A6M33BIO - Biometrie

Biometrické metody založené na rozpoznávání hlasu

Doc. Ing. Petr Pollák, CSc.

18. prosince 2023 - 15:21

Obsah přednášky

- **Úvod**
  - Možnosti hlasové biometrické identifikace řečníka
  - Základní popis vzniku řeči (hlasu)
- **Řečové charakteristiky a příznaky pro identifikaci**
  - Spektrální charakteristiky, formanty
  - Kepstrum
  - Expertní identifikace, spektrografické metody
- **Automatického rozpoznávání řečníka**
  - Verifikace vs. identifikace
- **Reprezentace řečníka a algoritmy klasifikace**
  - GMM, GMM-UBM
  - i-vektory
  - x-vektory (DNN)
- **Příklady systémů verifikace**

I. část

Produkce a základní charakteristiky řeči

Možnosti hlasové identifikace řečníka

Motivace pro realizaci hlasové identifikace mluvčího

1) přesná identifikace totožnosti mluvčího

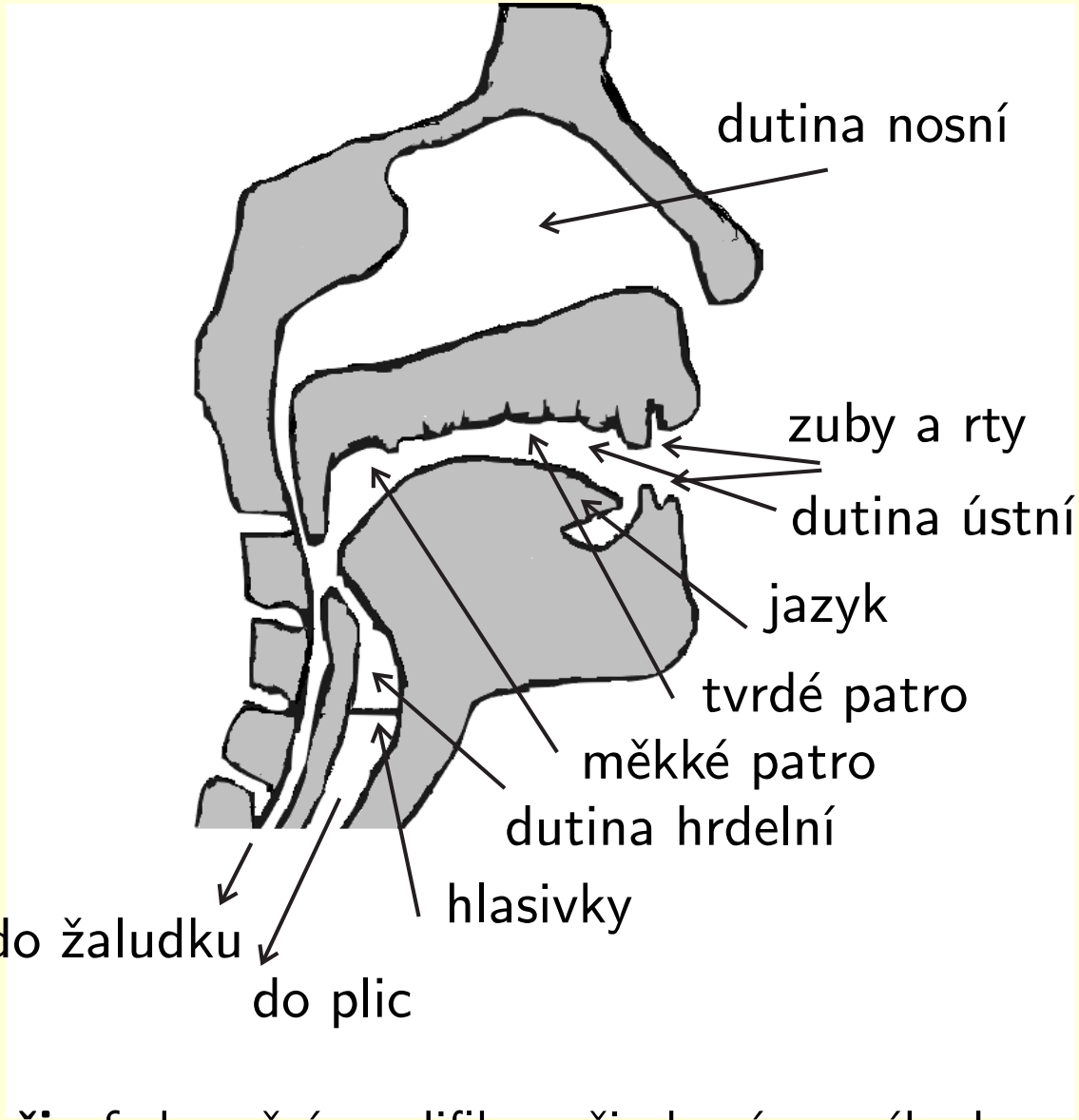
- kriminalistická a soudní praxe - forenzní aplikace
  (v minulosti subjektivní fonetická a lingvistická analýza)
- verifikace pro přístup k zabezpečeným systémům
  (osobn9 PC, mobilní telefony, bankovní účty, přístup do chráněných objektů/systémů, bezpečnostní kontroly, apod.)

2) identifikace mluvčího s největší podobností hlasu

- *př. - identifikace volajících v call-centrech*
- *komplexní rozpoznávače řeči*
  (LVCSR - diktovací systémy, transkripční systémy pro přepis rozhlasových/TV zpravodajství)
  - modely pro konkrétního mluvčího (GMM-HMM ASR)
  (skupinové modely - pohlaví, nářečí, apod.)
  - reprezentace mluvčího na vstupu ASR na bázi DNN

Produkce řeči - fyziologie hlasového ústrojí

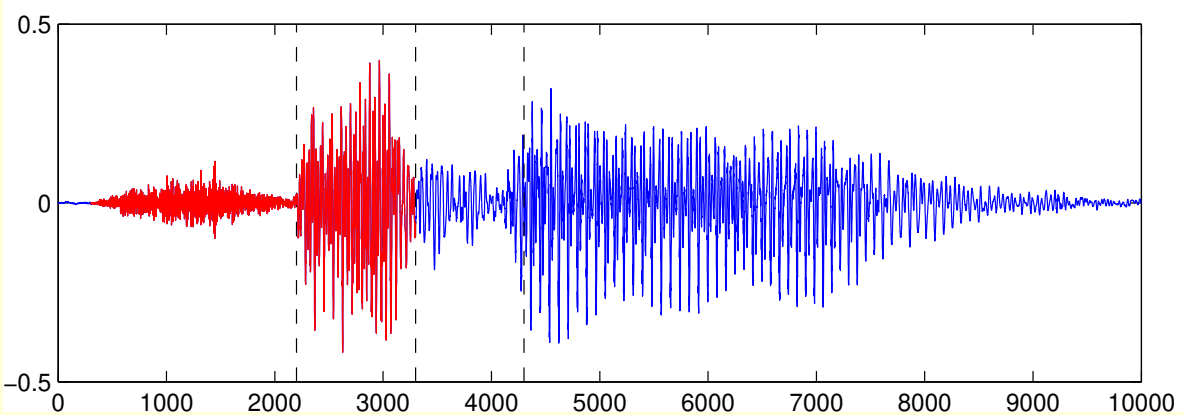
Artikulační orgány hlasového ústrojí člověka



