

A6M33BIO - Biometrie

Biometrické metody založené na rozpoznávání hlasu I

Doc. Ing. Petr Pollák, CSc.

17. prosince 2024 - 16:57

- **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, kepsrální 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**

SRE - Speaker REcognition

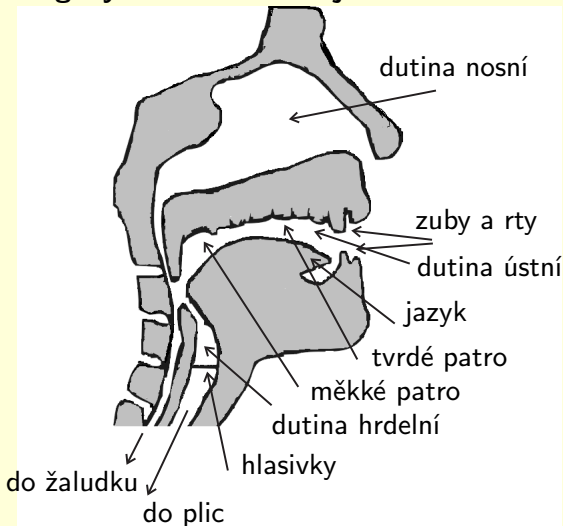
1) přesná identifikace/verifikace totožnosti mluvčího

- 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

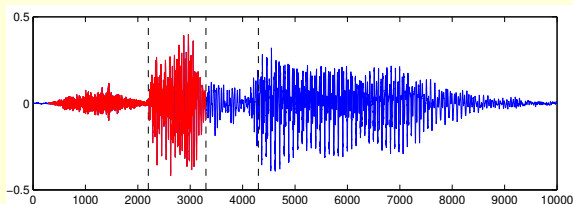
Artikulační orgány hlasového ústrojí člověka



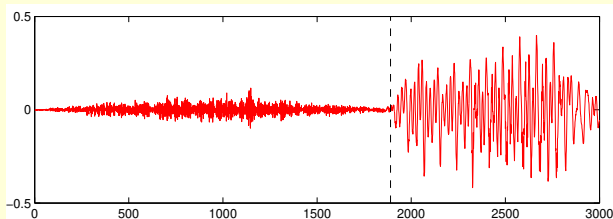
Produkce řeči : frekvenční modifikace širokopásmového buzení proudem vzduchu průchodem dutinami (rezonátory) hlasového ústrojí

Řečové hlásky v časové oblasti

Slovo “šedý”



Slabika “še”

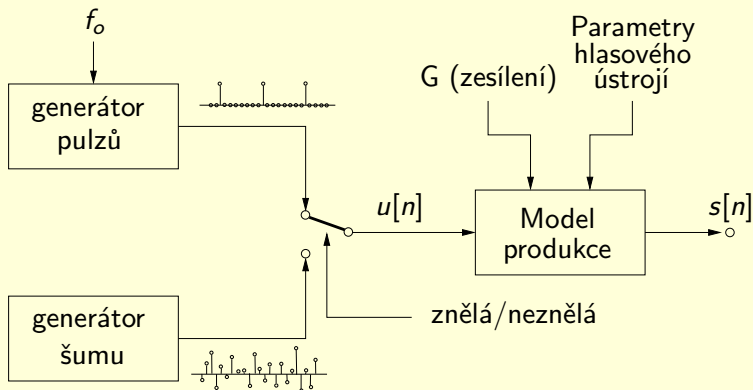


Hlásky “š” ... neznělá, šumový charakter

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

Hlásky “t” ... plovivní, 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 vlastností hlasivek

Možnosti biometrické identifikace na bázi hlasu

Originalita hlasu - výška a barva

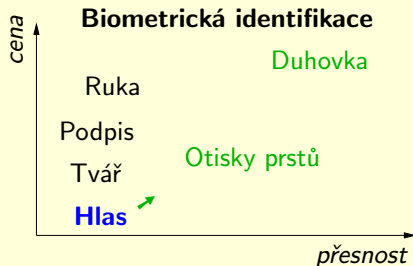
- dané fyzikálními rozměry (anatomii) hlasového ústrojí → 😊

Originalita stylu - doba trvání hlásek, intonace, apod.

- dané dynamikou pohybu hlasového ústrojí → 😊

Obecná **variabilita** jednotlivých realizací - **PROBLÉM** → 😞

Možnost **napodobení hlasu** - **PROBLÉM** → 😞



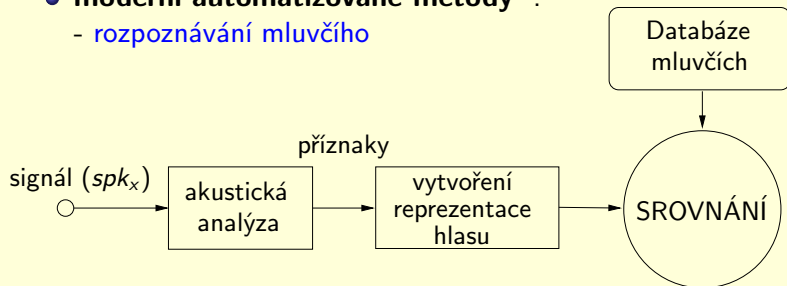
Motivace pro použití hlasové identifikace

