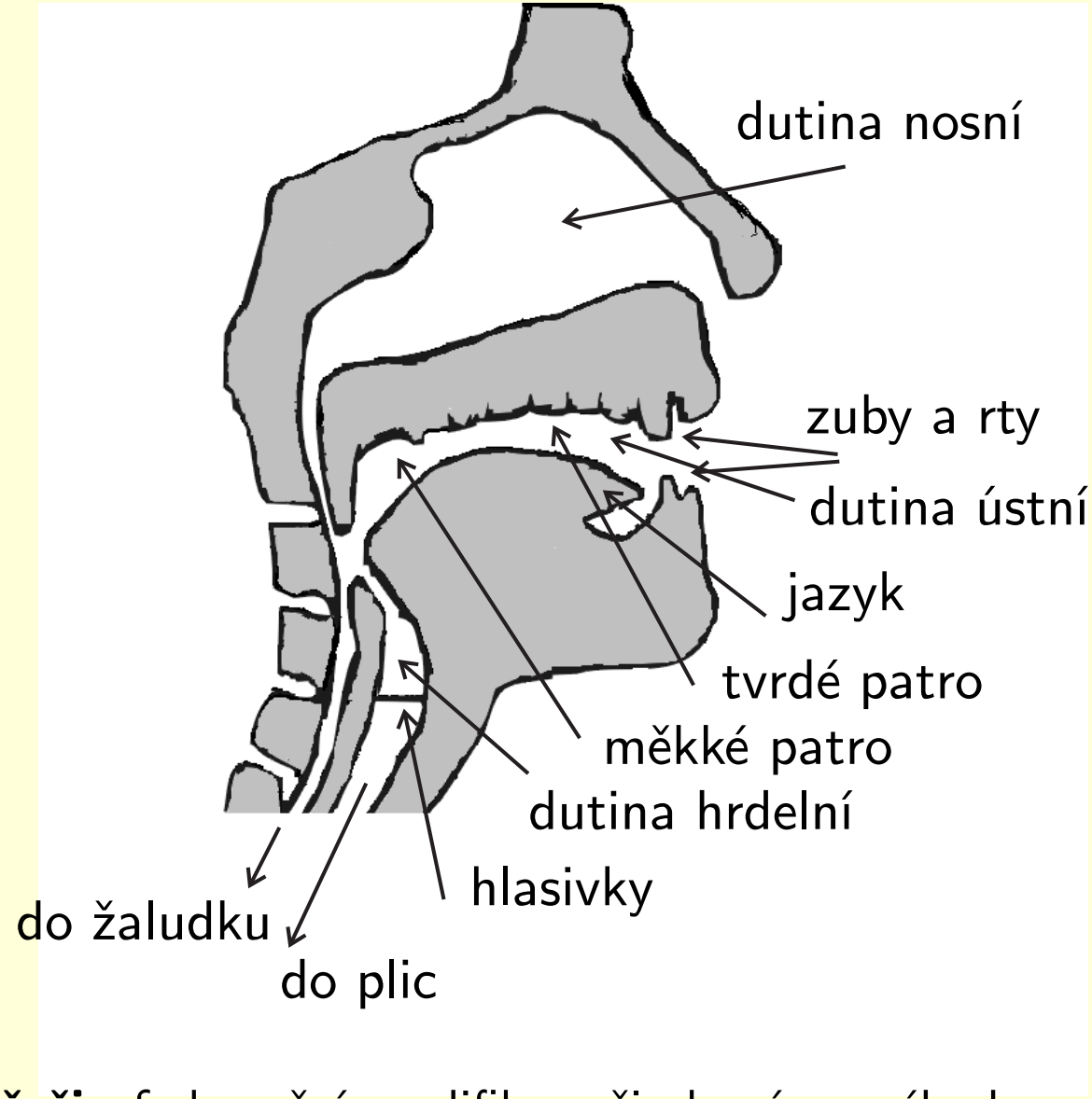
- kriminalistická a soudní praxe - forenzní aplikace (náhrada subjektivní fonetické a lingvistické analýzy)
- identifikace pro přístup k zabezpečeným systémům (osobní PC, mobilní telefony, bankovní účty, přístup do chráněných objektů/systémů, bezpečnostní kontroly, apod.)
- identifikace volajících v call-centrech

2) identifikace mluvčího s největší podobností hlasu

- komplexní rozpoznávače řeči - ASR (LVCSR - diktovací systémy, transkripční systémy pro přepis rozhlasových/TV zpravodajství)
- modely pro identifikovaného mluvčího (GMM-HMM)
  (skupinové modely - pohlaví, nářečí, apod.)
- reprezentace mluvčího na vstupu ASR na bázi DNN

Produkce řeči - fyziologie hlasového ústrojí

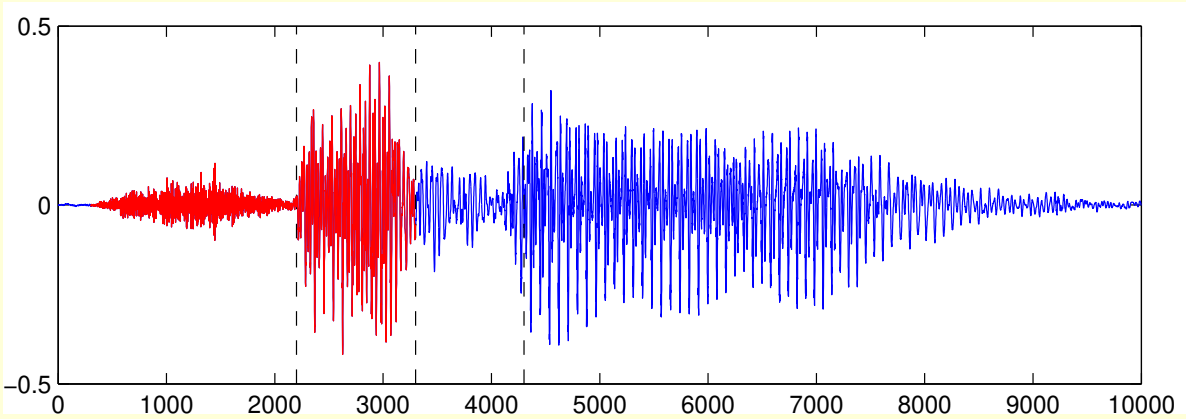
Artikulační orgány hlasového ústrojí člověka



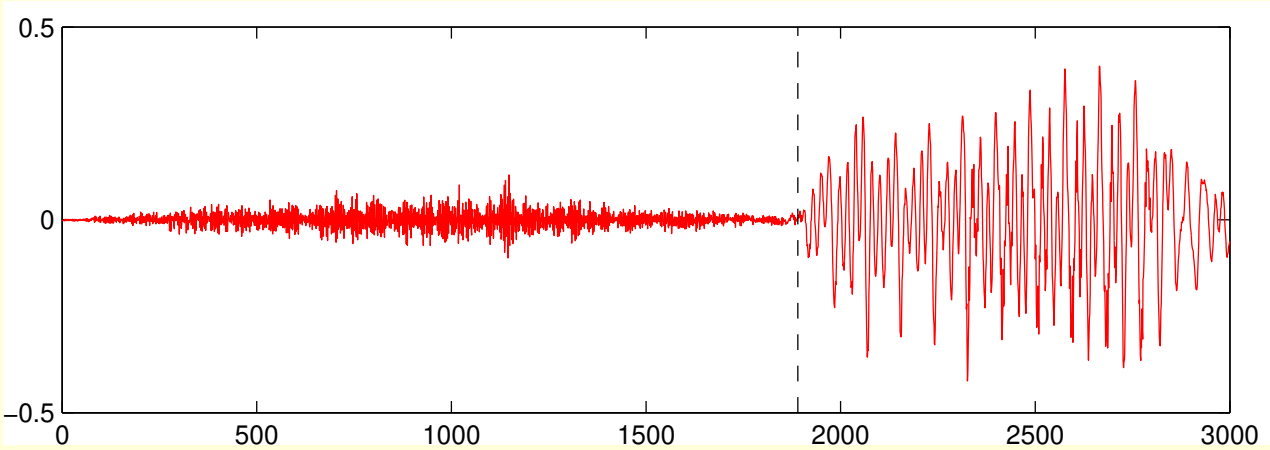
Produkce řeči : frekvenční modifikace širokopásmového buzení proudem vzduchu průchodem dutinami (rezonátory) hlasového ústrojí

Řečové hlásky v časové oblasti

Slovo **“šedý”**



Slabika **“še”**

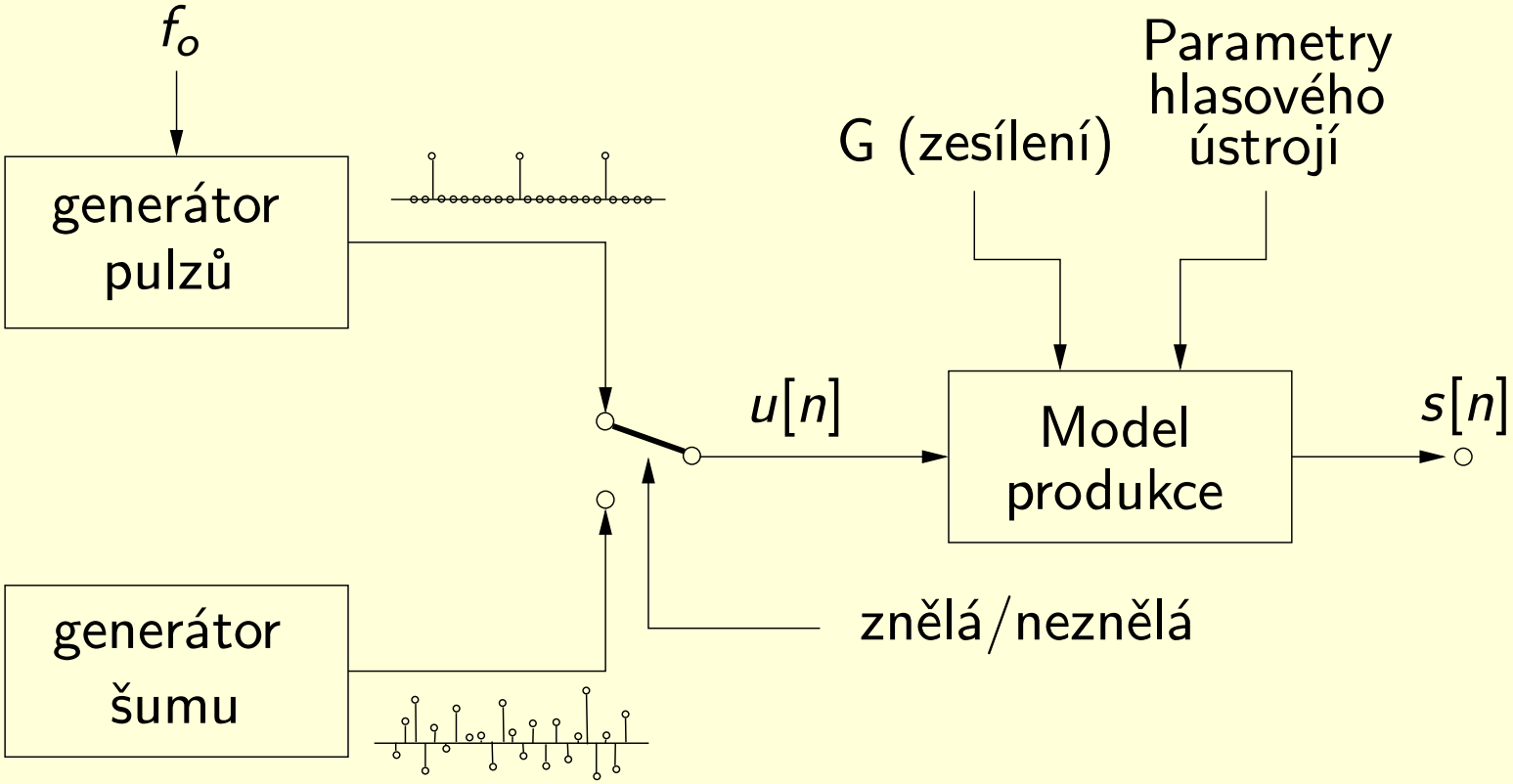


Hláska **“š”** ... neznělá, šumový charakter

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

Hláska “t” ... plozivní, okluze (závěr) + exploze, znělá i neznělá

Produkce řeči - signálový model vzniku řeči



Závislost na mluvčím :

- anatomická/fyzikální jedinečnost hlasového ústrojí

1. vokální trakt (barva hlasu)
   - souvislost s anatomií rezonátorů (dutin) hlasového ústrojí
2. generování hlasivkových pulzů (výška hlasu - intonace)
   - souvislost s vlastnosti hlasivek

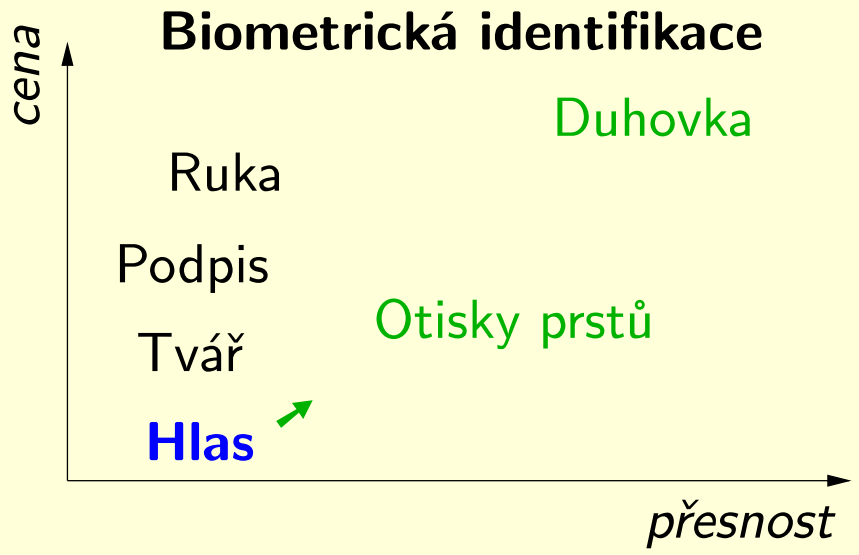
Možnosti biometrické identifikace na bázi hlasu

Originalita hlasu - výška a barva
- dané fyzikálními rozměry (anatomií) hlasového ústrojí → ☺

Originalita stylu - doba trvání hlásek, intonace, apod.
- dané dynamikou pohybu hlasového ústrojí → ☺