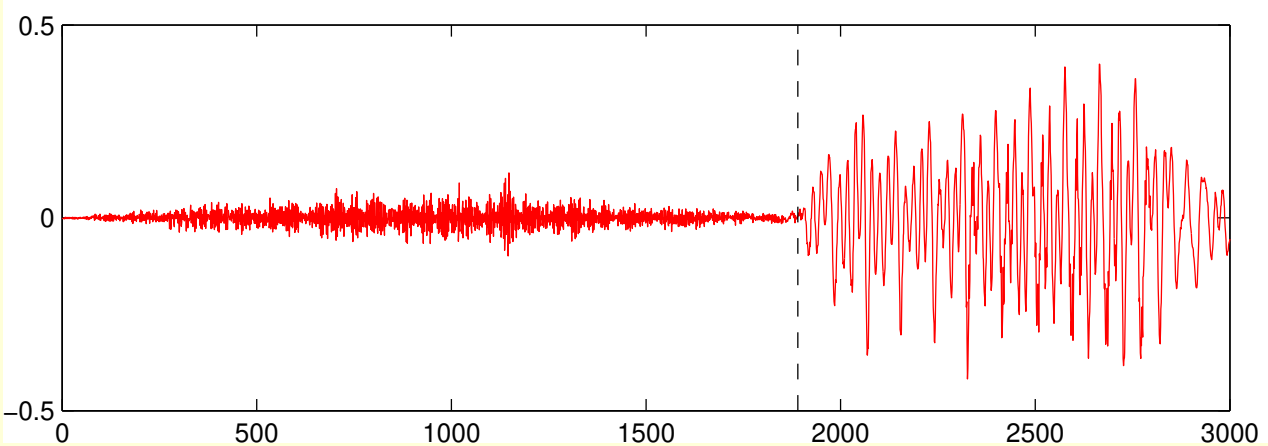
Produkce řeči : frekvenční modifikace širokopásmového buzení proudem vzduchu průchodem dutinami (rezonátory) hlasového ústrojí

Řečové hlásky v časové oblasti

Slovo **“šedý”**



Slabika **“še”**

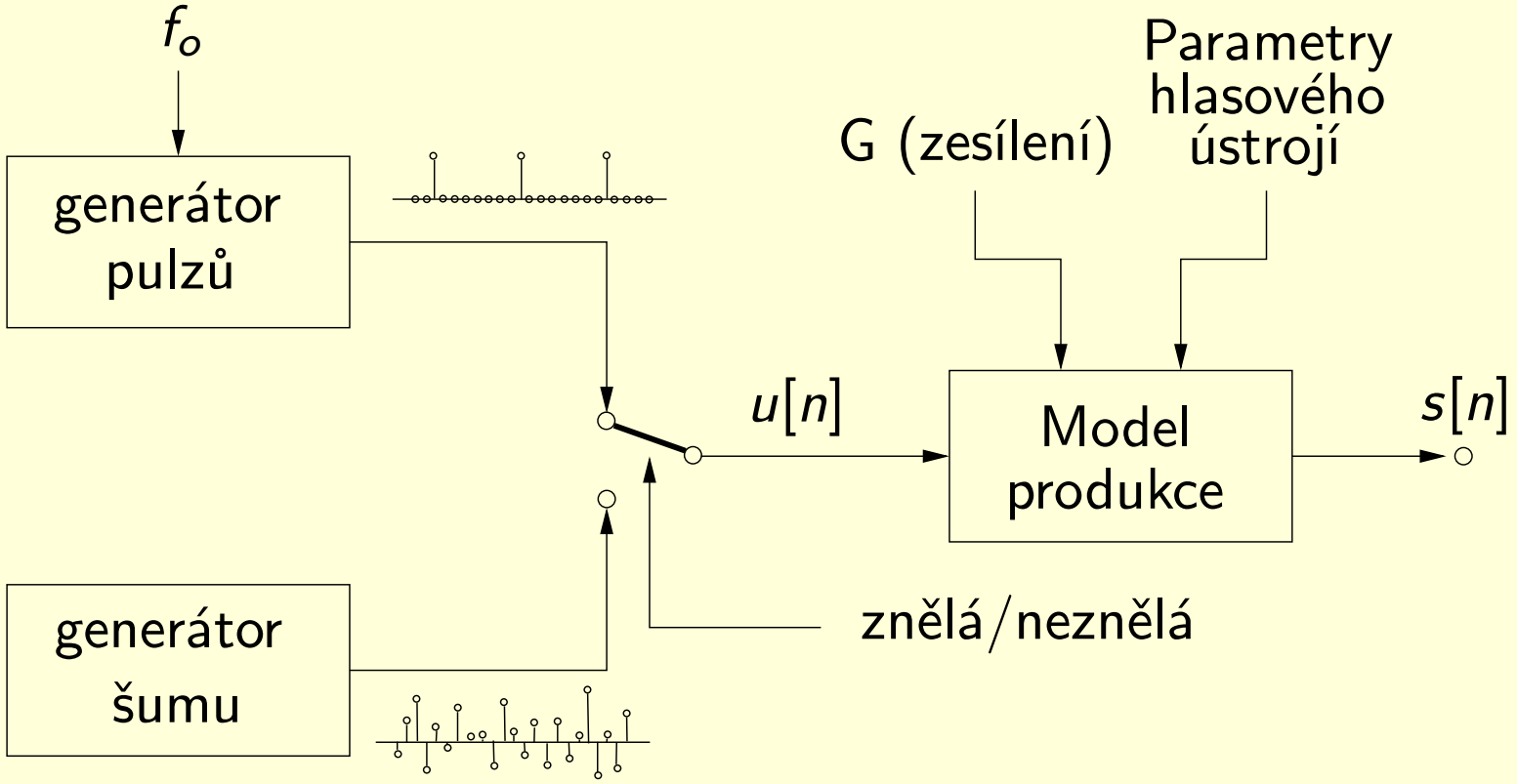


Hláska **“š”** … neznělá, šumový charakter

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

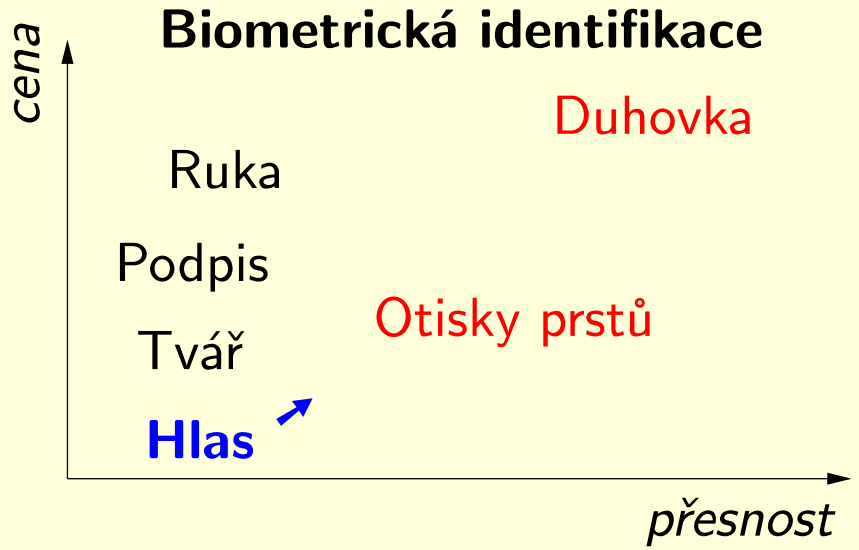
Možnosti biometrické identifikace na bázi hlasu