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

Možnosti identifikace mluvčího

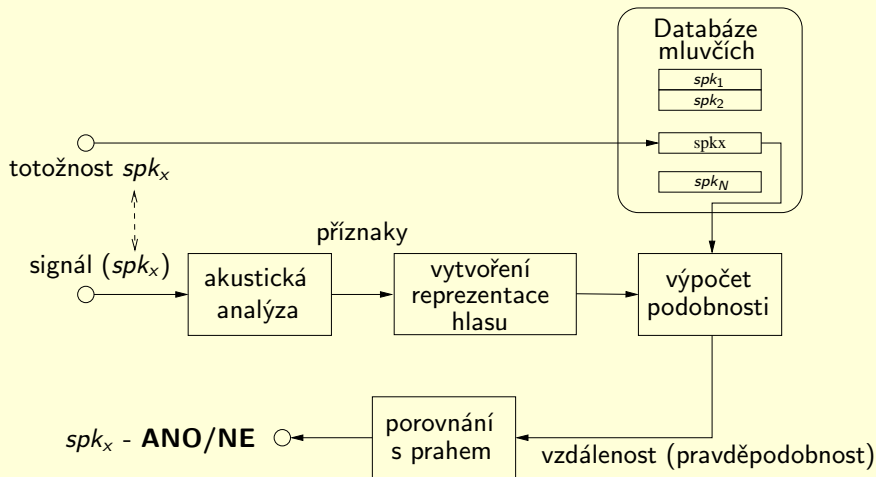
- *historické klasické přístupy ve forenzní praxi :*
 - *expertní rozhodování (fonetici, lingvisté)*

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



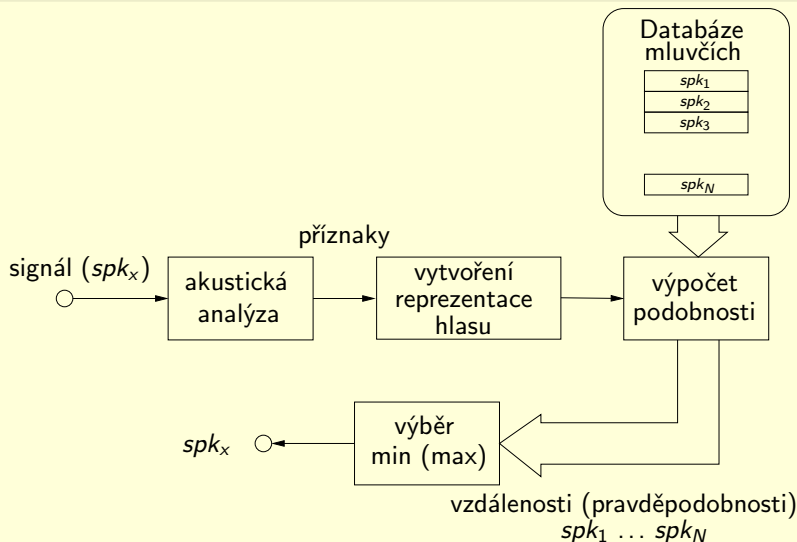
- **řečové příznaky**
 - spektrum, kepsstrum, formanty, základní tón, apod.
- **reprezentace mluvčího a metody srovnání**
 - DTW, VQ, GMM, HMM, i-vektory, ANN/DNN (x-vektory)

Verifikace mluvího



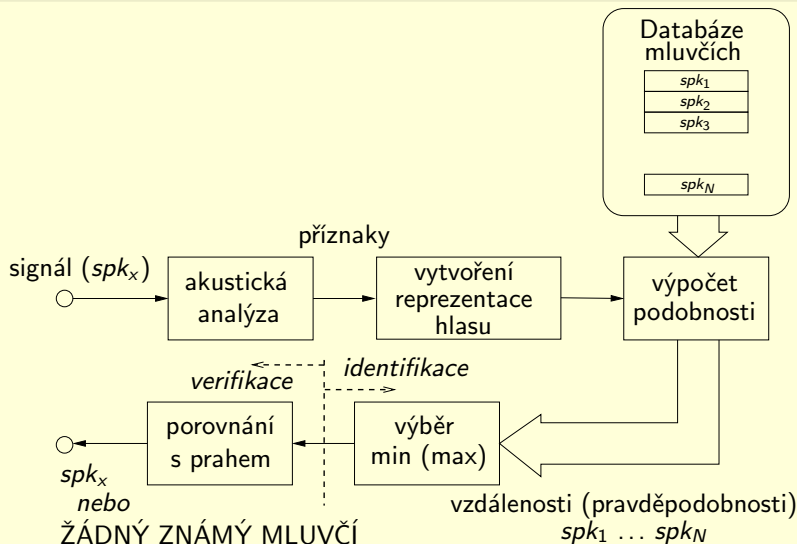
- ověření předpokládané totožnosti mluvího
- **VÝSLEDEK** = **přijetí** / **odmítnutí** předpokl. totožnosti