Obecná variabilita jednotlivých realizací každého mluvčího → ☹

Napodobení hlasu, umělá **podvržená řeč** (*deepfake speech*) → ☹



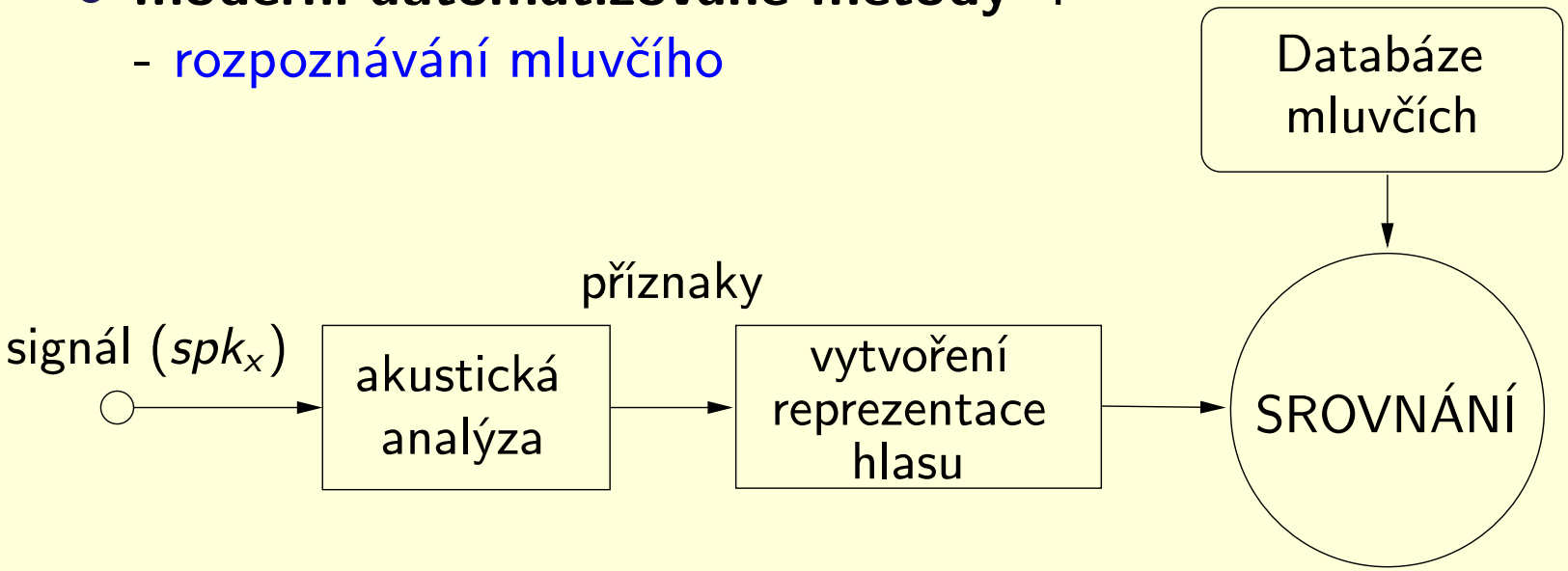
Motivace pro použití hlasové identifikace

- přirozenost komunikace, relativně jednodušší realizace
- jediná volba při dostupnosti pouze hlasového záznamu

Možnosti identifikace mluvčího

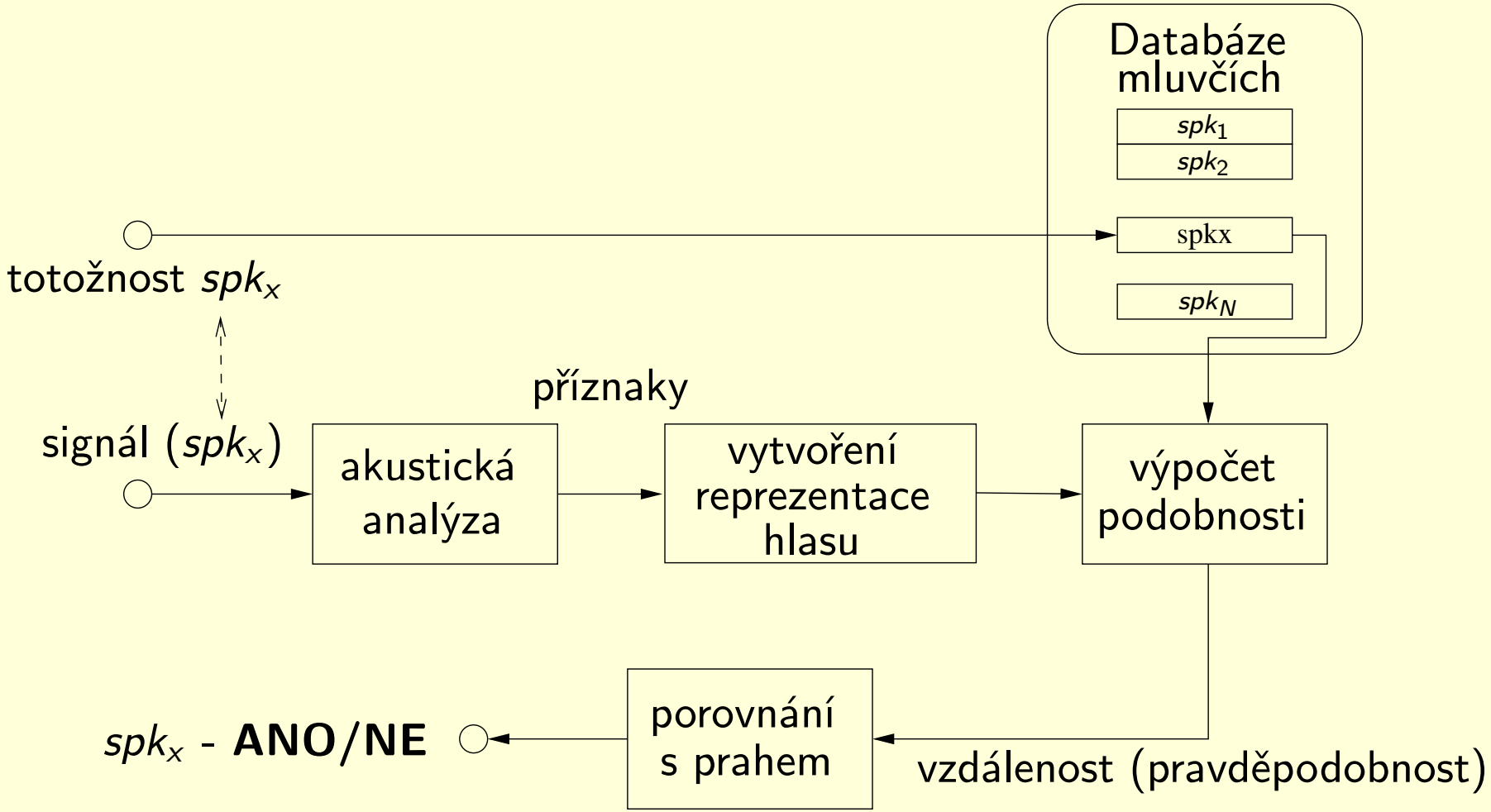
- *historické klasické přístupy ve forenzní praxi :*
  - expertní rozhodování (fonetici, lingvisté)

- **moderní automatizované metody** :
  - rozpoznávání mluvčího



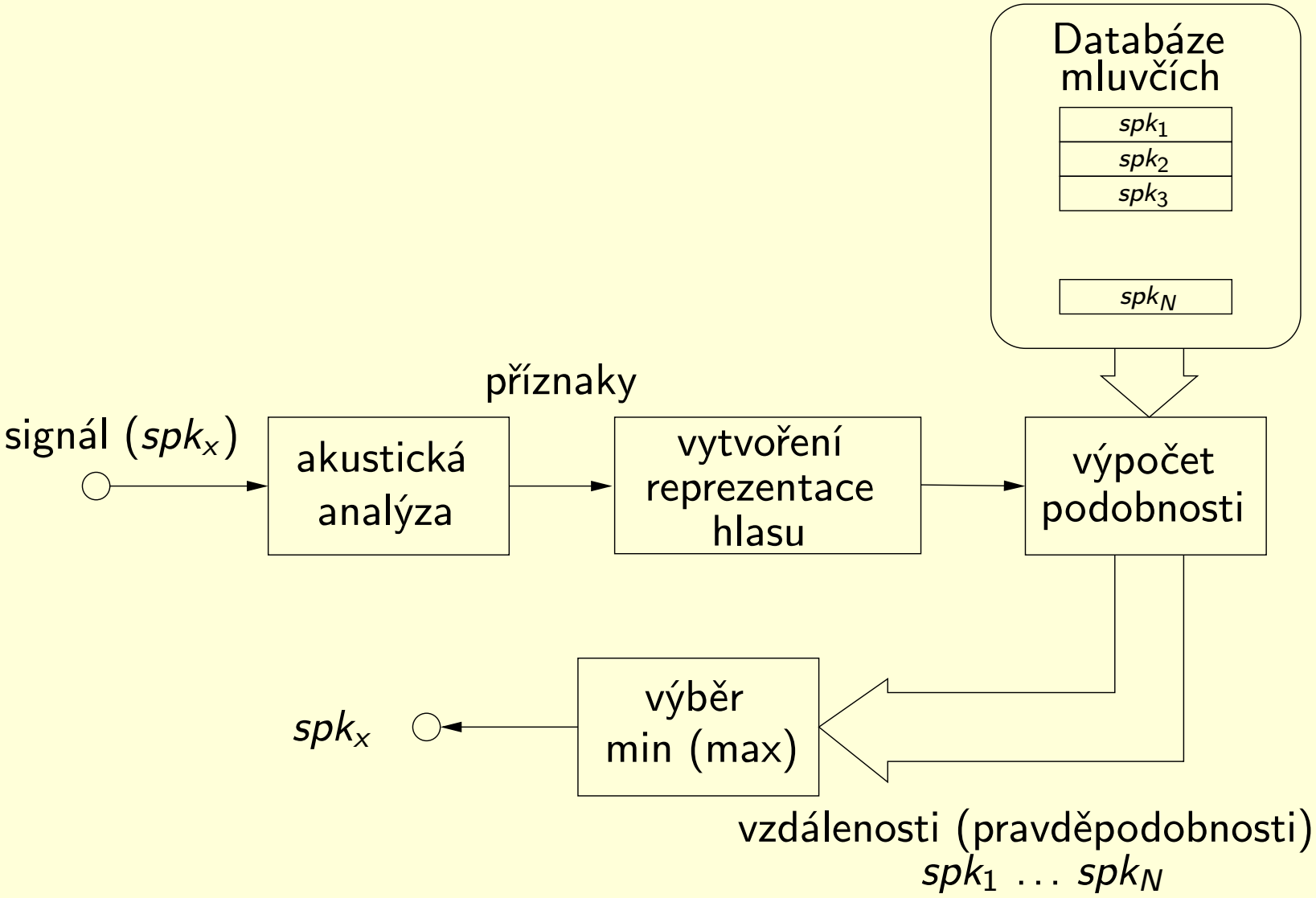
- **řečové příznaky**
  - spektrum, kepstrum, formanty, základní tón, apod.
- **reprezentace mluvčího a metody srovnání**
  - DTW, VQ, GMM, HMM, i-vektory, ANN/**DNN** (**x-vektory**)

Verifikace mluvčího



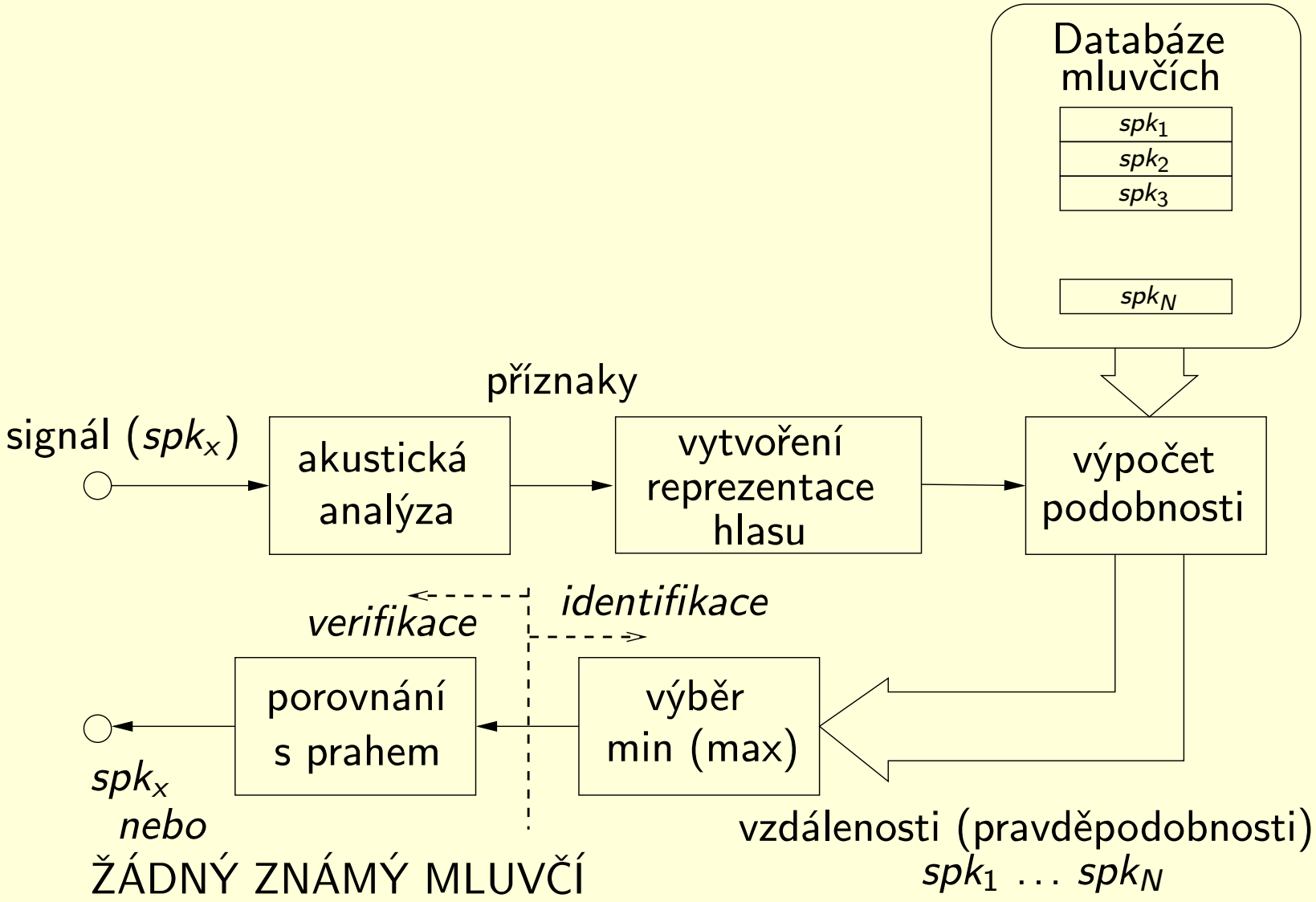
- ověření předpokládané totožnosti mluvčího
- **VÝSLEDEK** = **přijetí / odmítnutí** předpokl. totožnosti

Identifikace mluvčího (v uzavřené množině)