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 - **PROBLÉM** → ☹

Možnost napodobení hlasu - **PROBLÉM** → ☹

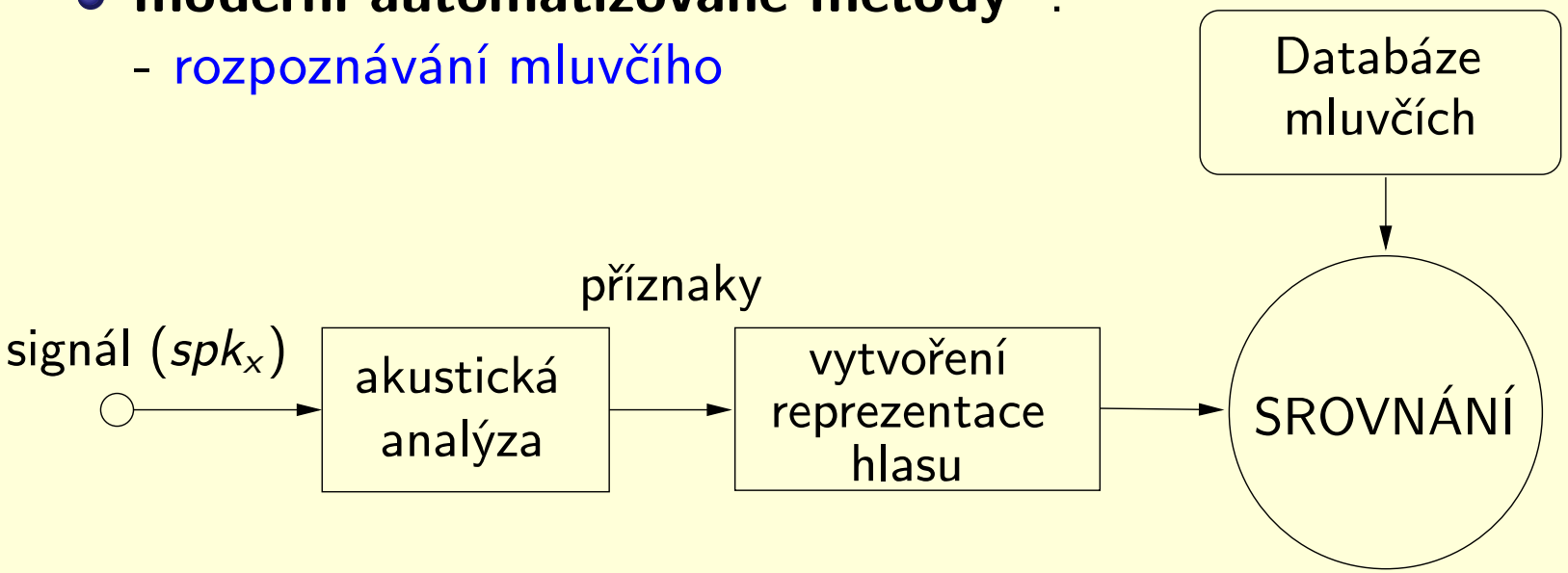


Motivace pro použití hlasové identifikace
- přirozenost komunikace, relativně jednodušší realizace
- jediná volba při dostupnosti pouze hlasového záznamu

Možnosti identifikace mluvčího

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

- **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)

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

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

Řečové příznaky pro hlasovou identifikaci - PROBLÉMY

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)

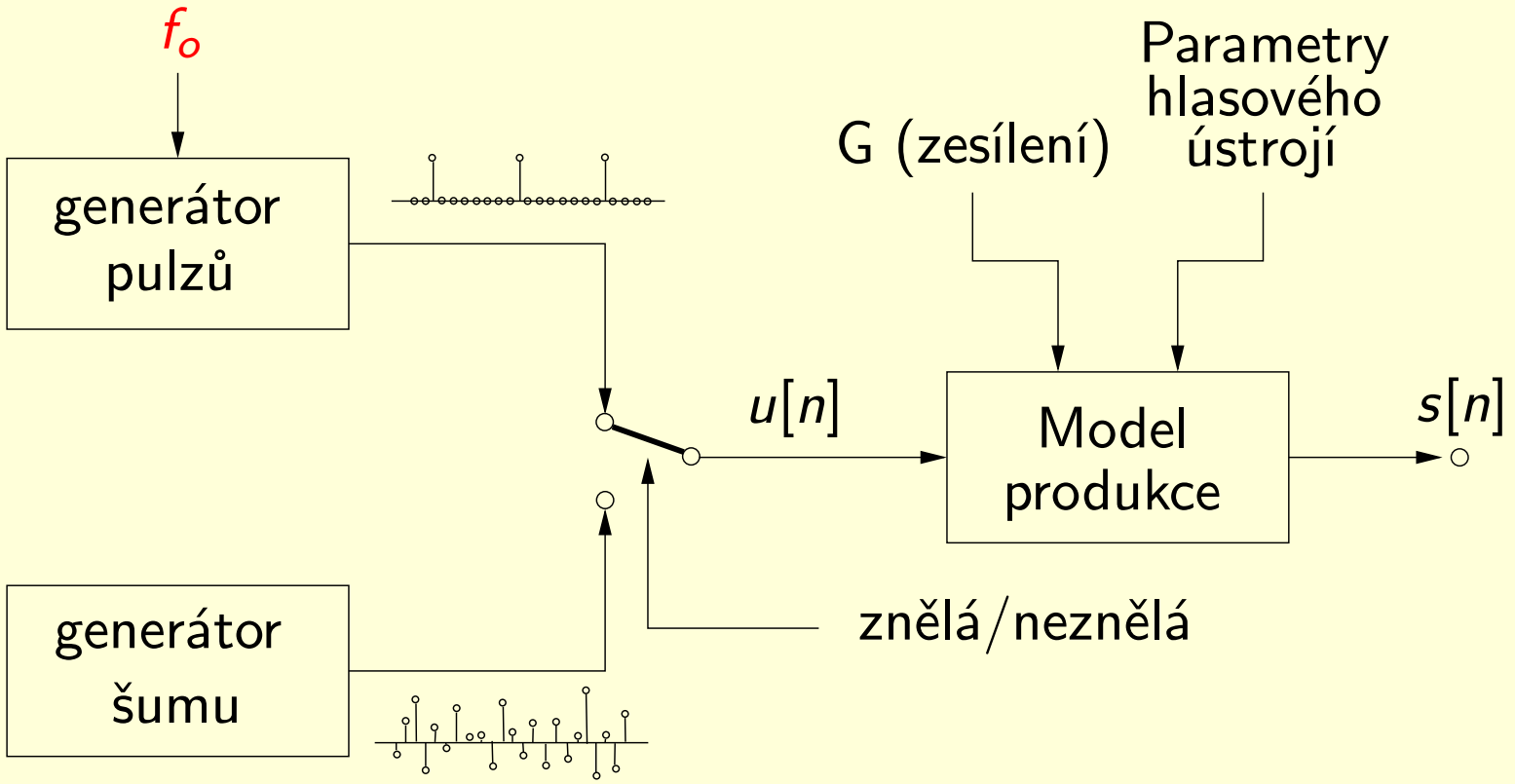
Získané charakteristiky řečníka

\+ 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)
  $\rightarrow$ hrubá charakteristika mluvčího

Odhad základního tónu řeči

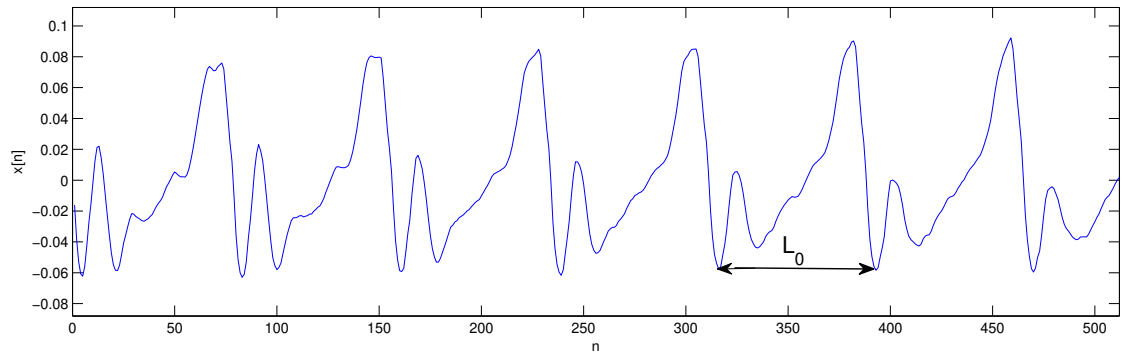
f_o ..... základní tón (frekvence) řeči ..... $f_o = \frac{1}{T_o}$