Identifikace mluvčího (v uzavřené množině)



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

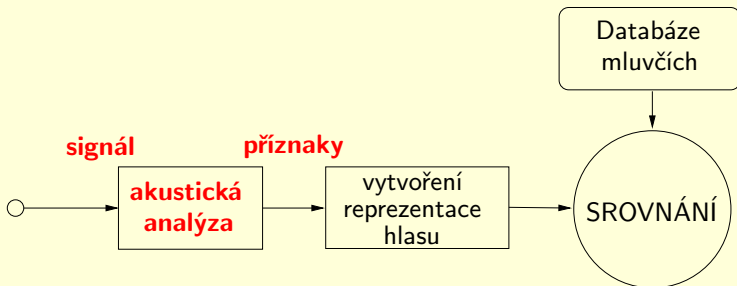
Identifikace mluvího (v otevřené množině)



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

II. část

Řečové charakteristiky a možnosti využití pro 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



Vnitřní charakteristiky - související s vytvářením řeči

Získané charakteristiky - souvisejí s dynamikou pohybů hlasového traktu (dané prostředím)

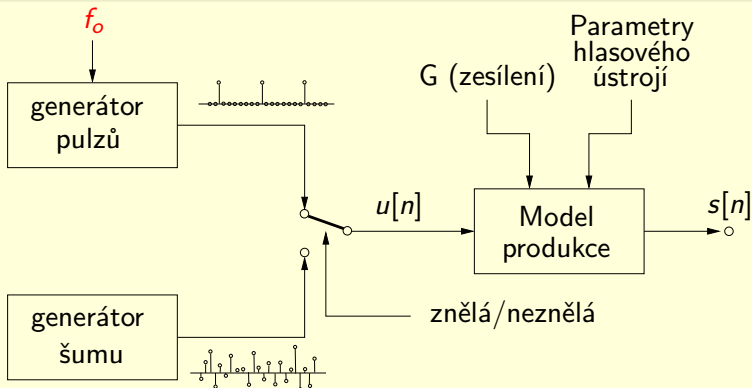
Vnitřní charakteristiky řečníka

- + obtížně cíleně ovlivnitelné
(dané fyzikálními rozměry hlasového ústrojí)
 - ovlivnitelné zdravotním stavem
(např. nosní dutina : neměnné rozměry při artikulaci,
mírné nachlazení = zásadní změna)
-

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
- způsob řeči lze snadno napodobit

Základní tón řeči



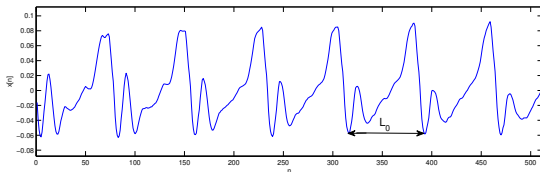
- základní frekvence $f_o = 1/T_o$
- pro znělé hlásky s harmonickou strukturou
- souvisí s kmitáním hlasivek
- hodnota f_o je ovlivněna vlastnostmi hlasivek (pružnost, hmotnost, délka)
→ hrubá charakteristika mluvčího

Odhad základního tónu řeči

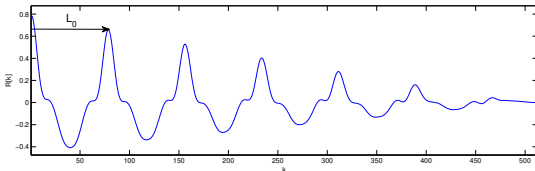
f_o základní tón (frekvence) řeči $f_o = \frac{1}{T_o}$
 T_o (L_o) základní perioda (v sekundách vs. ve vzorcích)

Nejčastější metoda odhadu - na bázi autokorelační funkce
(hledání postranního maxima autokorelační funkce)

segment signálu

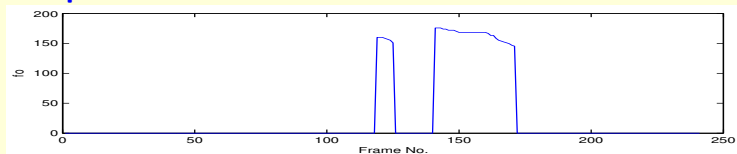


odhad autokorelační funkce

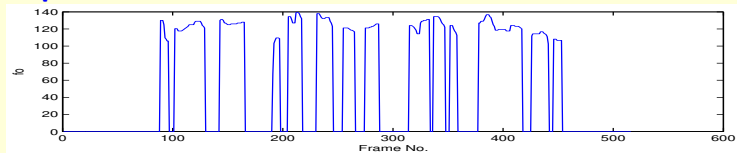


Průběh základního tónu v promluvě

Krátká promluva - slovo

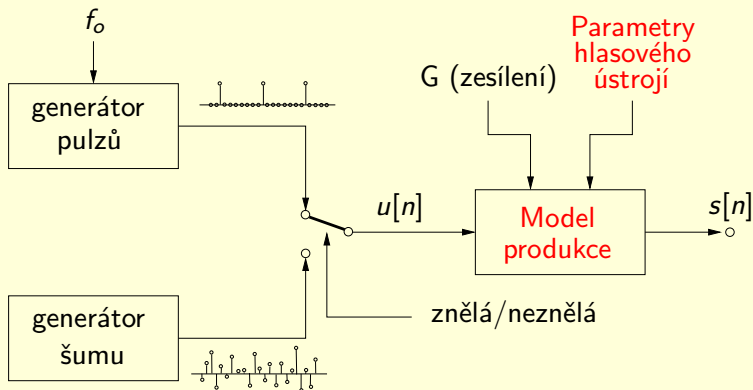


Delší promluva - věta



Průběh f_0 v promluvě → získaná (naučená) charakteristika
Průměrná hodnota f_0 → vnitřní charakteristika (výška hlasu)