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

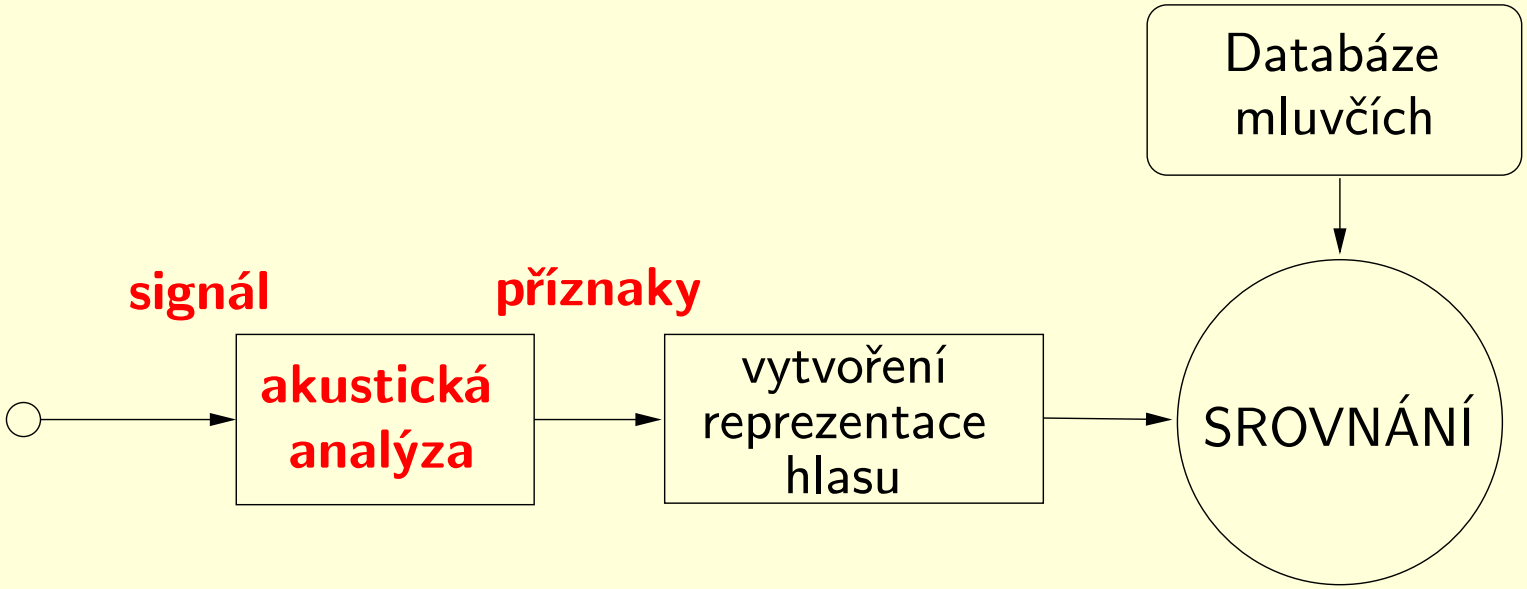
Identifikace mluvčího (v otevřené množině)



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

II. část

Řečové charakteristiky a možnosti využití pro identifikaci



Řečové příznaky pro hlasovou identifikaci

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 charakteristiky - související s vytvářením řeči
(fyziologií vokálního traktu)

Získané charakteristiky - reprezentují dynamiku pohybů hlasového traktu
(dané prostředím života, regionem, vzděláním, výchovou)

Řečové příznaky pro hlasovou identifikaci - PROBLÉMY

Vnitřní charakteristiky řečníka

+ obtížně cíleně ovlivnitelné
  - dané fyzikálními rozměry hlasového ústrojí

– ovlivnitelné zdravotním stavem, stresem
  Př. 1: nosní dutina : neměnné rozměry při artikulaci,
  avšak pro mírné nachlazení = zásadní změna
  Př. 2: "sevřené hrdlo" při stresu

Získané charakteristiky řečníka

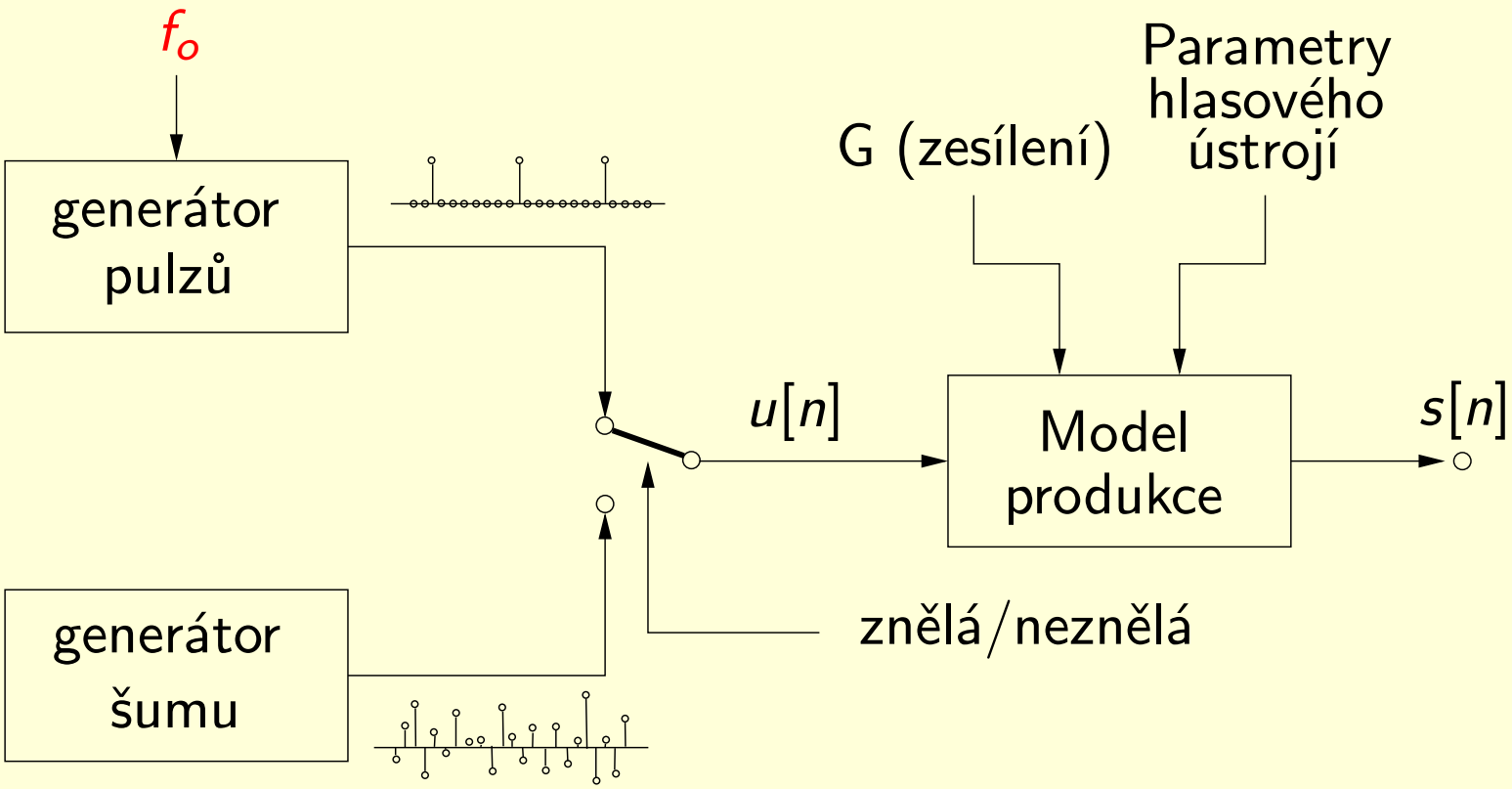
+ styl mluvy
  - časování, intonace, hrubost, živost, síla, srozumitelnost
  → jako celek komplexní charakteristika řečníka
  (používáno člověkem při přirozené identifikaci)

– nemusí být snadno modelovatelné různými modely

V obou případech

– hlas i styl mluvy jednotlivce lze snadno napodobit
(přirozeně či uměle)

Základní tón řeči



- základní frekvence $f_o = 1/T_o$
- pro znělé hlásky s harmonickou strukturou
- souvisí s kmitáním hlasivek
- hodnota f_o je ovlivněna vlastnostmi hlasivek (pružnost, hmotnost, délka)
  $\rightarrow$ hrubá charakteristika mluvčího

Odhad základního tónu řeči

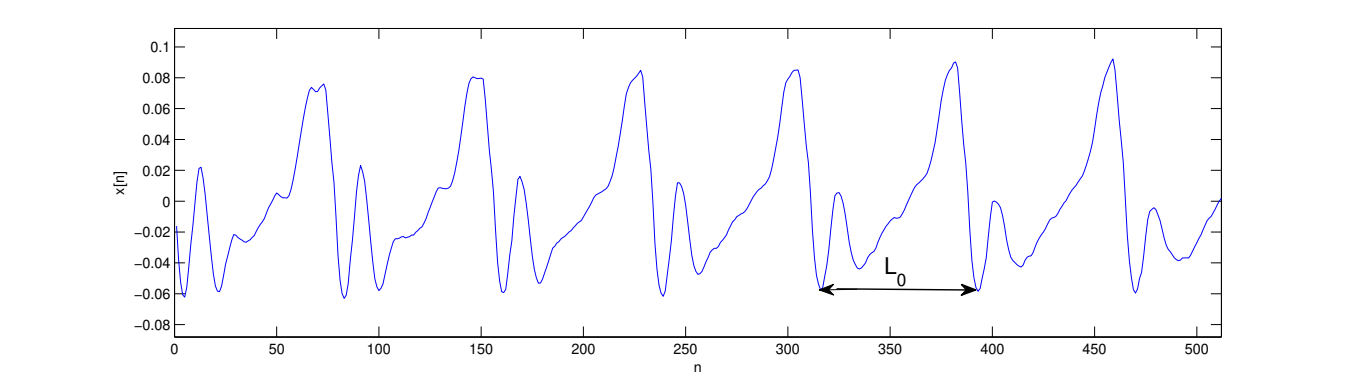
f_o ..... základní tón (frekvence) řeči ..... $f_o = \dfrac{1}{T_o}$

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

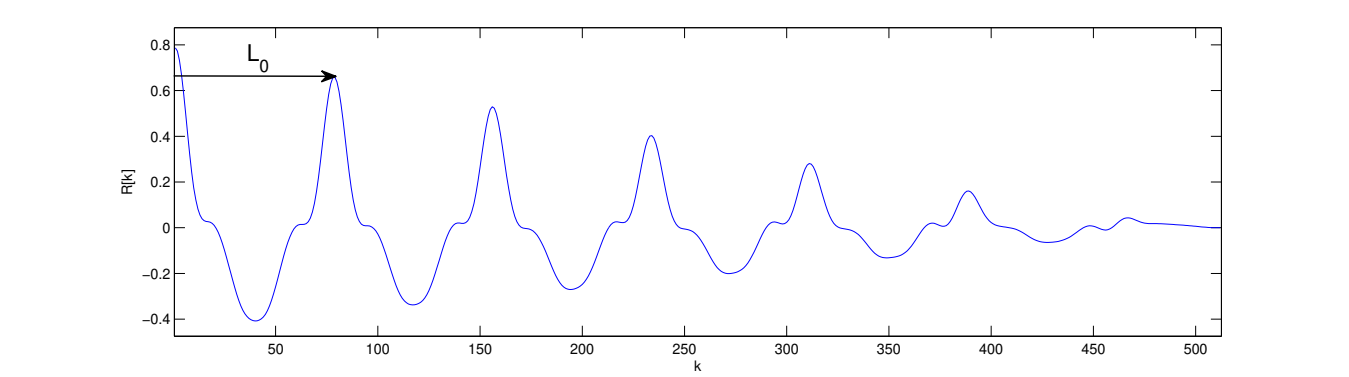
Nejčastější metoda odhadu - na bázi autokorelační funkce

(hledání postranního maxima autokorelační funkce)