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

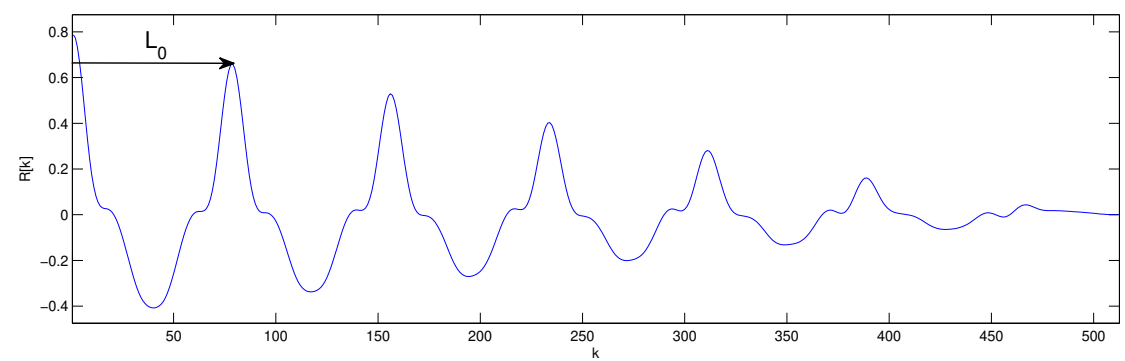
Nejčastější metoda odhadu - na bázi autokorelační funkce

(hledání postranního maxima autokorelační funkce)