Spektrální charakteristiky řeči



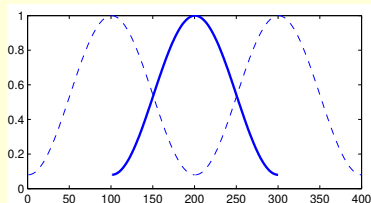
- spektrální charakteristiky souvisí s vokálním traktem
- otázka vhodné reprezentace pro identifikaci

Odhad spektra na bázi DFT:

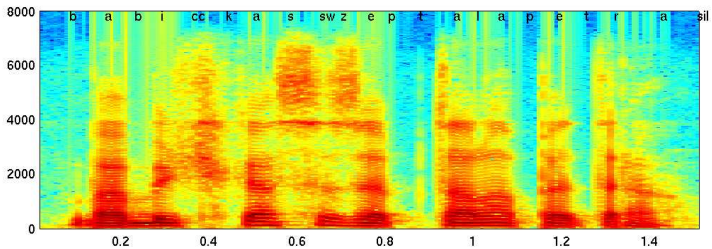
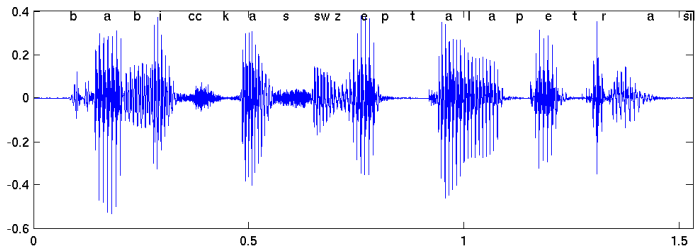
- **řeč je obecně nestacionární signál** $\Rightarrow$ nutná segmentace a sledování vývoje krátkodobého spektra (spektrogram)
- **řeč je kvazistacionární**
(tj. stacionární v krátkém časovém intervalu - cca 10-100 ms)
 $\Rightarrow$ 20-30 ms - typická délka krátkodobého segmentu
- **DFT spektrum je ovlivněno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

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

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

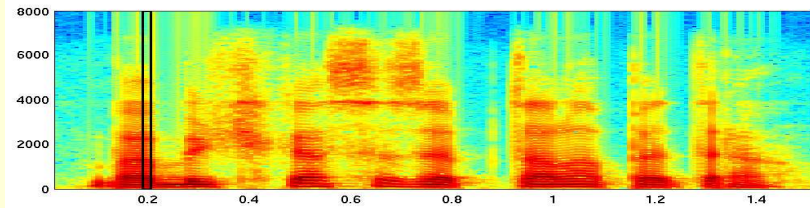


Časový průběh a spectrogram řeči



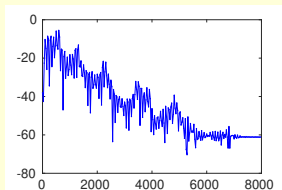
Přehled možností spektrální reprezentace promluvy

Spektrogram celé promluvy



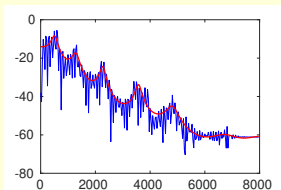
Spektrální reprezentace vybraného segmentu

DFT spektrum:



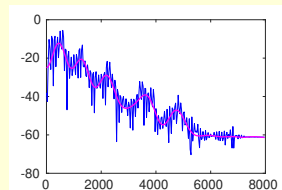
256 vzorků spektra
(amplitudové sp.)

LPC spektrum:



16 koeficientů a_k
(autoregresní koef.)

Kepstrální koeficienty:



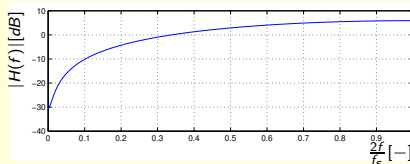
20 koeficientů c_n
(reálné kepstrum)

Preemfáze signálu - kompenzace útlumu vyšších kmitočtů

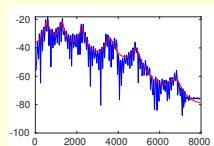
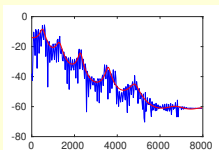
Sklon amplitudového spektra - vyšší kmitočty - nižší energie

Preemfázový filtr (FIR 1.řád):