segment signálu

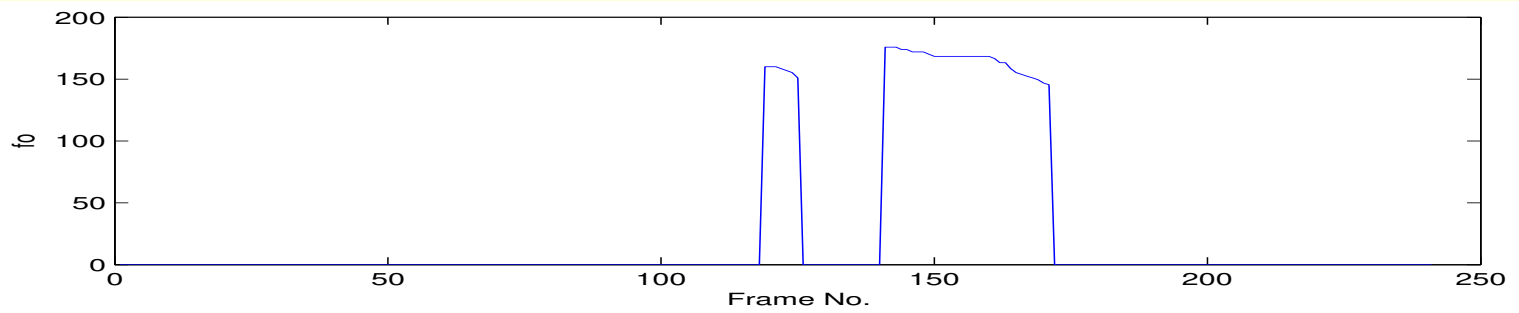


odhad autokorelační funkce

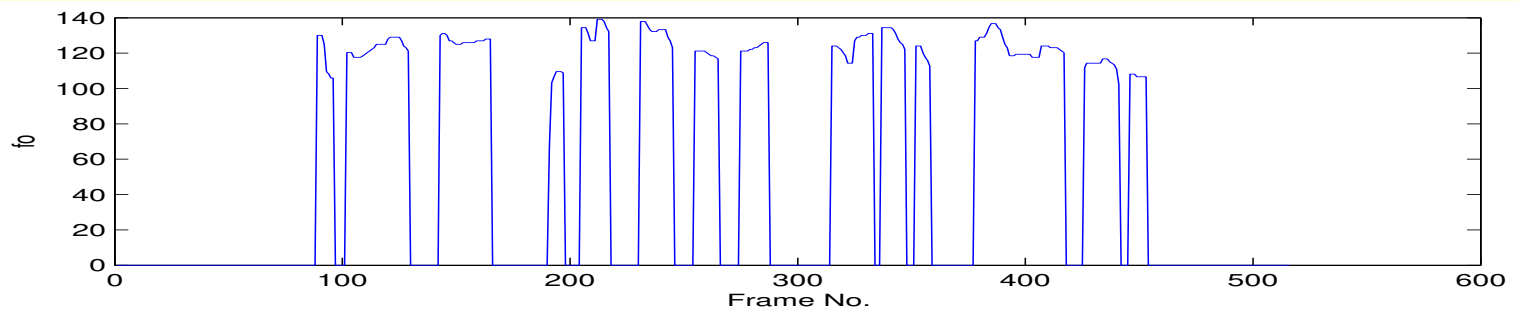


Průběh základního tónu v promluvě

Krátká promluva - slovo



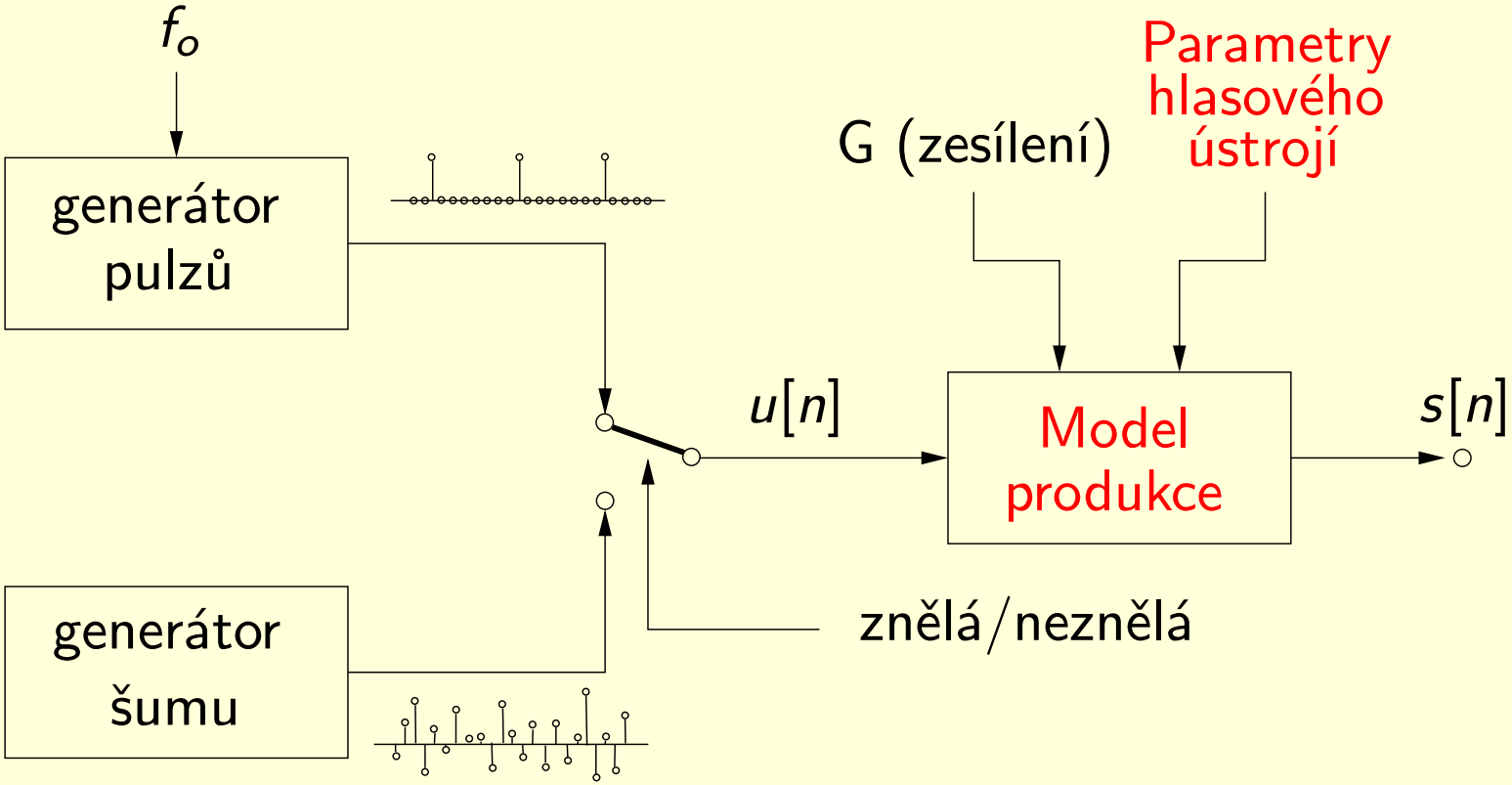
Delší promluva - věta



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

Průměrná hodnota f_o $\rightarrow$ vnitřní charakteristika (výška hlasu)

Spektrální charakteristiky řeči



- spektrální charakteristiky souvisí s vokálním traktem
- otázka vhodné reprezentace pro identifikaci

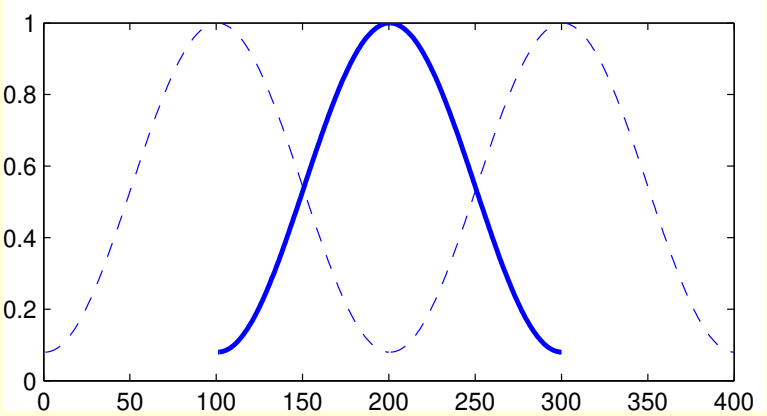
Spektrální reprezentace řeči na bázi DFT

Odhad spektra na bázi DFT:

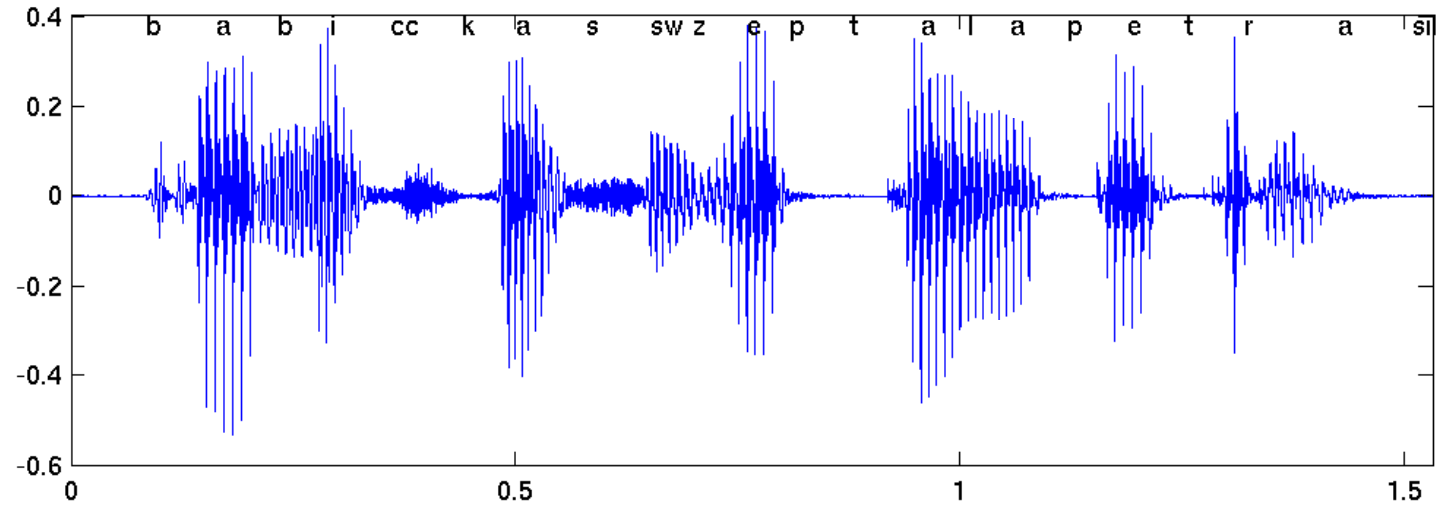
- **řeč je obecně nestacionární signál** $\Rightarrow$ nutná segmentace a sledování vývoje krátkodobého spektra (spektrogram)
- **řeč je kvazistacionární**
  (tj. stacionární v krátkém časovém intervalu - cca 10-100 ms)
  $\Rightarrow$ 20-30 ms - typická délka krátkodobého segmentu
- **DFT spektrum je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

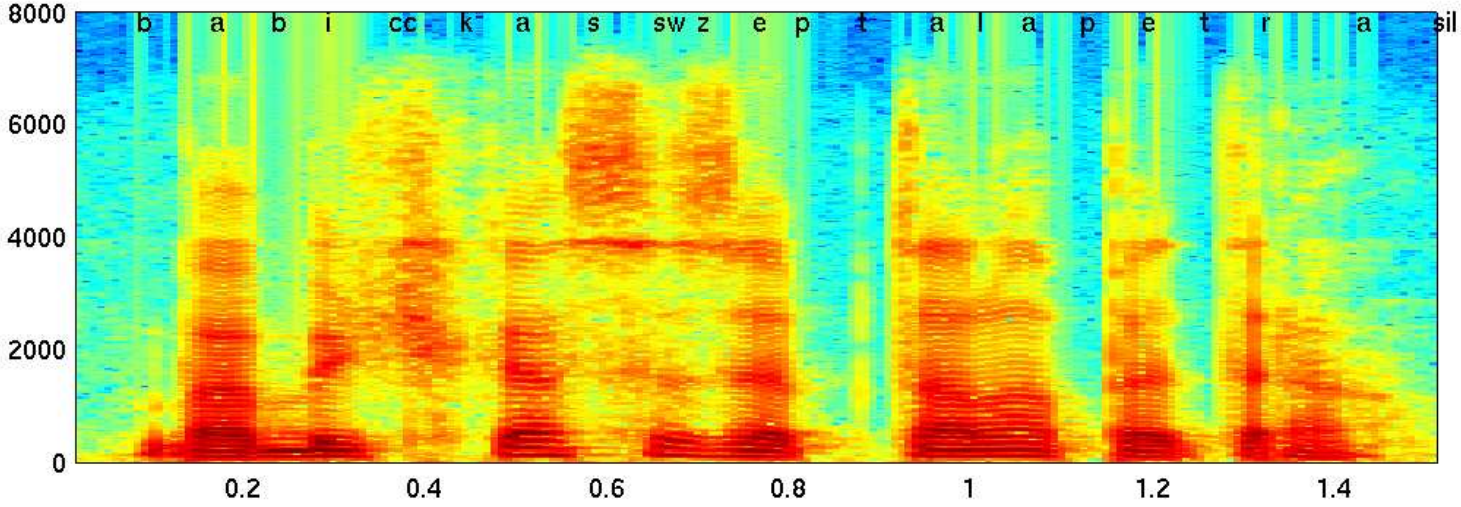
$$w[n] = 0,54 - 0,46 \cos \frac{2\pi n}{N}$$

pro $0 \leq n \leq N - 1$.



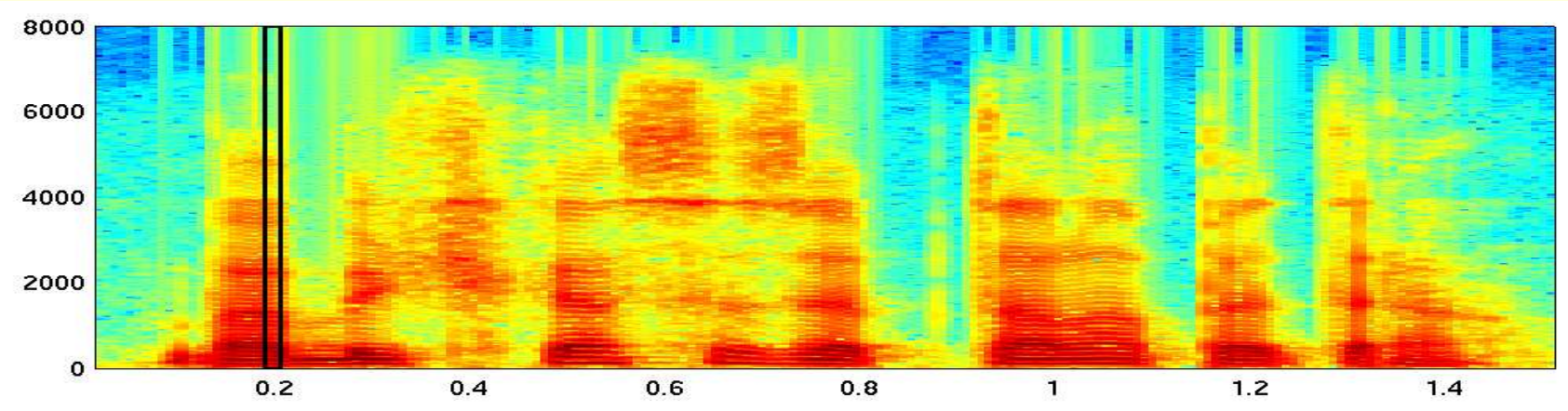
Časový průběh a spectrogram řeči





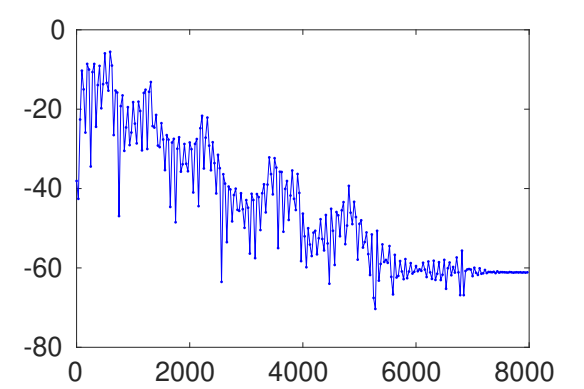
Přehled možností spektrální reprezentace promluvy

Spektrogram celé promluvy



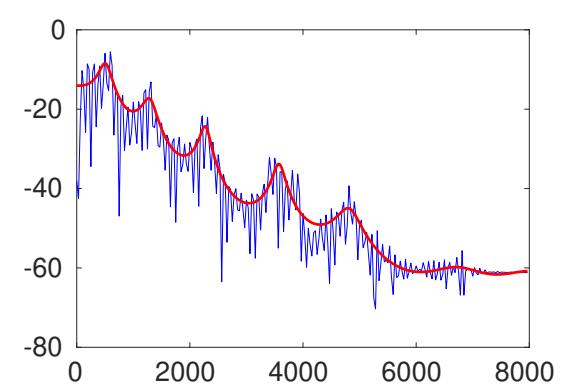
Spektrální reprezentace vybraného segmentu

DFT spektrum:



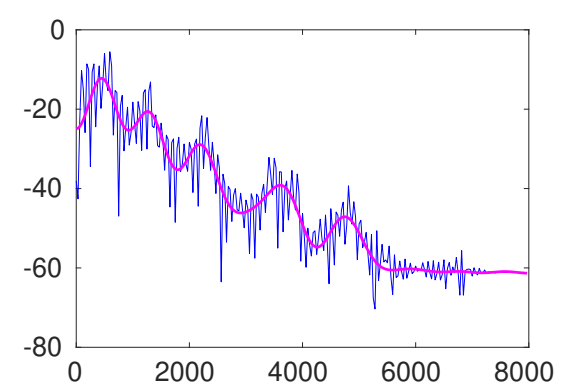
256 vzorků spektra
(amplitudové sp.)