segment signálu

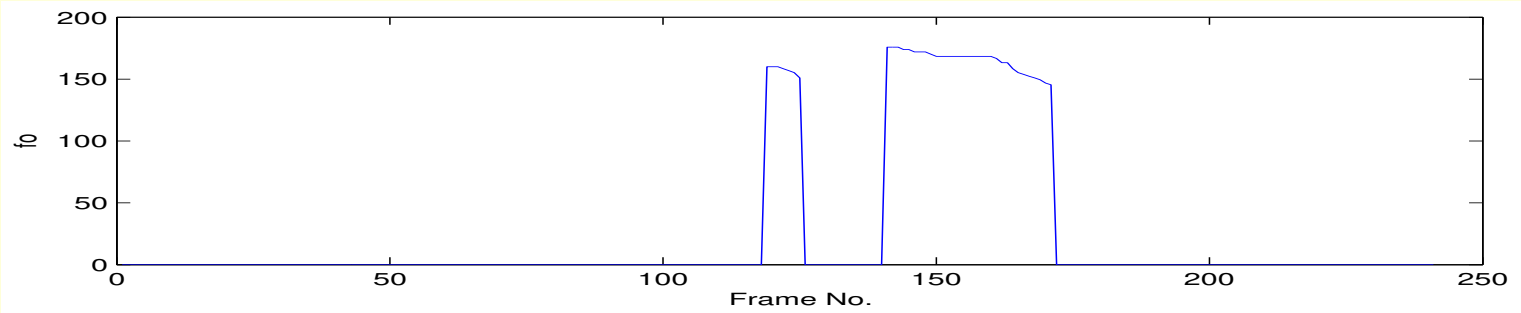


odhad autokorelační funkce

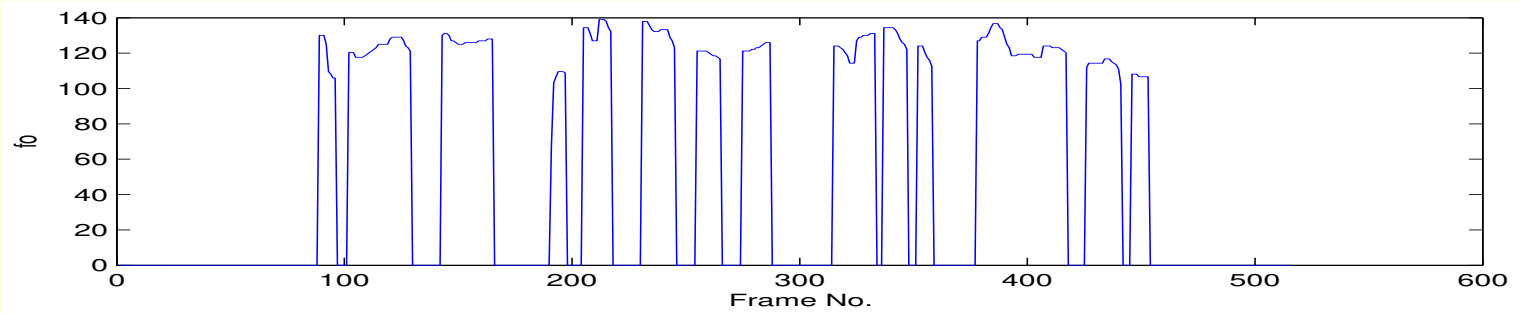


Průběh základního tónu v promluvě

Krátká promluva - slovo



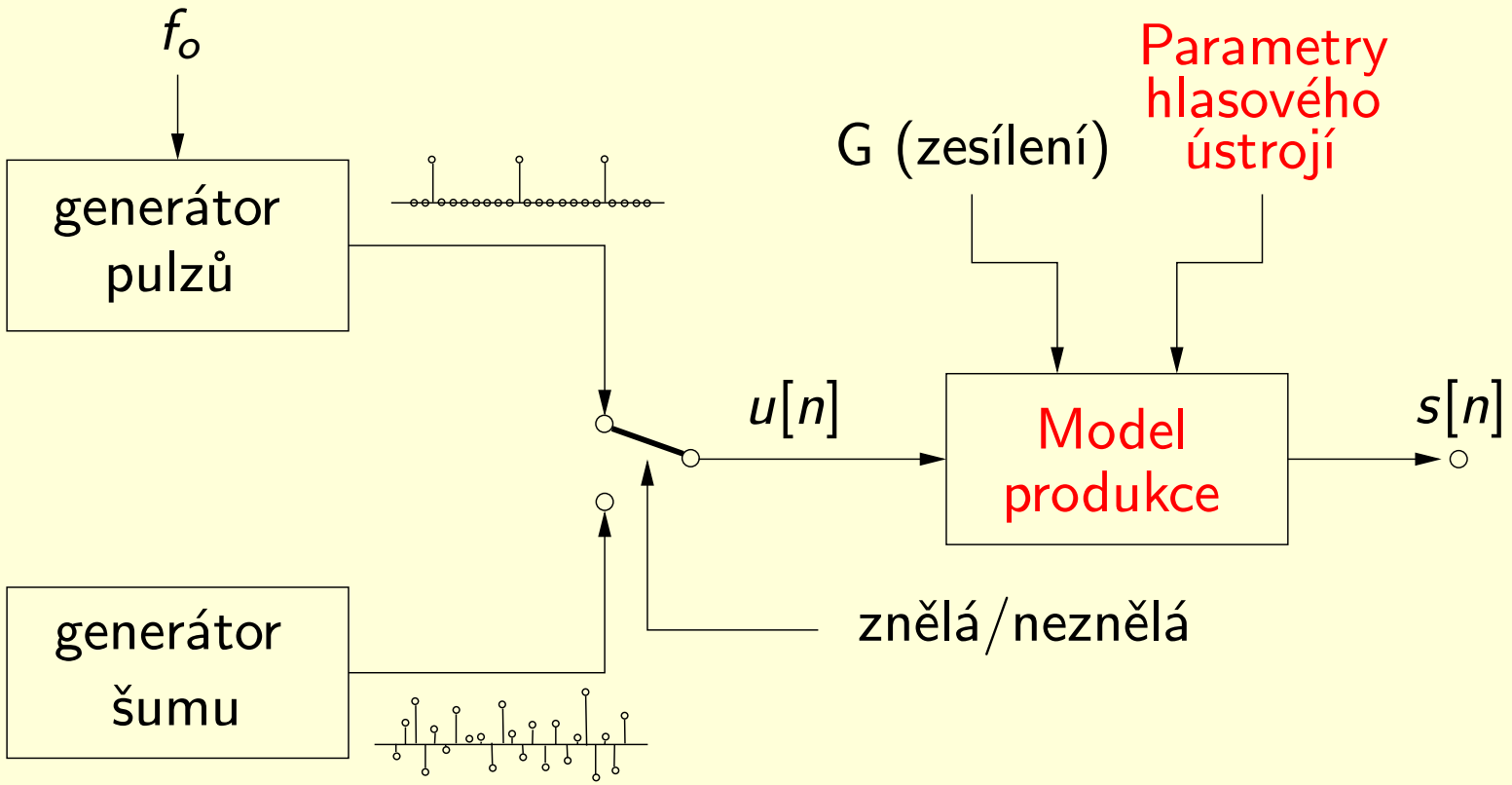
Delší promluva - věta



Průběh f_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

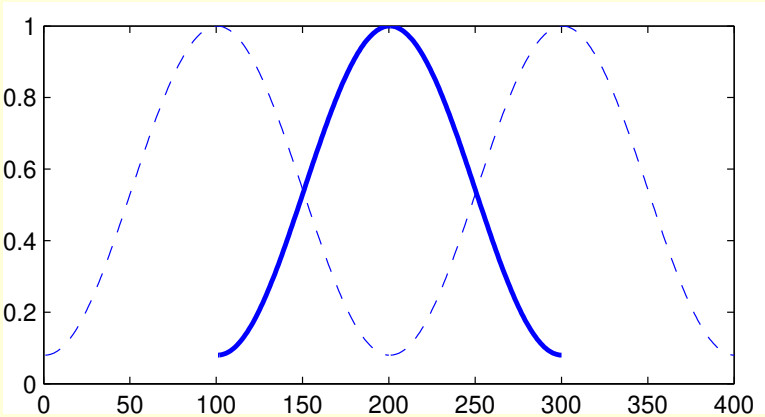
Spektrální reprezentace řeči na bázi DFT

Odhad spektra na bázi DFT:

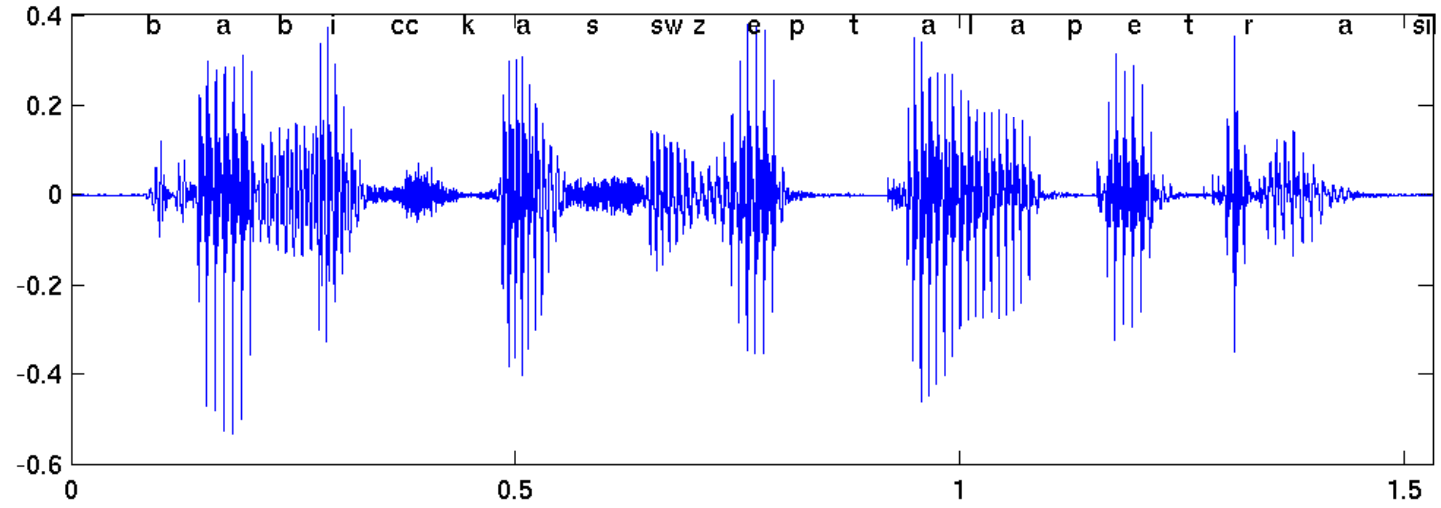
- **řeč je obecně nestacionární signál** $\Rightarrow$ nutná segmentace a sledování vývoje krátkodobého spektra (spektrogram)
- **řeč je kvazistacionární**
  (tj. stacionární v krátkém časovém intervalu - cca 10-100 ms)
  $\Rightarrow$ 20-30 ms - typická délka krátkodobého segmentu
- **DFT spektrum je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

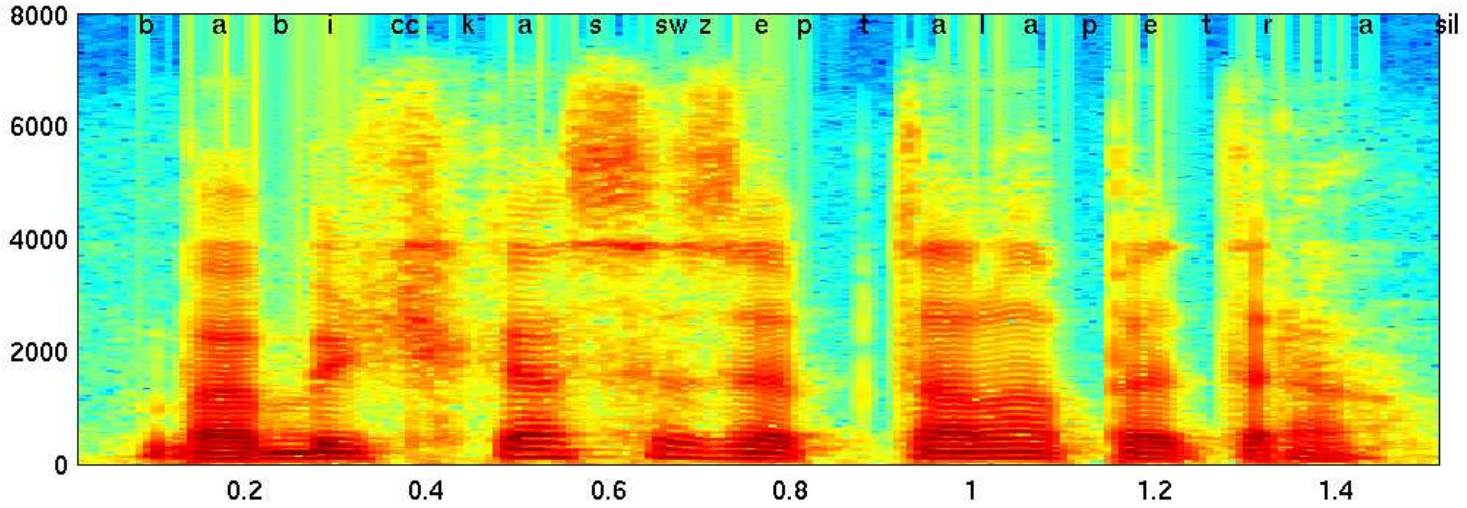
$$w[n] = 0,54 - 0,46 \cos \frac{2\pi n}{N}$$