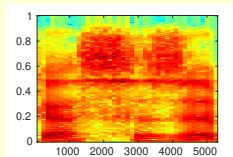
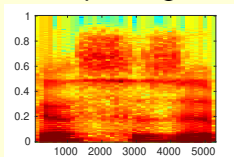
$$s'[n] = s[n] - m \cdot s[n - 1] \\ (m = 0.97)$$



Vliv preemfáze v krátkodobém spektru (DFT a LPC)



Vliv preemfáze ve spektrogramu



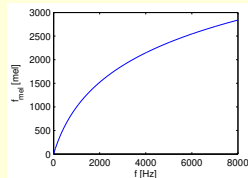
Banka filtrů s melovskou nelineární frekvenční osou

Hlavní cíl → počítá se výkon (energie) ve zvolených pásmech
Nelineární zkreslení frekvenční osy (nelineární vnímání frek.)

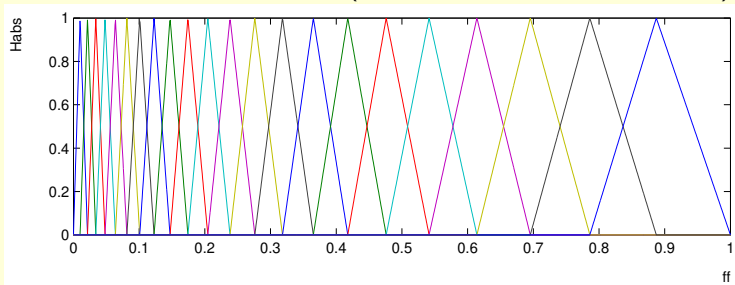
Melodická stupnice:

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

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



Trojúhelníková melovská BF (používaná pro výpočet MFCC)



BF je realizovaná na bázi DFT

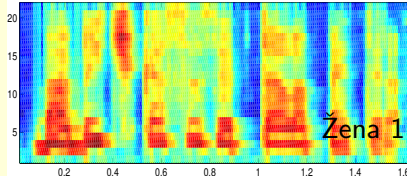
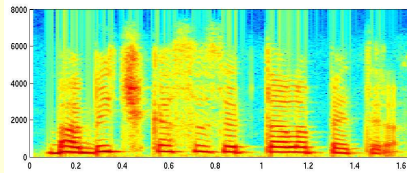
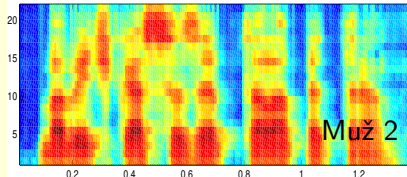
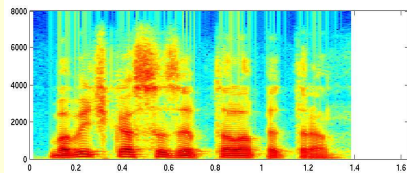
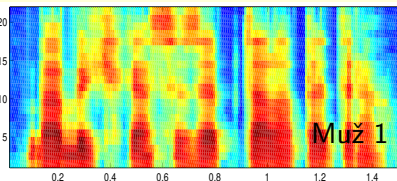
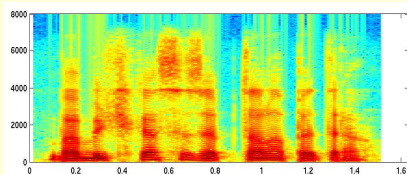
- ⇒ filtry jsou dány vahami DFT čar pro dané rozlišení (NDFT) a f_s
- ⇒ princip výpočtu je stejný pro všechny BF
- ⇒ pro jinou BF pouze jiné konkrétní váhy (např. PLP BF)

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

M - počet pásem 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



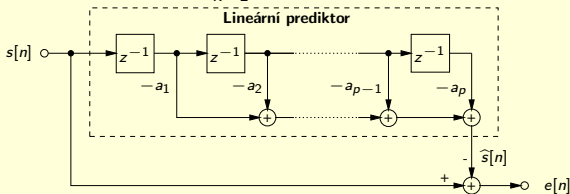
Výstup BF - **pásmový spektrogram** - častý příznak na vstupu DNN

- možná nevýhoda - vyhlazování rozdílů mezi mluvčími

- při použití časového kontextu - zlepšení (zahrnuje styl řeči)

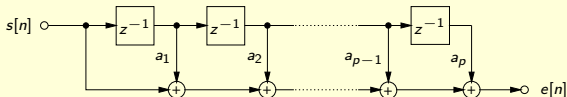
Lineární prediktivní analýza

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



Chybový signál (míra kvality prediktoru)

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



IDEA: přesnější predikce → 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.