LPC spektrum:



16 koeficientů a_k
(AR koef., formanty)

Kepstrální koeficienty:



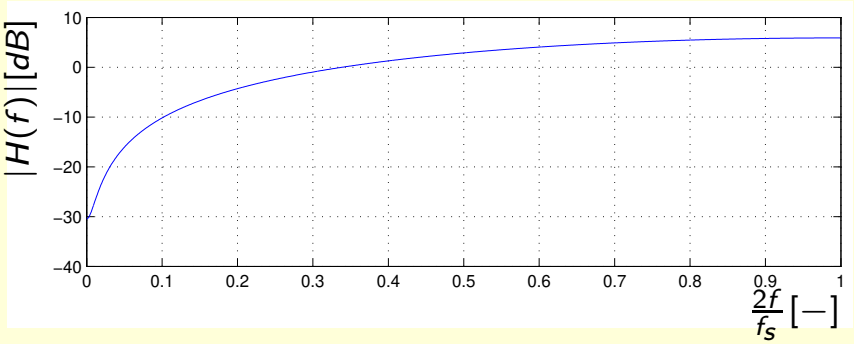
20 koeficientů c_n
(reálné kepstrum)

Preemfáze signálu - kompenzace útlumu vyšších kmitočtů

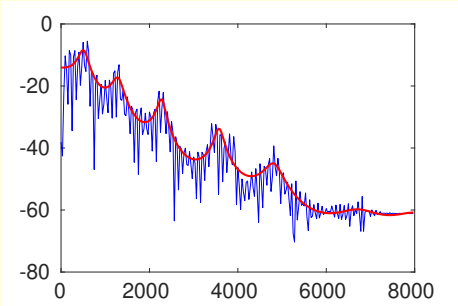
Sklon amplitudového spektra - vyšší kmitočty - nižší energie

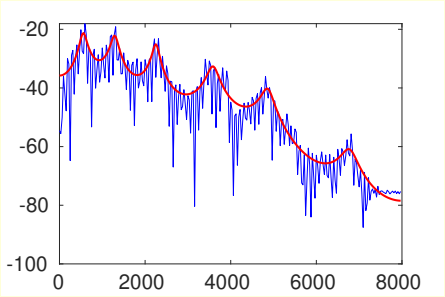
Preemfázový filtr (FIR 1.řád):

$$s'[n] = s[n] - m \cdot s[n-1]$$
$$(m = 0.97)$$

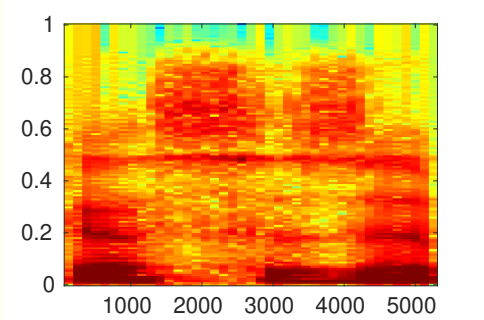


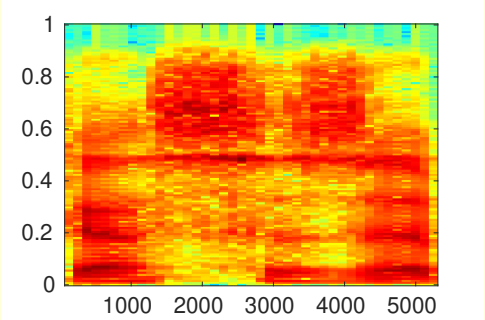
Vliv preemfáze v krátkodobém spektru (DFT a LPC)





Vliv preemfáze ve spektrogramu





Banka filtrů s melovskou nelineární frekvenční osou

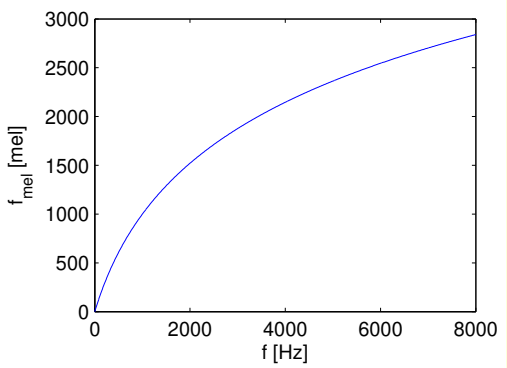
Hlavní cíl → počítá se výkon (energie) ve zvolených pásmech

Nelineární zkreslení frekvenční osy (nelineární vnímání frek.)

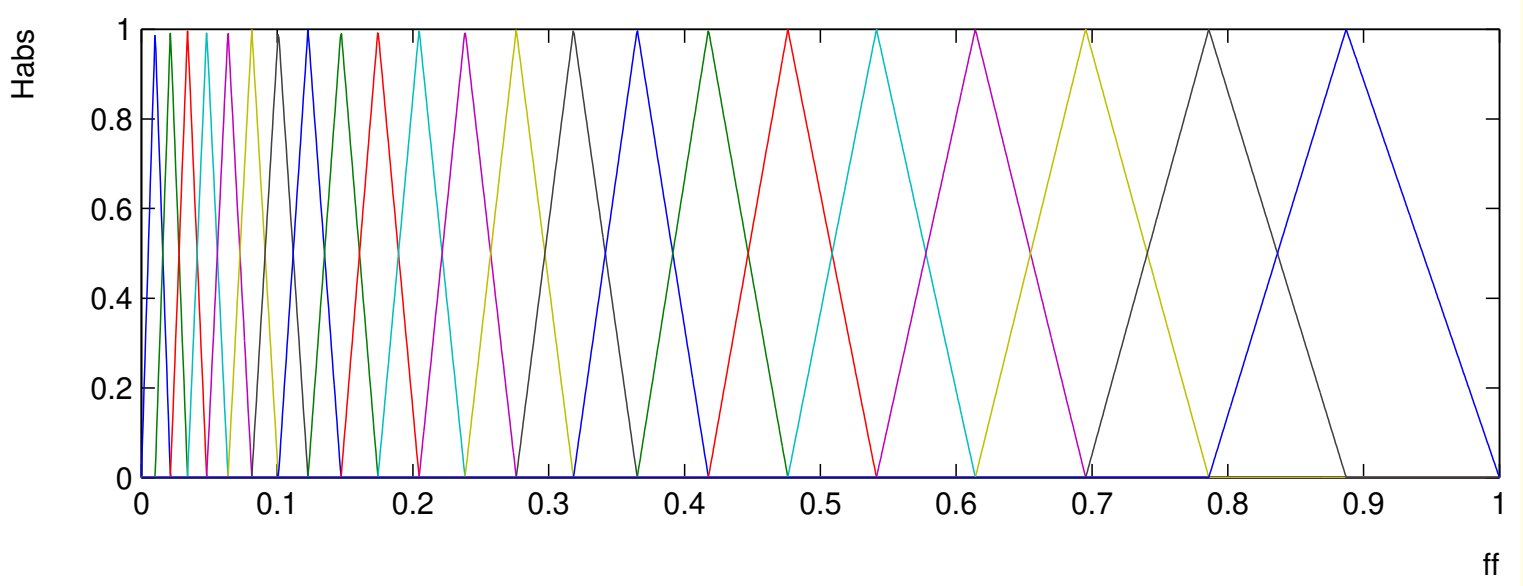
Melodická stupnice:

$$f_{mel} = \mathrm{Mel}\,(f) = 2595 \log_{10} \left(1 + \frac{f}{700}\right)$$

$$f = \mathrm{InvMel}\,(f_{mel}) = 700 \cdot \left(10^{\frac{f_{mel}}{2595}} - 1\right)$$



Trojúhelníková melovská BF (používaná pro výpočet MFCC)



Banka filtrů s melovskou nelineární frekvenční osou

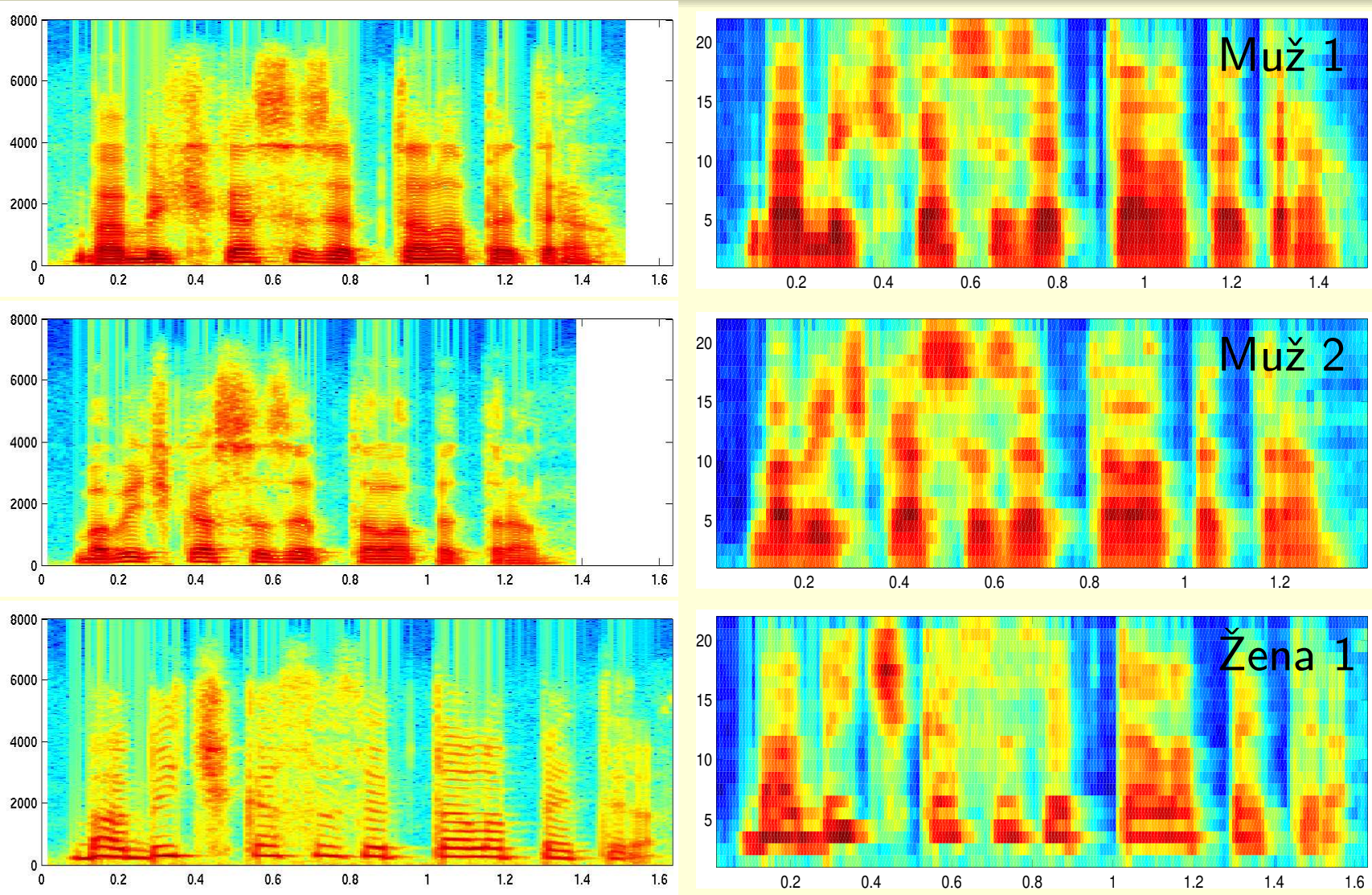
BF je realizovaná na bázi DFT

⇒ filtry jsou dány vahami DFT čar pro dané rozlišení (NDFT) a f_s
⇒ princip výpočtu je stejný pro všechny BF
⇒ pro jinou BF pouze jiné konkrétní váhy (např. PLP BF)

$$G_{mel}[j] = \sum_{k=0}^{N/2} |S[k]|^2 H_{mel,j}[k] \quad \text{pro j} = 1, \ldots, \text{M}$$

<u>M - počet pásem</u> ..... typické hodnoty 20-30 pásem
- podle f_s a počtu bodů DFT
- 22 pro $f_s = 8$ kHz a segment 25 ms
- 30 pro $f_s = 16$ kHz a segment 25 ms

Variabilita promluvy v melovském spektrogramu

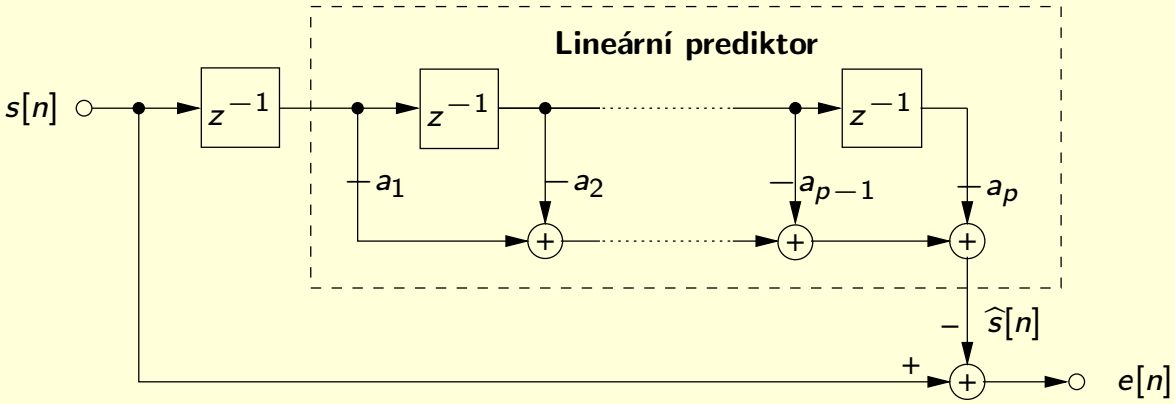


Výstup BF - pásmový spektrogram - častý příznak na vstupu DNN

- pro DNN aplikace často více pásem (40+)
- jeden vektor v jednom čase - vnitřní (statické) příznaky = charakt. hlasu
- zahrnutý časový kontext - naučené (dynamické) příznaky = charakt. stylu

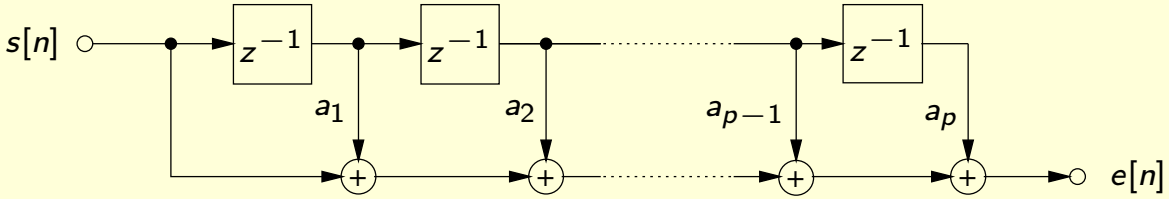
Lineární prediktivní analýza

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 $\rightarrow$ nižší úroveň chybového signálu

Kritérium - výkon chybového signálu

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

Hledání koeficientů a_k $\rightarrow$ minimalizace chyby predikce
$\rightarrow$ hledání minima kriteriální funkce 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 (Yule-Walker)
  - nejčastěji používaný přístup
  - Levinson-Durbinův algoritmus (rychlý rekurentní výpočet)
- *pro řeč často Burgův algoritmus*