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



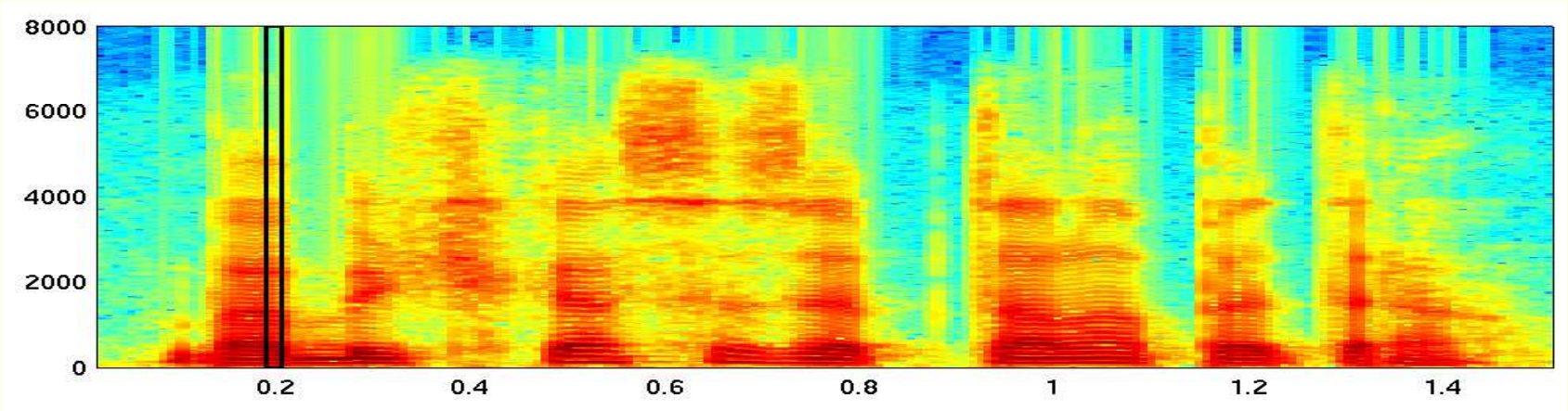
Časový průběh a spectrogram řeči





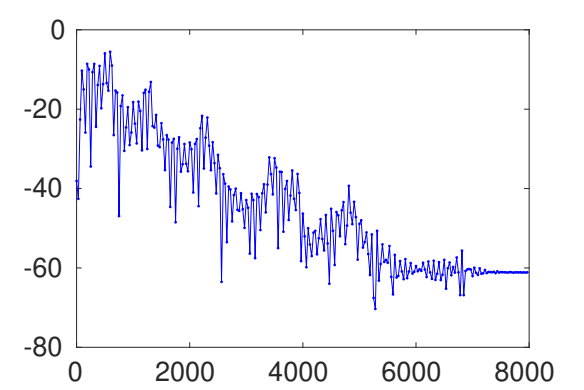
Přehled možností spektrální reprezentace promluvy

Spektrogram celé promluvy



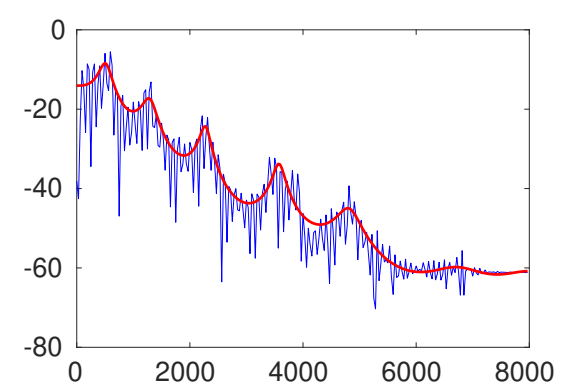
Spektrální reprezentace vybraného segmentu

DFT spektrum:



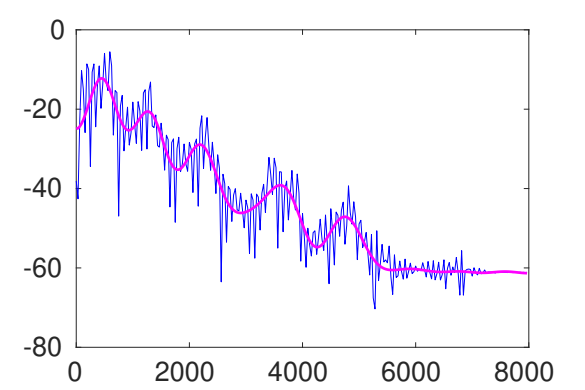
256 vzorků spektra
(amplitudové sp.)

LPC spektrum:



16 koeficientů a_k
(autoregresní koef.)

Kepstrální koeficienty:



20 koeficientů c_n
(reálné kepstrum)