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

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

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

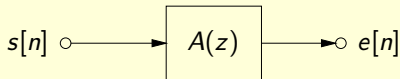
$R[k]$ autokorelační koeficienty analyzovaného signálu

VÝSLEDEK:

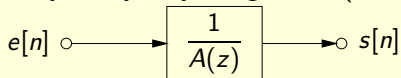
a_k autoregresní koeficienty (AR model signálu)

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

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

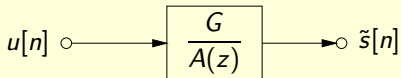


Syntéza se skutečným chybovým signálem (ideální případ)



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

→ tvar $S_{\tilde{s}}(e^{j\Theta})$ je kompletně zahrnut v AR modelu



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

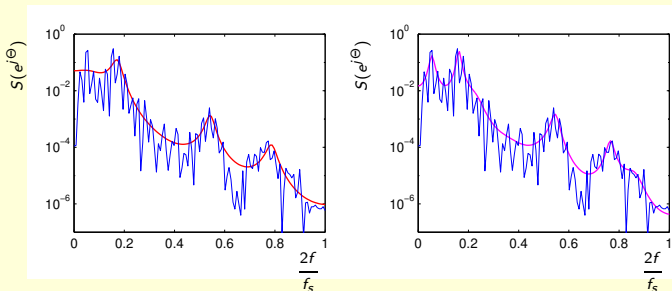
$$S_{\tilde{s}}(e^{j\Theta}) = |H(e^{j\Theta})|^2 = \frac{G^2}{|A(e^{j\Theta})|^2}$$



koeficienty a_k komprimovaná spektrální reprezentace

Srovnání LPC a DFT spektra

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



- 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
→ typické hodnoty: $p = 10$ pro $f_s = 8$ kHz, $p = 16$ pro $f_s = 16$ kHz
- pozice špiček v LPC spektru = **formantové kmitočty**

Formanty (formantové frekvence)

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



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

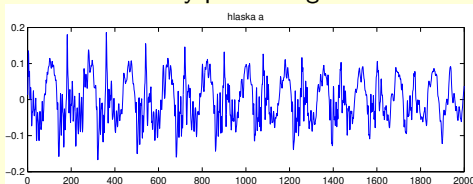
$$F_i = \frac{(2i - 1) \cdot c}{4 \cdot VTL}$$

Odhad na bázi LPC:

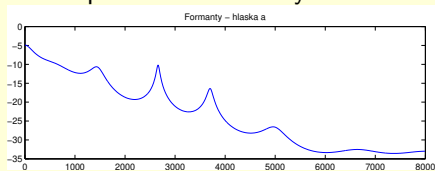
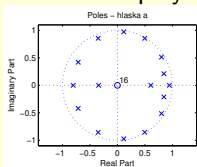
- z pólů (p_i) přenosové funkce $H(z) = \frac{G}{A(z)}$
- F_i - formantová frekvence (centrální kmitočet rezonátoru)
$$F_i = f_s \cdot \arg p_i / 2\pi$$
- B_i - šířka pásma formantu
$$B_i = -f_s \cdot \ln |p_i| / 2\pi$$

Odhad formantů na bázi LPC - příklad

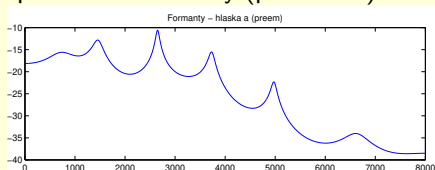
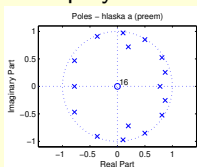
Časový průběh signálu



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-transformace

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

Výpočet pomocí DFT (numerický odhad) :

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

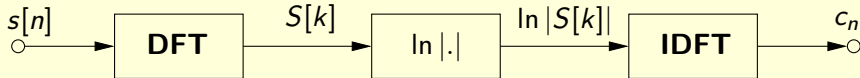
Obecné vlastnosti:

- $\hat{c}[n]$ 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

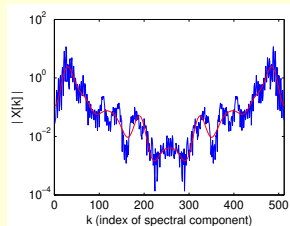
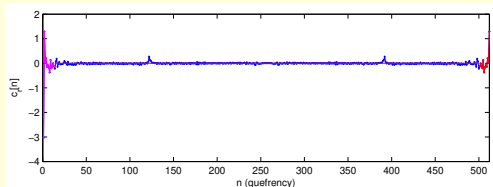
Názvosloví = slovní přesmyčky:

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

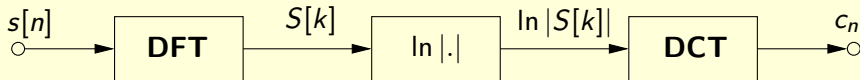
Vlastnosti reálného DFT kepra



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

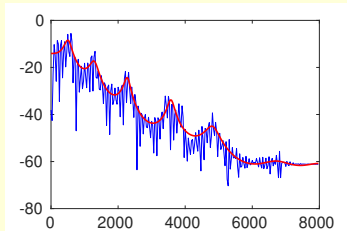


DCT místo IDFT při výpočtu reálného kepra reálného signálu

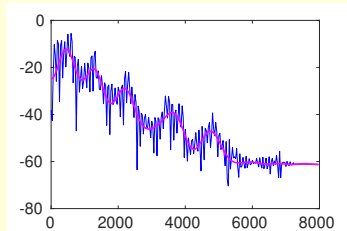


Kepstrální analýza pro zpracování řeči

LPC spektrum:



Vyhlazený odhad z reálného kepstra:



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



obecně vhodné příznaky pro rozpoznávání

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

⇒ **použití i pro identifikaci mluvčího**
(textově nezávislá identifikace)

Mel-kepstrum - melodická frekvenční stupnice

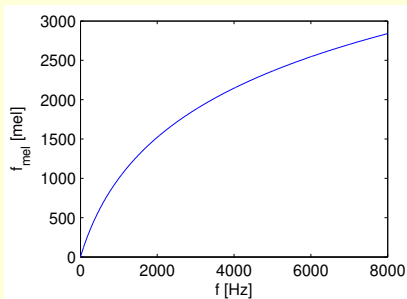
Vhodnější výpočet kepstra :

≈ **modelování nonlinearity vnímání frekvence lidským sluchem**

Nelineární zkreslení frekvenční osy - *melodická stupnice*

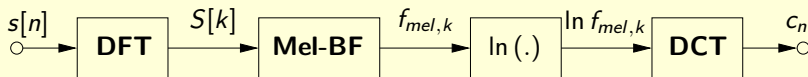
$$f_{mel} = \text{Mel}(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right)$$

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



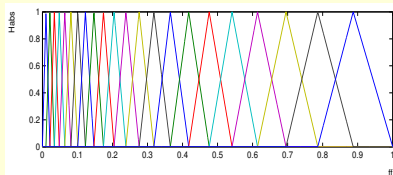
MFCC - Melovské keprstrální koeficienty

Blokové schéma výpočtu mel-keprstrálních koeficientů:



Výpočet energie v jednom pásmu

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



Výpočet keprstra pomocí DCT

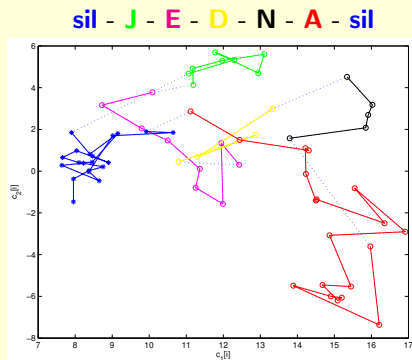
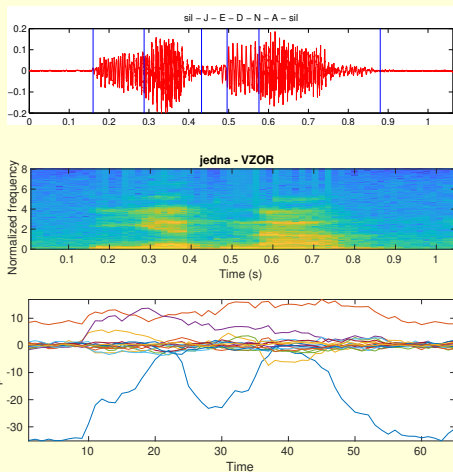
$$c_i = \sqrt{\frac{2}{P}} \sum_{j=1}^P g_j \cos \left(\frac{\pi i}{P} (j - 0.5) \right)$$

MFCC - nejrozšířenější příznaky používané pro

- rozpoznávání 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 kepler pro podobné zvuky (jednotlivé hlásky)

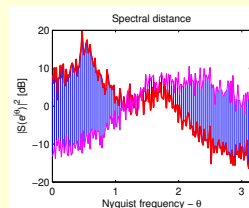


Rozlišení mluvčích v keprstrální oblasti později

Spektrální a keprální vzdálenost

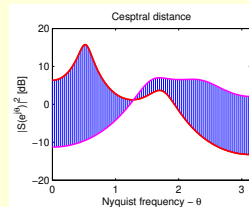
Spektrální vzdálenost (L_2 -norma)

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



Kepstrální vzdálenost

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



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

Varianty keprávní vzdálenosti

- Euklidovská vzdálenost:
$$CD = \sqrt{\sum_{k=0}^L (c_s[k] - c_x[k])^2}$$

- Eukl. vzdálenost bez $c[0]$:
$$CD = \sqrt{\sum_{k=1}^L (c_s[k] - c_x[k])^2}$$

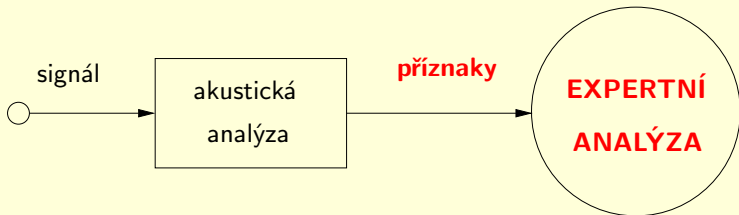
- kvadrát Euklidovské vzdálenosti bez $c[0]$:
$$CD = \sum_{k=1}^L (c_s[k] - c_x[k])^2$$

- vyřazení $c[0] \rightarrow$ odstranění závislosti na intenzitě hlasu
 $\Rightarrow$ **vhodné pro SRE**