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

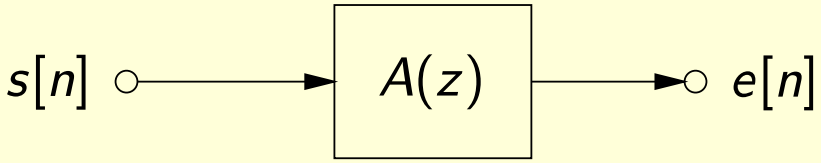
<u>VÝSLEDEK:</u>

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

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

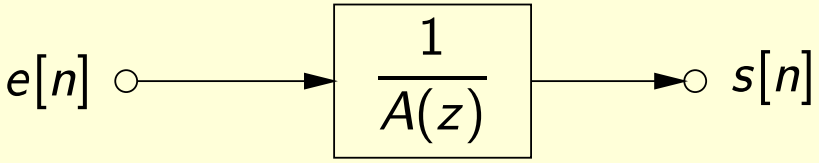
AR model signálu

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

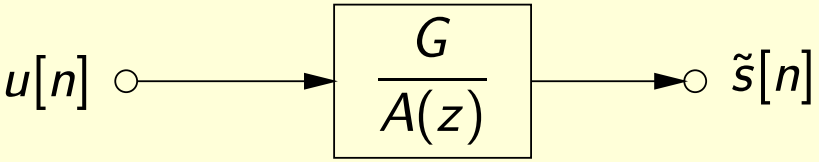


Chybový signál $e[n]$ je dekorelovaný (bílý šum, pulsy)

Syntéza se skutečným chybovým signálem (ideální případ)



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



$H(z) = \dfrac{G}{A(z)}$ .... AR model signálu

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

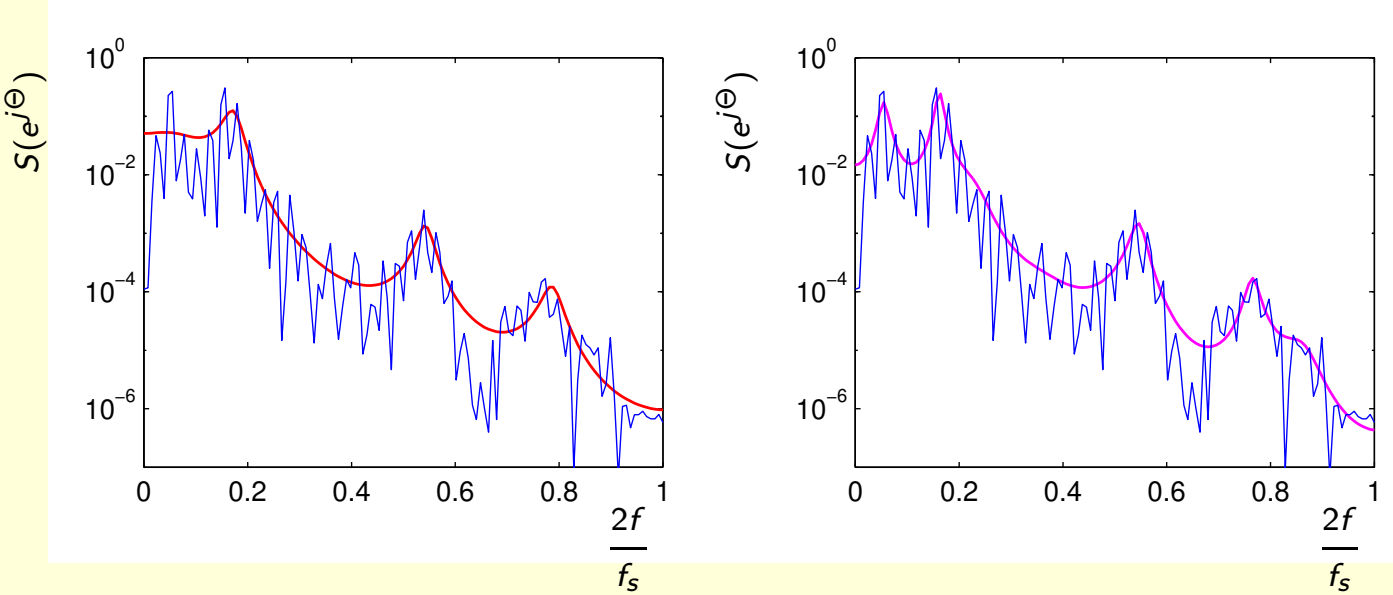
$$S_{\tilde{S}}(e^{j\Theta}) = |H(e^{j\Theta})|^2 = \frac{G^2}{|A(e^{j\Theta})|^2}$$

$$\Downarrow$$

koeficienty a_k .... komprimovaná spektrální reprezentace

Srovnání LPC a DFT spektra

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



- AR model: "all-pole" filtr, modeluje pouze špičky ve spektru (rezonátory v dutinách vokálního traktu)
- obecná špička = dvojice komplexně združených pólů
- vyšší řád AR modelu = více špiček v LPC spektru
  $\rightarrow$ typické hodnoty: $p = 10$ pro $f_s = 8$ kHz, $p = 16$ pro $f_s = 16$ kHz
- pozice špiček v LPC spektru = formantové kmitočty
  $\rightarrow$ možná reprezentace mluvčího

Formanty (formantové frekvence)

- centrální kmitočty rezonátorů vokálního traktu
- významné špičky ve VYHLAZENÉM (LPC) spektru
- významné formanty F1 - F4 v pásmu do 4 kHz

$$\downarrow$$

Souvislost s fyziologií vokálního traktu = vhodný vnitřní příznak
(formantové frekvence jsou nepřímo úměrné délce vok. traktu)

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

Odhad na bázi LPC:

- z **pólů** (p_i) **přenosové funkce** $H(z) = \dfrac{G}{A(z)}$
- F_i - formantová frekvence (centrální kmitočet rezonátoru)

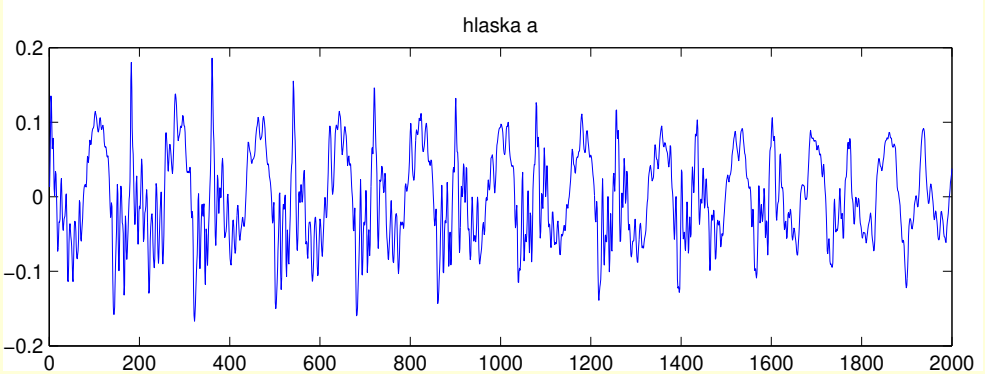
  $$F_i = f_s \cdot \arg\, p_i / 2\pi$$

- B_i - šířka pásma formantu

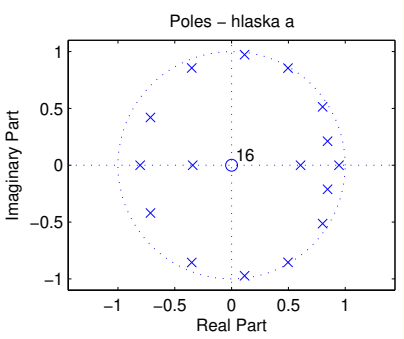
  $$B_i = -f_s \cdot \ln |p_i| / 2\pi$$

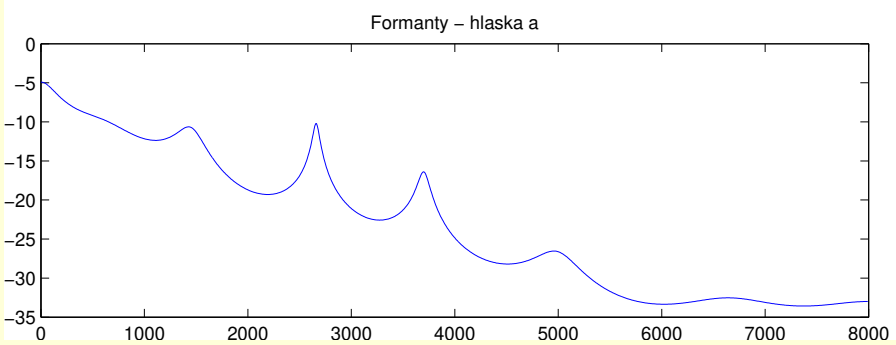
Odhad formantů na bázi LPC - příklad

Časový průběh signálu

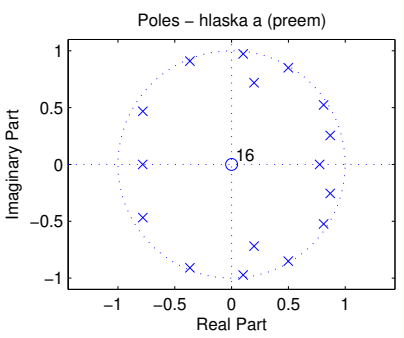


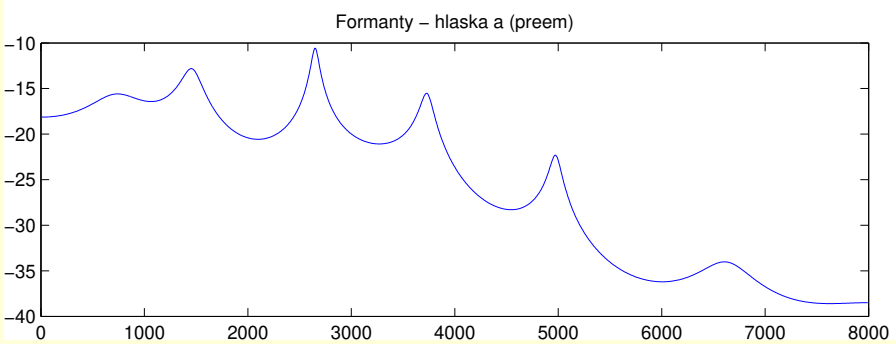
póly & LPC spektrum s formanty





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. .....

+ větší soubor příznaků, víceúrovňové rozhodování, větší přesnost

– **textově závislé příznaky**

Kepstrum - definice a základní vlastnosti, názvosloví

Základní definice pomocí Z-tranformace

$$\hat{c}[n] = \mathcal{Z}^{-1}\{\ln \mathcal{Z}\{x[n]\}\}$$

Výpočet pomocí DFT (numerický odhad) :

$$c_r[n] = \mathcal{IDFT}\{\ln |\mathcal{DFT}\{x[n]\}|\} \quad \ldots \text{ reálné kepstrum}$$

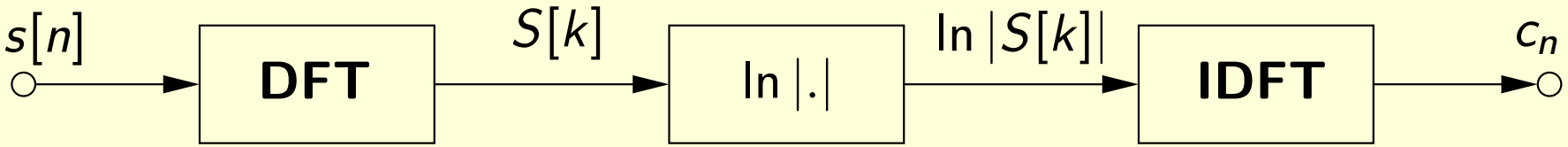
Obecné vlastnosti:

- $\hat{c}[n]$ .... nekonečně dlouhé, rychle ubývá k nule
- $c_r[n]$ .... konečně dlouhé, symetrické, inf. o ampl. spektru
- ve všech případech vždy reálné hodnoty pro reálný signál

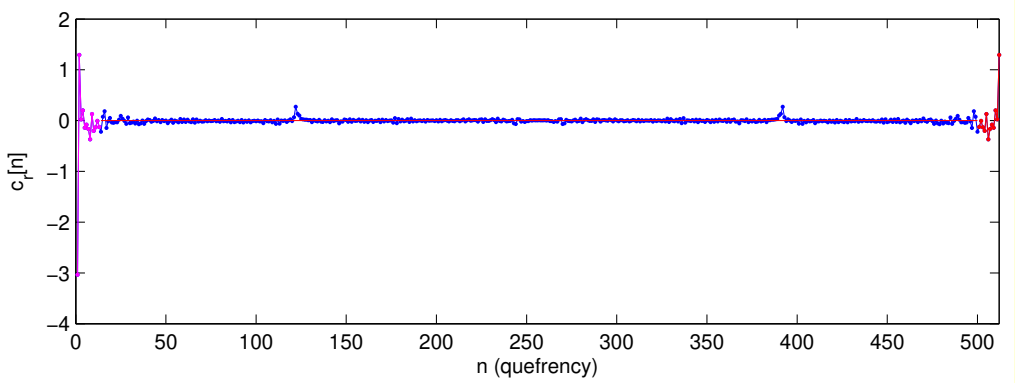
Názvosloví = slovní přesmyčky:

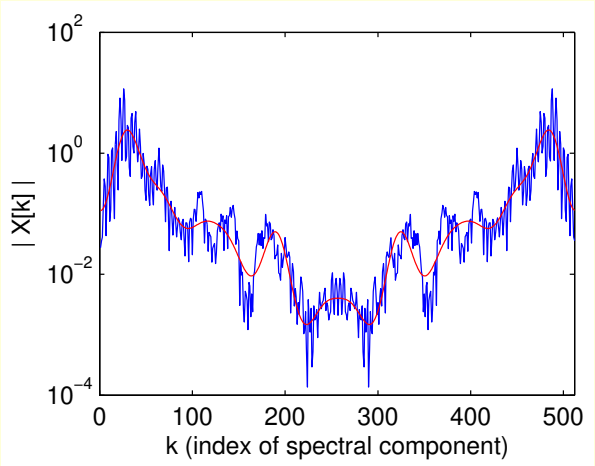
- **spektrum** vs. **kepstrum**
- kvefrence (frekvence) - základní proměnná kepstra [čas]
- liftr (filtr)
- liftrace (filtrace) = modifikace kepstra

Vlastnosti reálného DFT kepstra

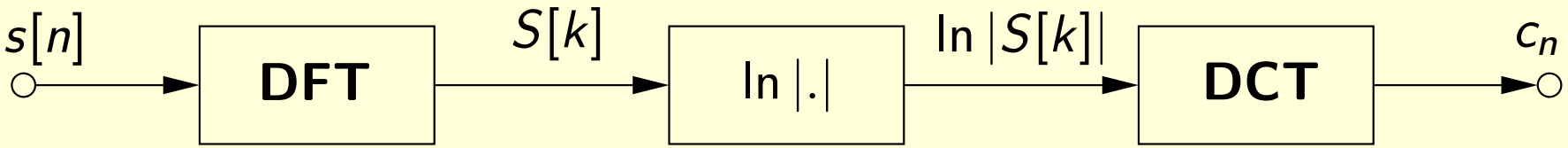


- DFT kepstrum - numerický výpočet (period. a symetr.)
- První část - informace o tvaru amplitudového spektra
  - spektrum neperiodické složky signálu, spektrální obálka, vyhlazené spektrum - $\overline{|X[k]|} = e^{DFT\{c_n \cdot w_n\}}$