Banky filtrů ve spektrální analýze

Hlavní cíl $\rightarrow$ počítá se výkon (energie) ve zvolených pásmech

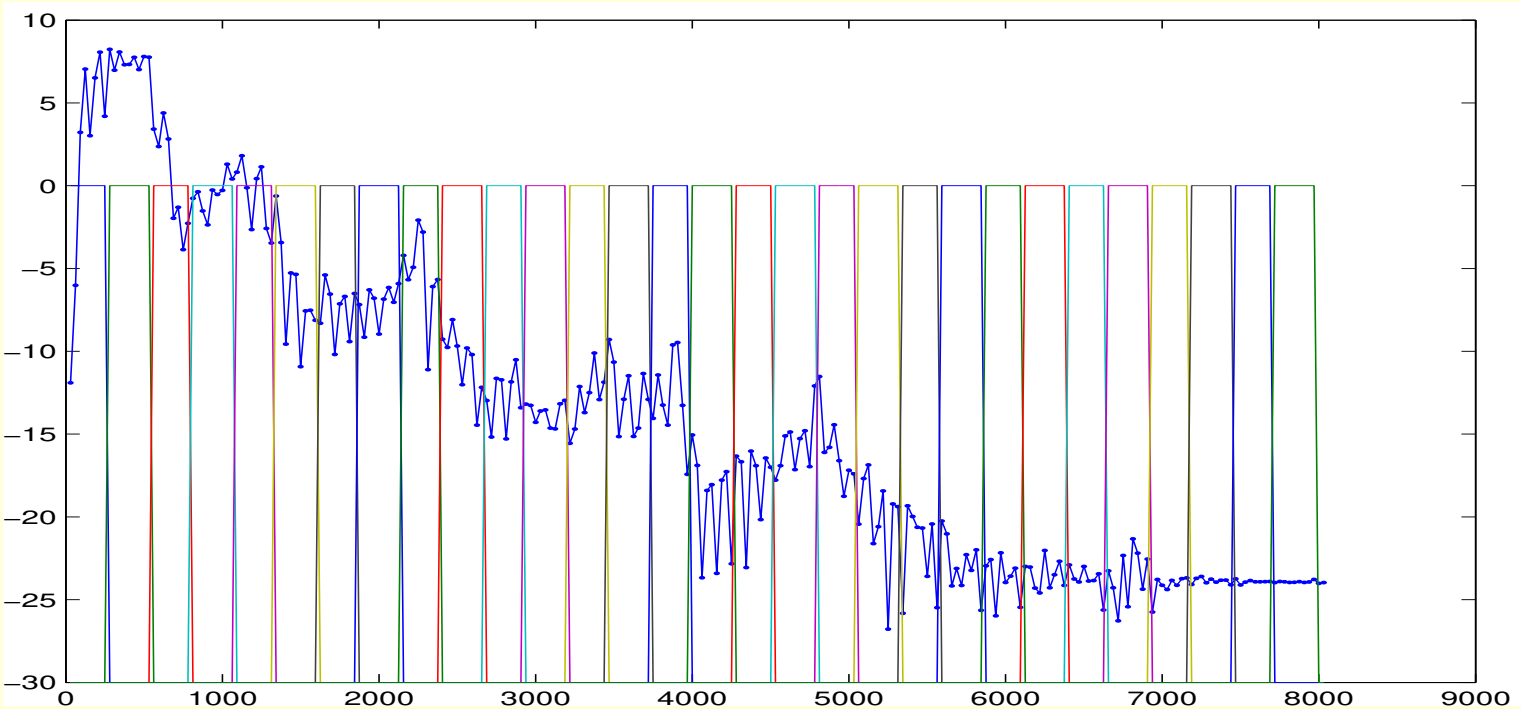
BF je realizovaná na bázi DFT

$\Rightarrow$ filtry jsou dány vahami DFT čar pro dané rozlišení (NDFT) a f_s

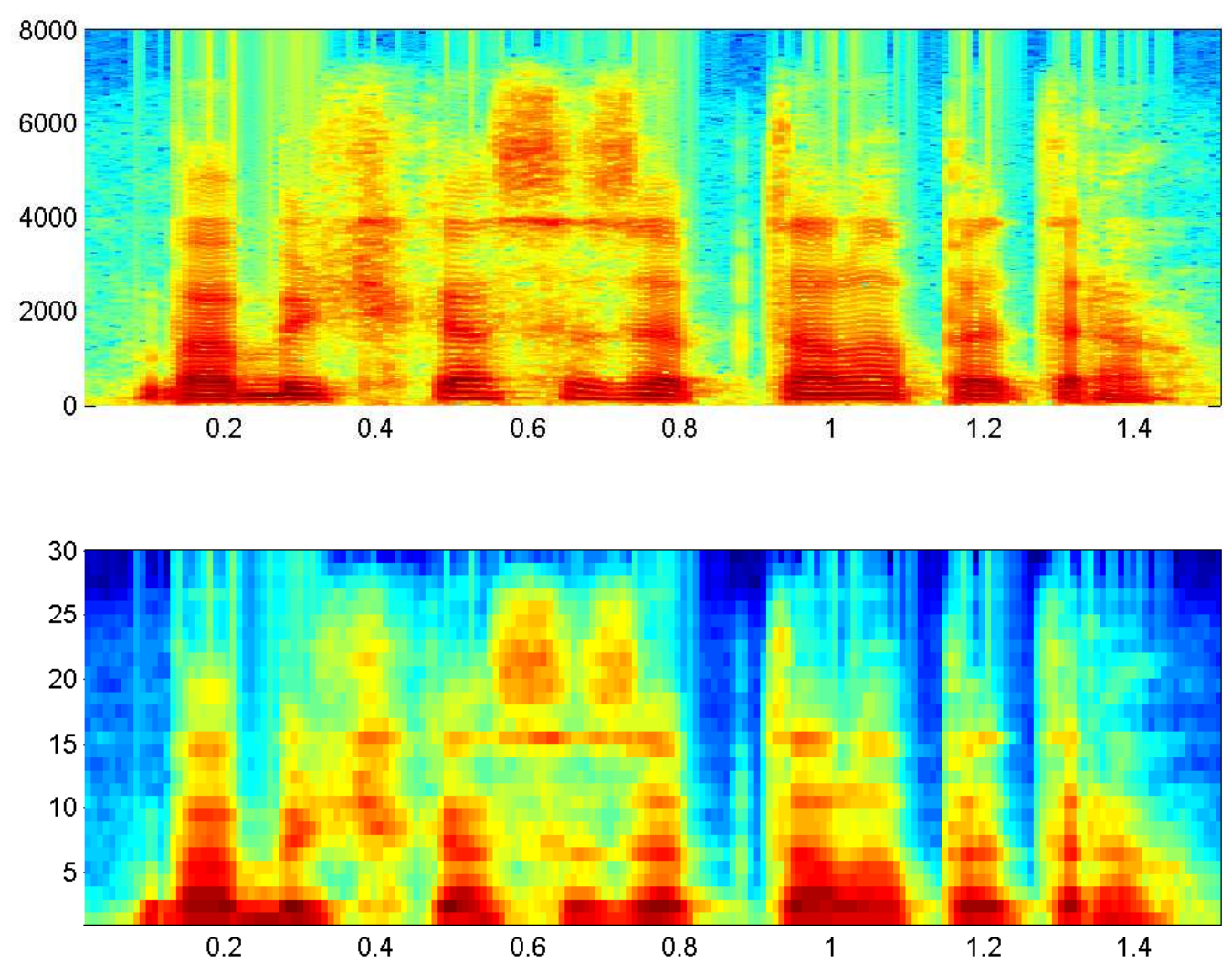
$$G_{mel}[j] = \sum_{k=0}^{N/2} |S[k]|^2 H_j[k] \quad \text{pro j} = 1, \ldots, \text{M}$$

M - počet pásem

- podle f_s, počtu bodů DFT a typu banky filtrů



Banky filtrů ve spektrální analýze



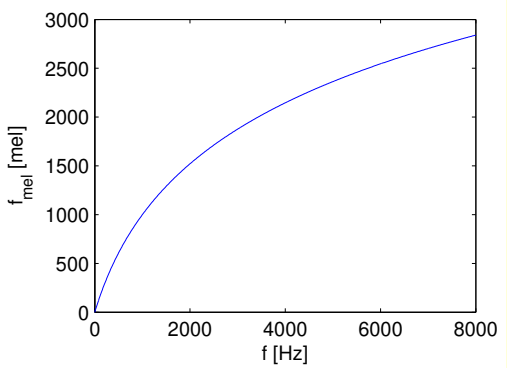
Lineární frekvenční osa - NEVÝHODA - hrubé rozlišení v DKP, jemné rozlišení v HKP (neodpovídá vnímání frekvence)

Banka filtrů s melovskou nelineární frekvenční osou

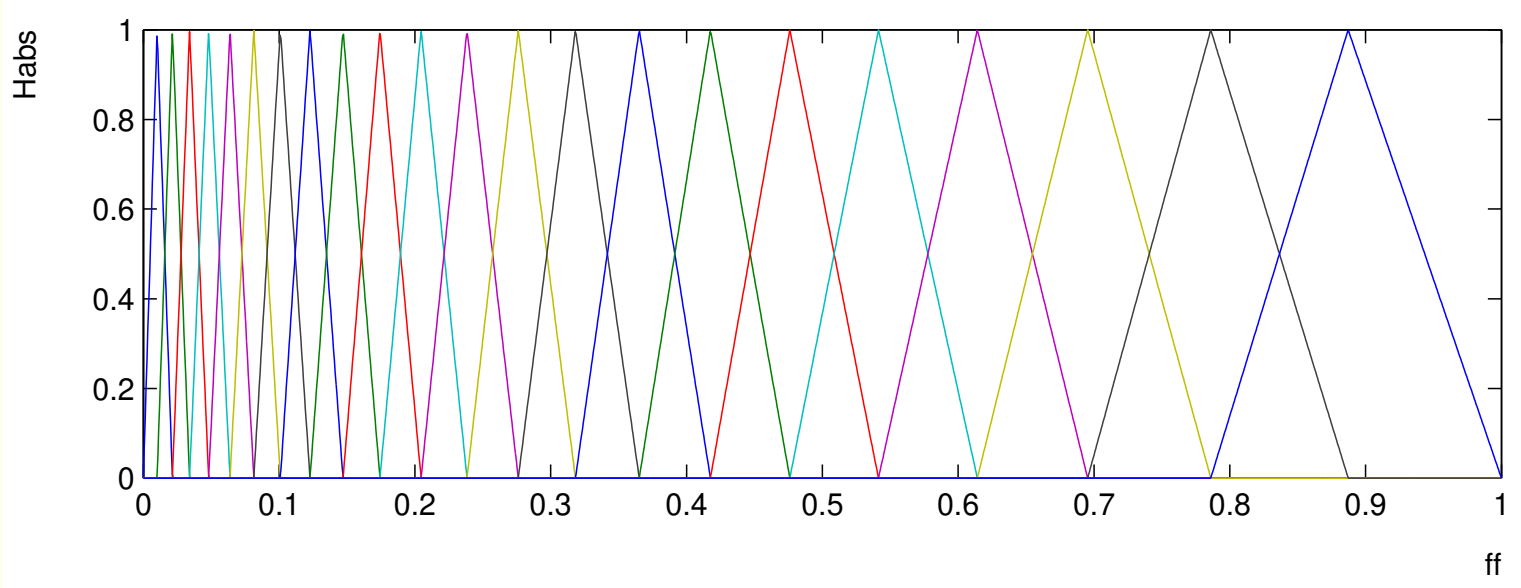
Nelineární zkreslení frekvenční osy - *melodická stupnice*

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

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



Trojúhelníková melovská BF (používaná pro výpočet MFCC)