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

III. část

Expertní metody v SRE



● Forenzní lingvistika a fonetika

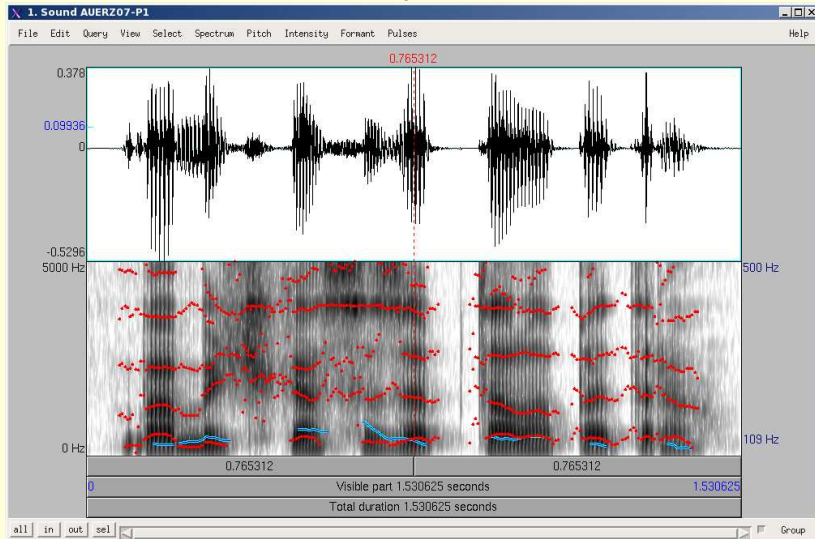
- sledování osobitých rysů projevu řečníka
- zaměření na artikulační zvláštnosti
- typické vedení melodie řeči (intonace)
- většinou na bázi poslechu

● Spektrografické metody

- Využívají možnost zobrazení diskutovaných hlasových charakteristik (spektrogramů, průběhu f_0 , trajektorií formantů, apod.)
- řešeno opět na expertní bázi
- detaily realizace vybraných hlásek

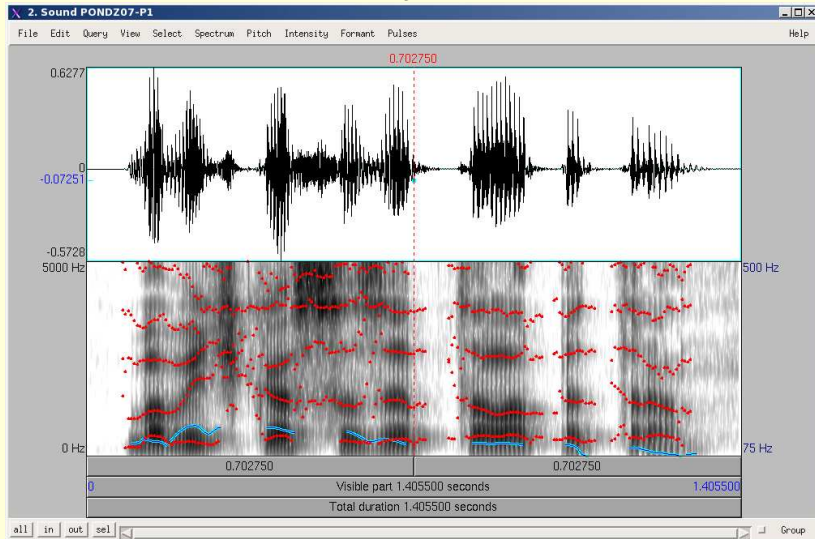
Formanty & základní tón - odhad Praat

Muž 1



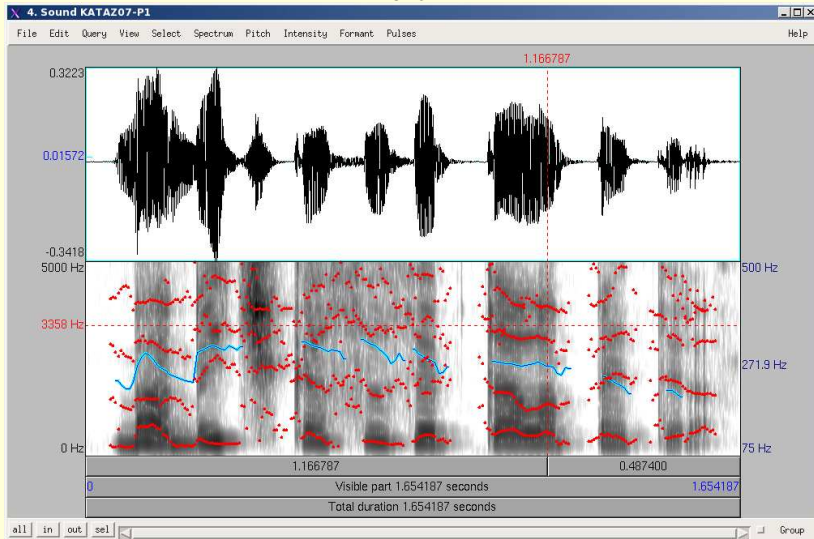
Formanty & základní tón - odhad Praat

Muž 2



Formanty & základní tón - odhad Praat

Žena 1



Děkuji vám za pozornost !