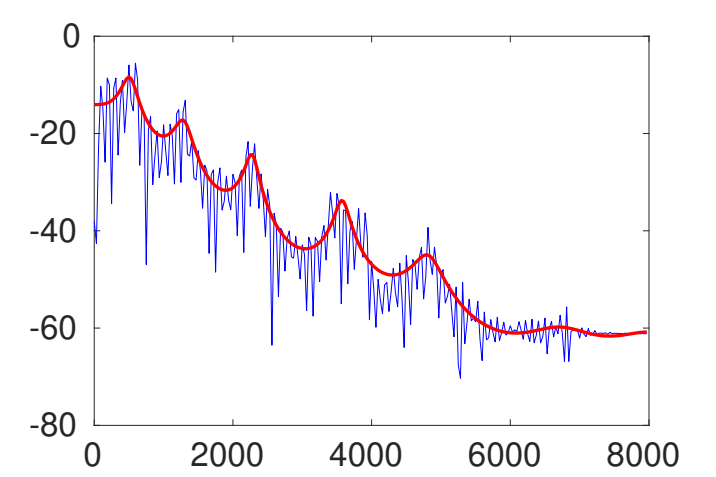


DCT místo IDFT při výpočtu reálného kepstra reálného signálu

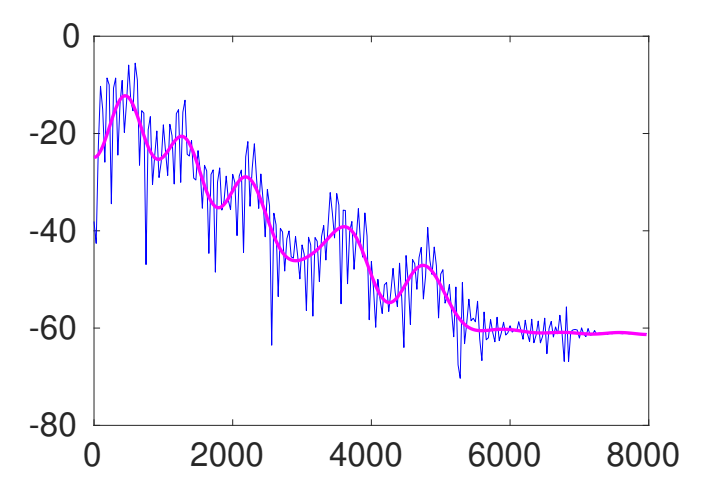


Kepstrální analýza pro zpracování řeči

LPC spektrum:



Vyhlazený odhad z reálného kepstra:



- První koeficienty nesou **komprimovanou informaci** o tvaru amplitudového spektra (12-20 kepstrálních koeficientů)
- **kepstra podobných segmentů tvoří shluky**

  $\Downarrow$

  obecně vhodné příznaky pro rozpoznávání

- **kepstrum** (amplit.spektrum) reprezentuje obecně informaci o **vokálním traktu**

  $\Longrightarrow$ **použití i pro identifikaci mluvčího**

  (textově nezávislá identifikace)

Mel-kepstrum - melodická frekvenční stupnice

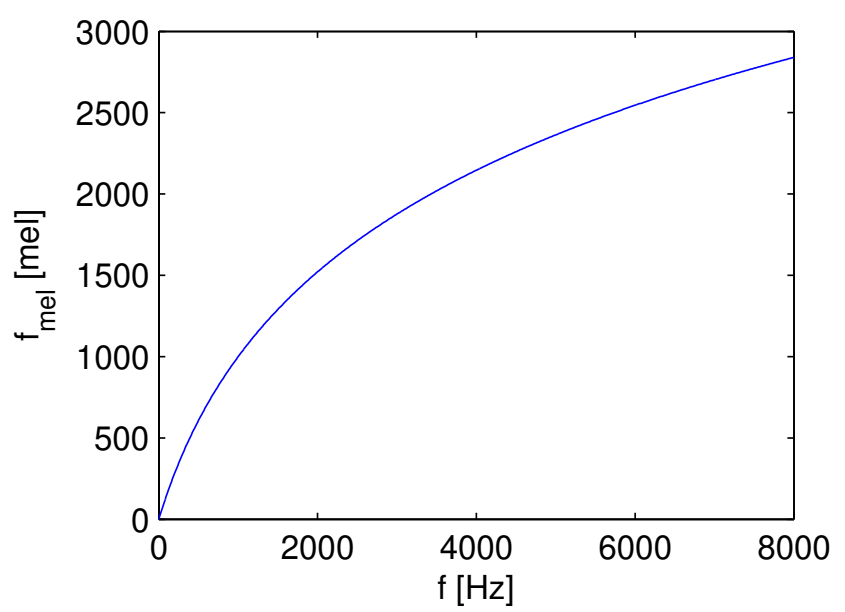
Vhodnější výpočet kepstra :

$\approx$ **modelování nelinearity vnímání frekvence lidským sluchem**

Nelineární zkreslení frekvenční osy - *melodická stupnice*

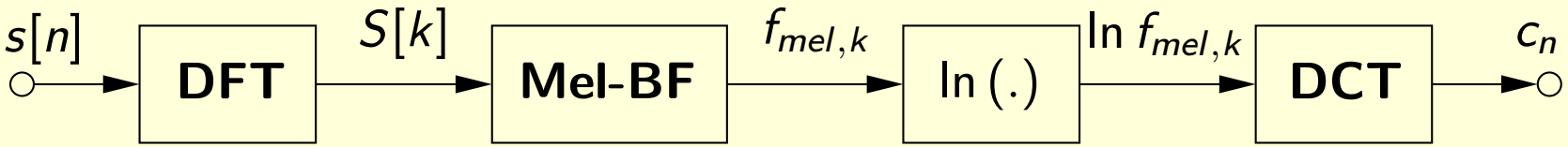
$$f_{mel} = \mathrm{Mel}\,(f) = 2595 \log_{10} \left(1 + \frac{f}{700}\right)$$

$$f = \mathrm{InvMel}\,(f_{mel}) = 700 \cdot (10^{\frac{f_{mel}}{2595}} - 1)$$

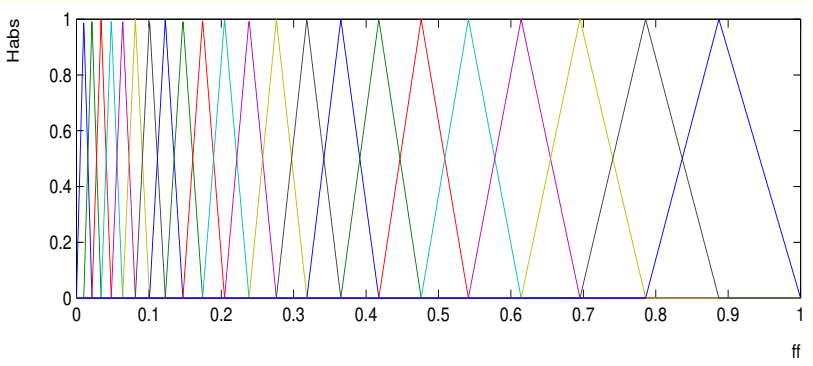


MFCC - Melovské kepstrální koeficienty

Blokové schéma výpočtu mel-kepstrálních koeficientů:



Výpočet energie v jednom pásmu



$$g_j = \ln \sum_{k=0}^{N/2} |S[k]|^2 H_{mel,j}[k] \,.$$

Výpočet kepstra pomocí DCT

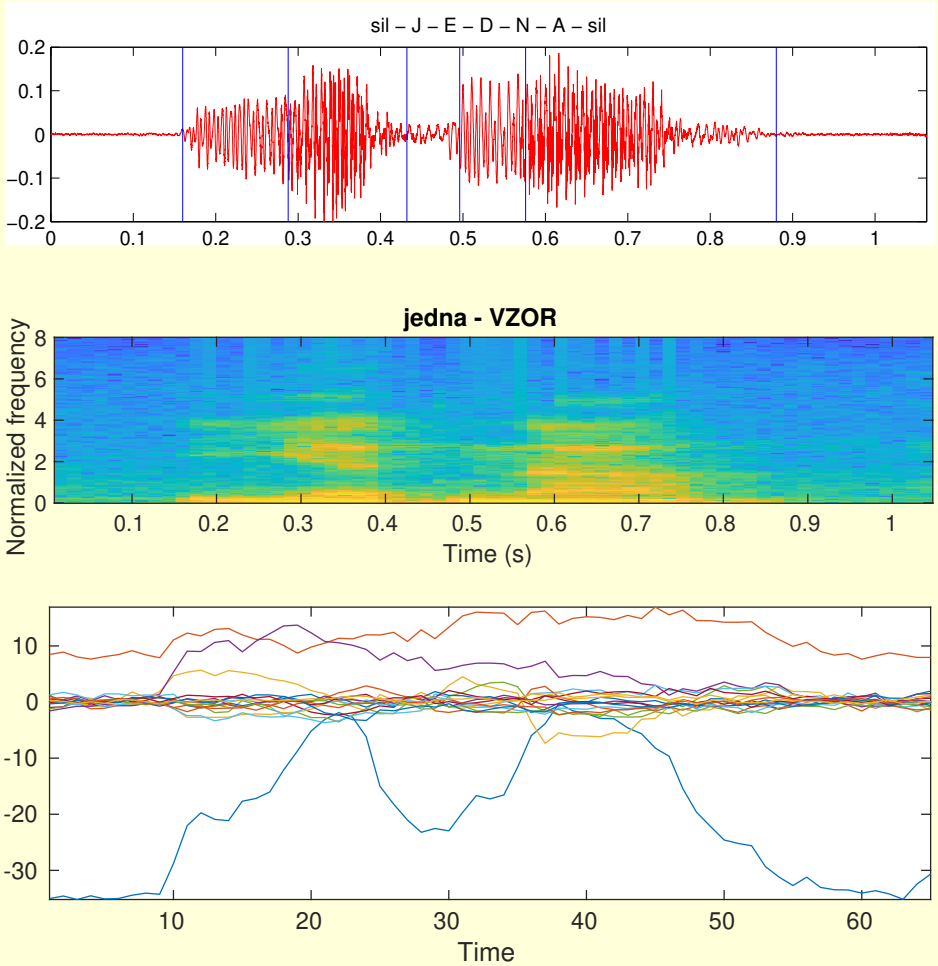
$$c_i = \sqrt{\frac{2}{P}} \sum_{j=1}^{P} g_j \cos\left(\frac{\pi i}{P}(j-0.5)\right)$$

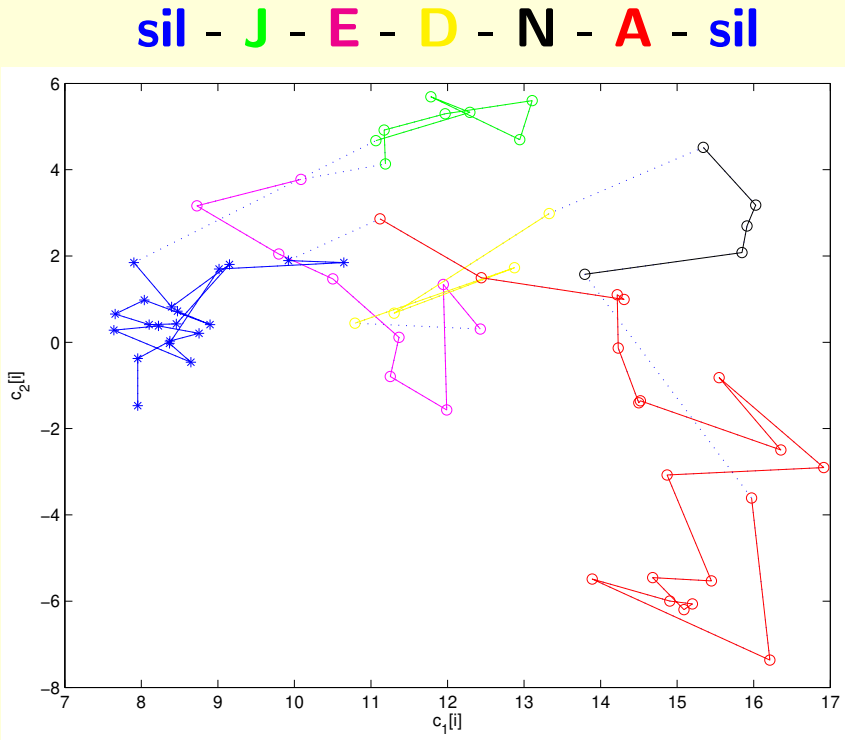
MFCC - nejrozšířenější příznaky používané pro

- *rozpoznánání řeči (ASR - Automated Speech Recognition)*
- rozpoznávání řečníka (SRE - *Speaker Recognition*) na bázi GMM (dekorelované příznaky)
- používané obvykle **textově nezávislou** hlasovou identifikaci

MFCC jednotlivých hlásek ve vybraném slově

Shluky kepster pro podobné zvuky (jednotlivé hlásky)



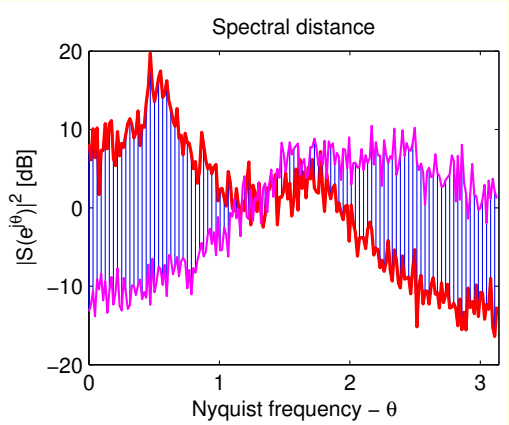


Rozlišení mluvčích v kepstrální oblasti později ....

Spektrální a kepstrální vzdálenost

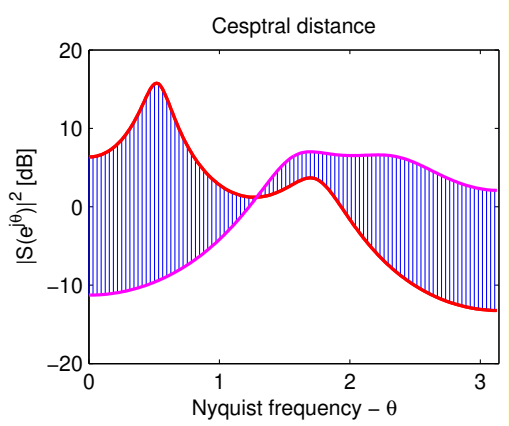
Spektrální vzdálenost (L_2-norma)

$$L_2 = \int_{-\pi}^{\pi} \ln \frac{|S_1(e^{j\theta})|^2}{|S_2(e^{j\theta})|^2} d\theta$$



Kepstrální vzdálenost

$$\mathrm{CD} = \sqrt{(c_s[0] - c_x[0])^2 + 2\sum_{k=1}^{L}(c_s[k] - c_x[k])^2}$$



- kvantifikace plochy ohraničné dvěma spektry (křivkami)
- CD aproximuje spektrální vzdálenost na bázi L_2-normy (jen první kepstrální koeficienty = vyhlazená spektra)

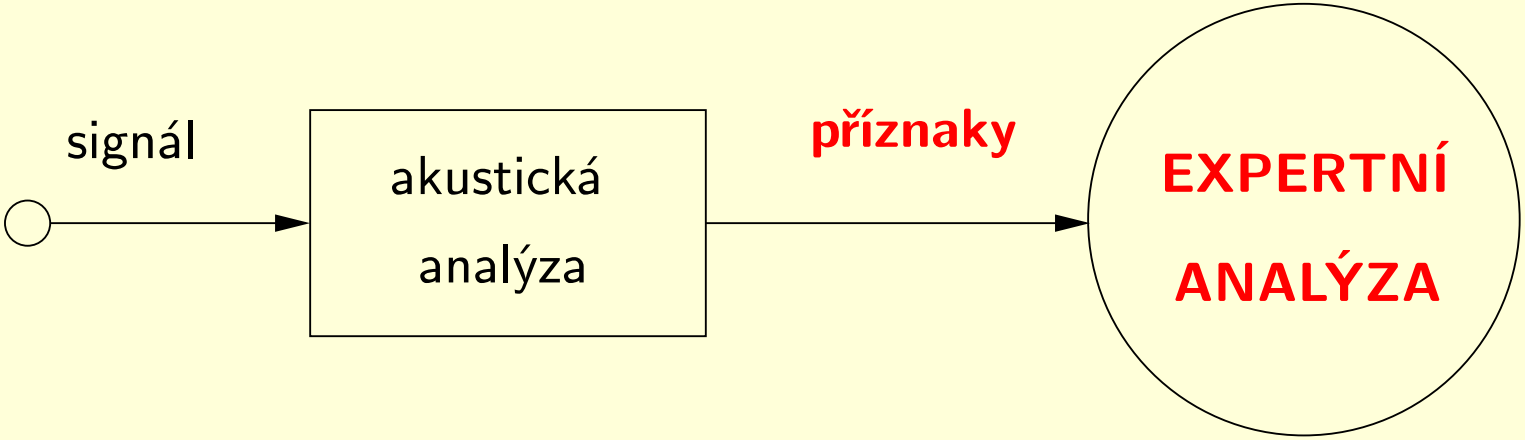
Varianty kepstrální vzdálenosti

- Euklidovská vzdálenost: $$\mathrm{CD} = \sqrt{\sum_{k=0}^{L}(c_s[k] - c_x[k])^2}$$
- Eukl. vzdálenost bez $c[0]$: $$\mathrm{CD} = \sqrt{\sum_{k=1}^{L}(c_s[k] - c_x[k])^2}$$
- kvadrát Euklidovské vzdálenosti bez $c[0]$: $$\mathrm{CD} = \sum_{k=1}^{L}(c_s[k] - c_x[k])^2$$
- vyřazení $c[0]$ $\rightarrow$ odstranění závislosti na intenzitě hlasu
  $\Longrightarrow$ **vhodné pro SRE**

- Vždy kvantifikace rozdílů ve spektru
- Varianty - různá citlivost a různé měřítko

III. část

Expertní metody v SRE



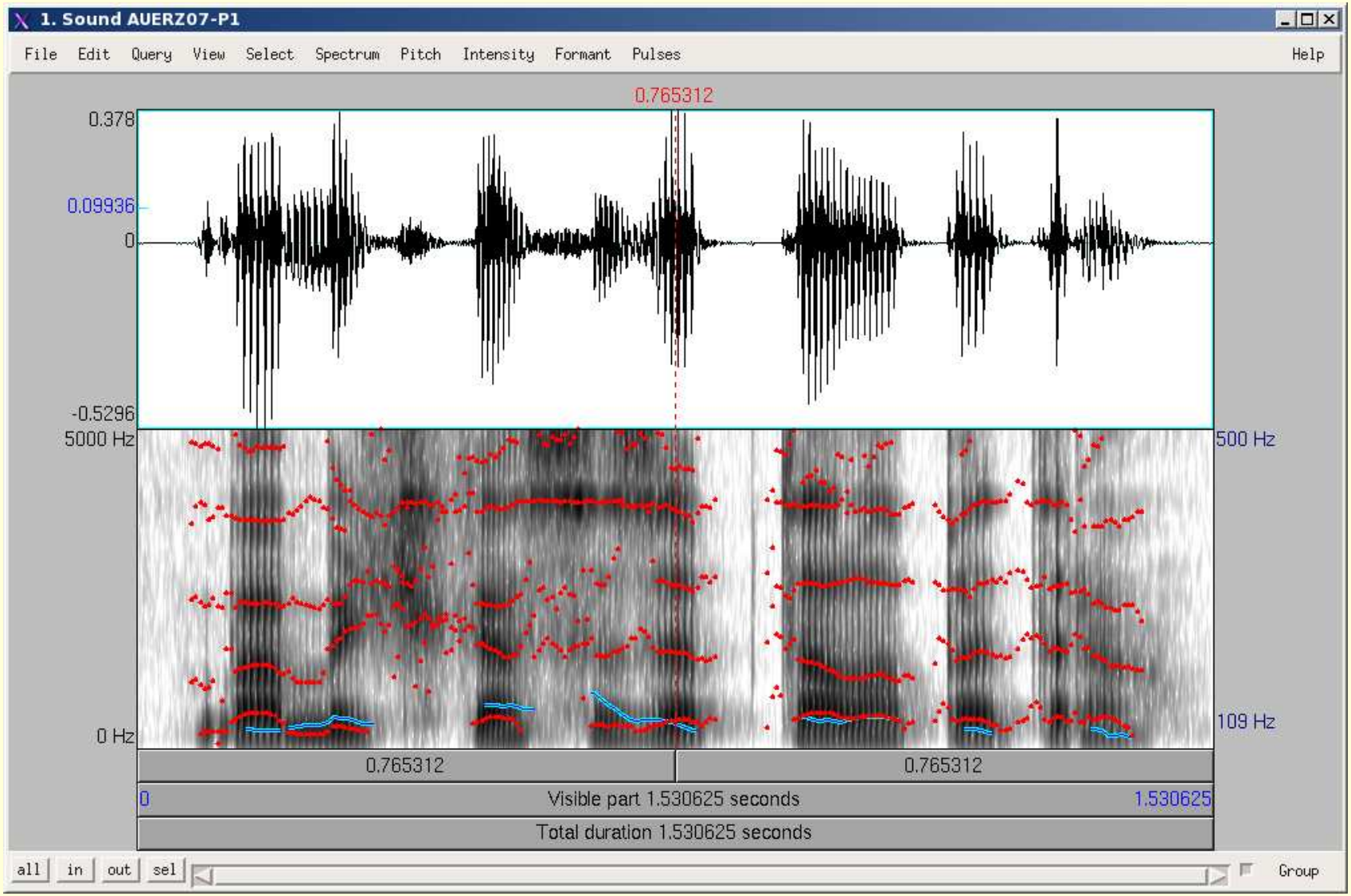
Expertní metody při rozpoznávání mluvčího

- **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 (spektrogramů, průběhu f_o, trajektorií formantů, apod.)
  - řešeno opět na expertní bázi
  - detaily realizace vybraných hlásek

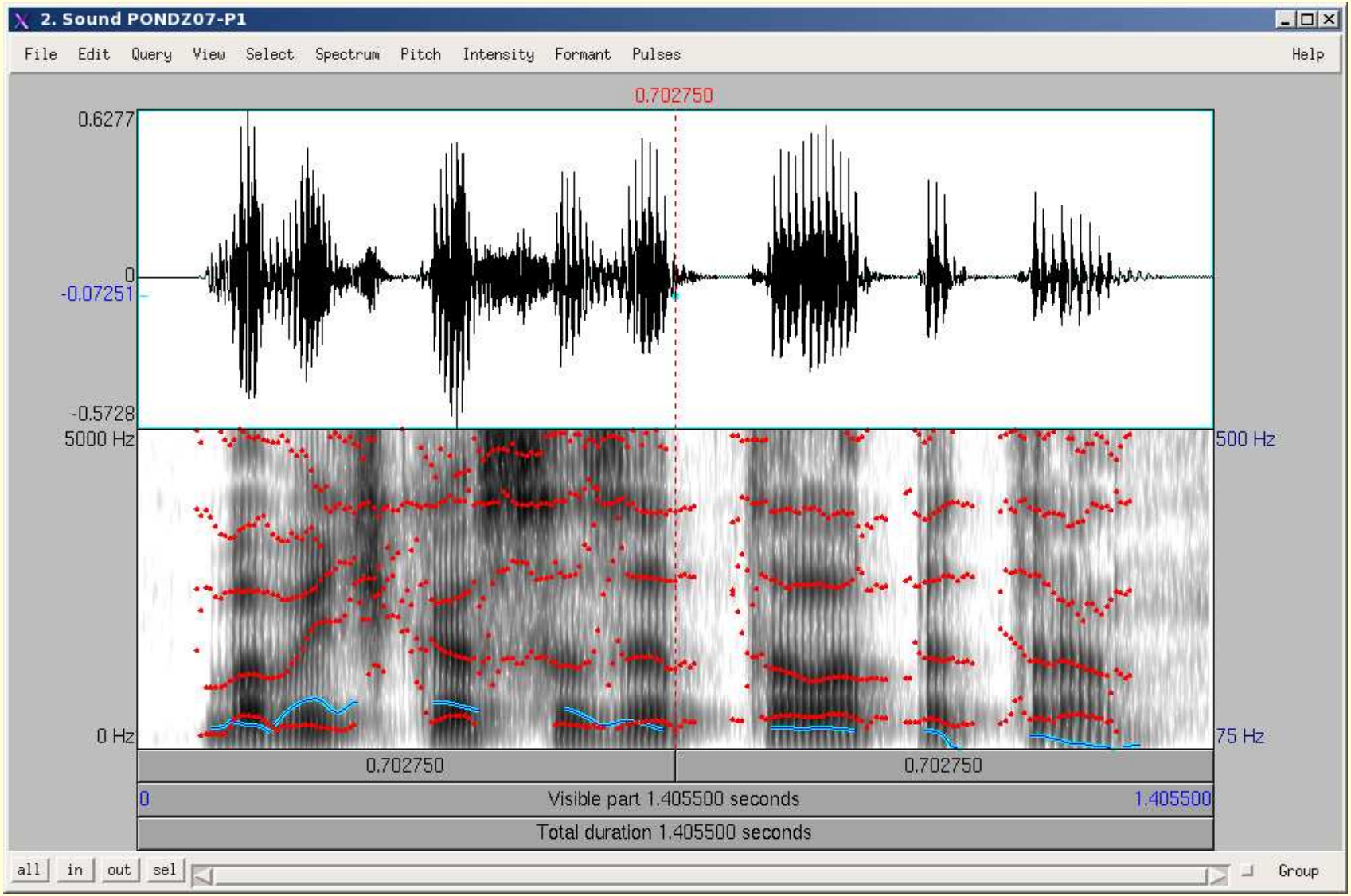
Formanty & základní tón - odhad Praat

Muž 1



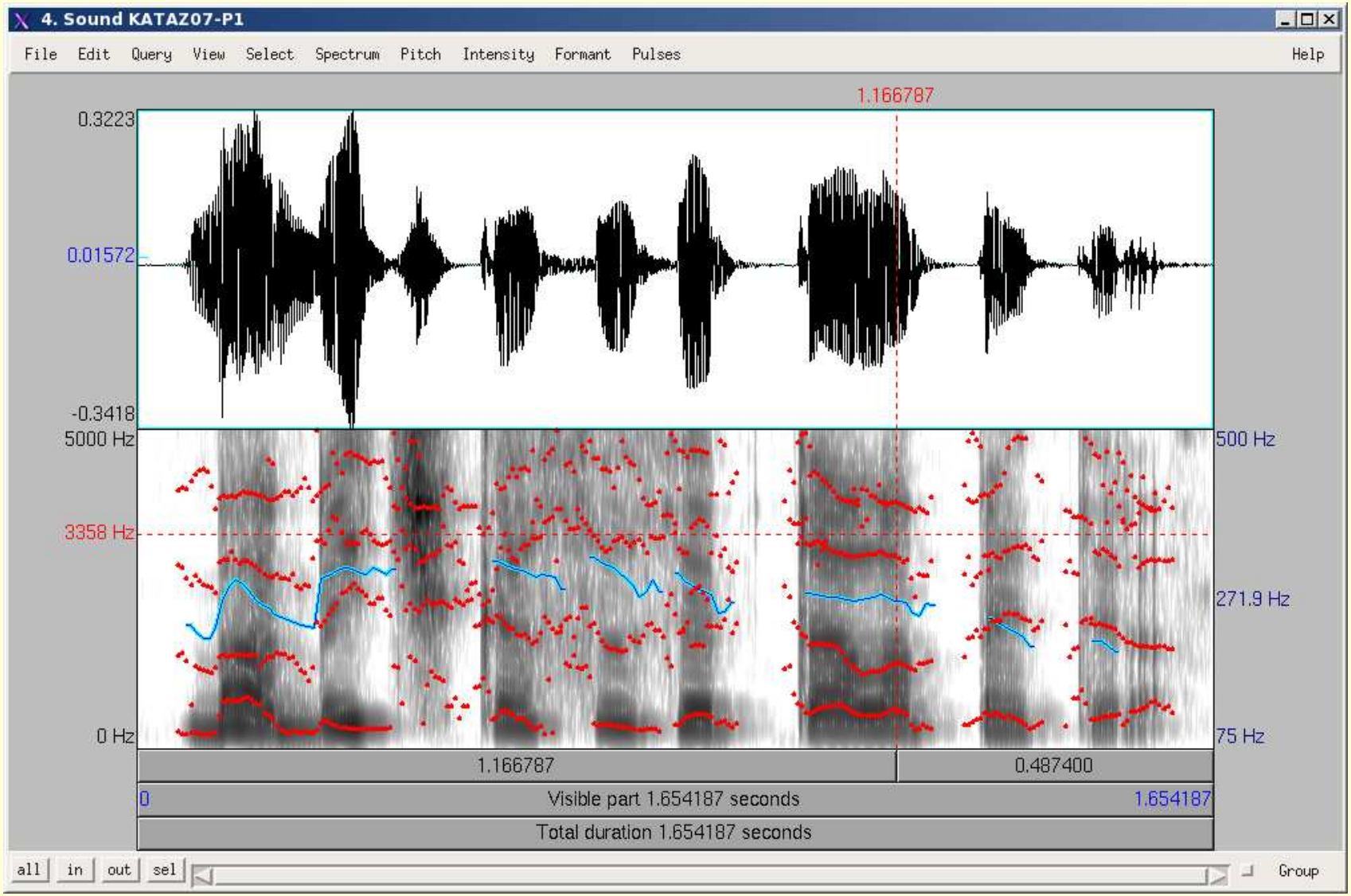
Formanty & základní tón - odhad Praat

Muž 2



Formanty & základní tón - odhad Praat

Žena 1



Děkuji vám za pozornost !