Banka filtrů s melovskou nelineární frekvenční osou

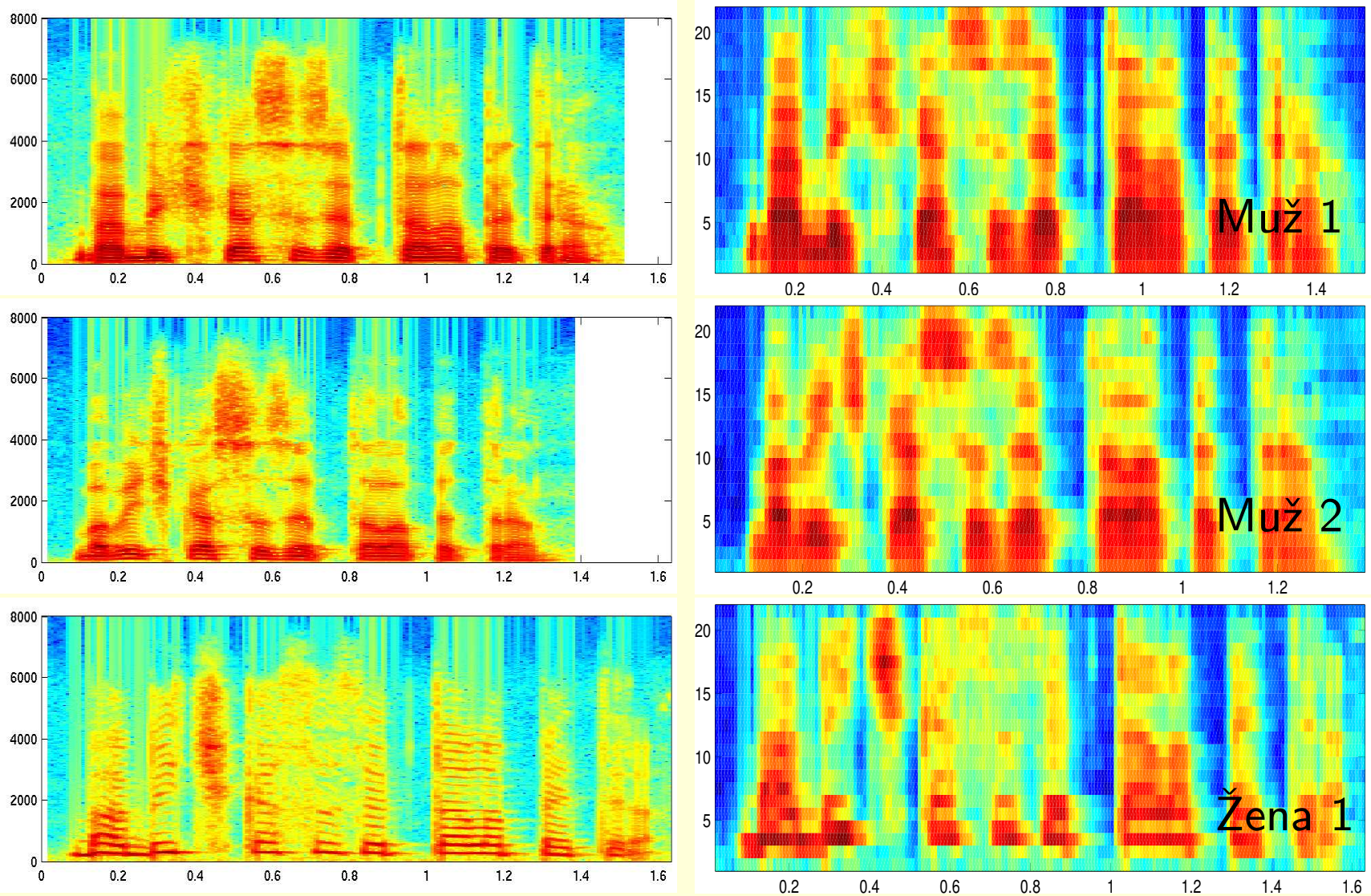
BF je opět realizovaná na bázi DFT

$\Rightarrow$ filtry jsou dány vahami DFT čar pro dané rozlišení (NDFT) a f_s
$\Rightarrow$ princip výpočtu je stejný pro všechny BF
$\Rightarrow$ pro jinou BF pouze jiné konkrétní váhy

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

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

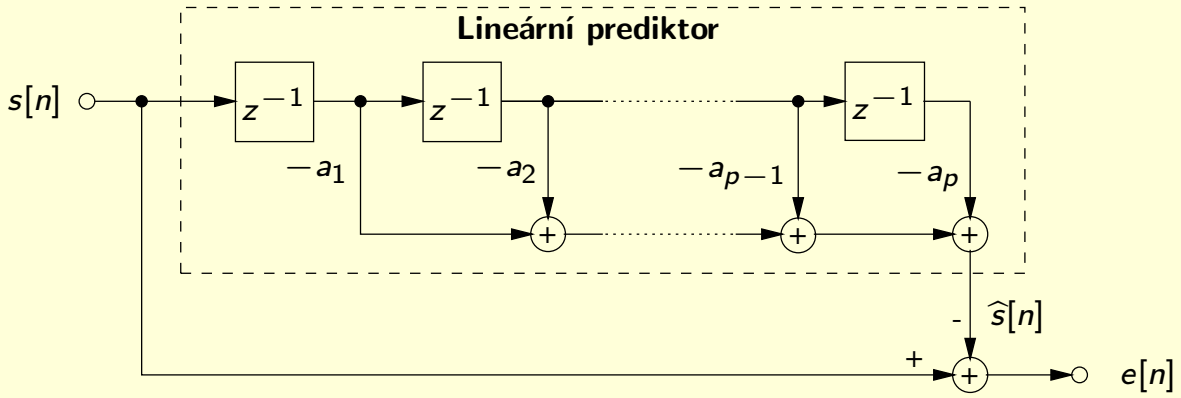
Variabilita promluvy v melovském spektrogramu



Výstup BF - pásmový spektrogram - častý příznak na vstupu DNN
- možná nevýhoda - vyhlazování rozdílů mezi mluvčími
- při použití časového kontextu - zlepšení (zahrnuje styl řeči)

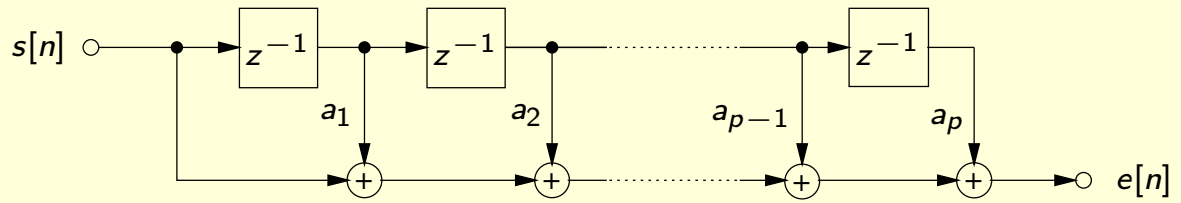
Lineární prediktivní analýza

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



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 $\equiv$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 (Yule-Walkerovy rovnice)
- *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

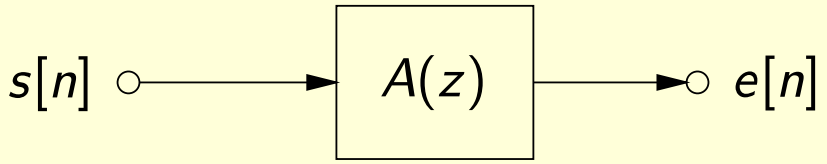
VÝSLEDEK:

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

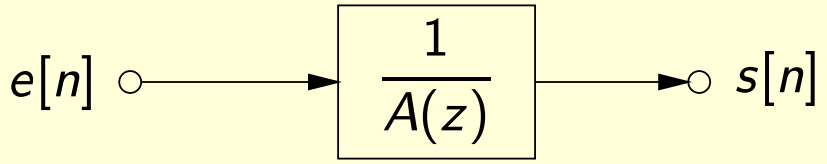
$P_p = R[0] + \sum_{k=1}^{p} a_k R[k]$ .... výkon chybového signálu

AR model signálu

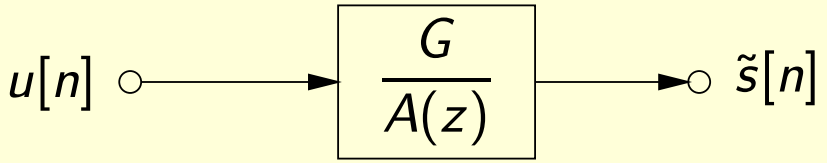
Dekorelační (analyzující) filtr : $A(z) = \sum_{k=0}^{p} a_k z^{-k}$



Syntéza se skutečným chybovým signálem (ideální případ)

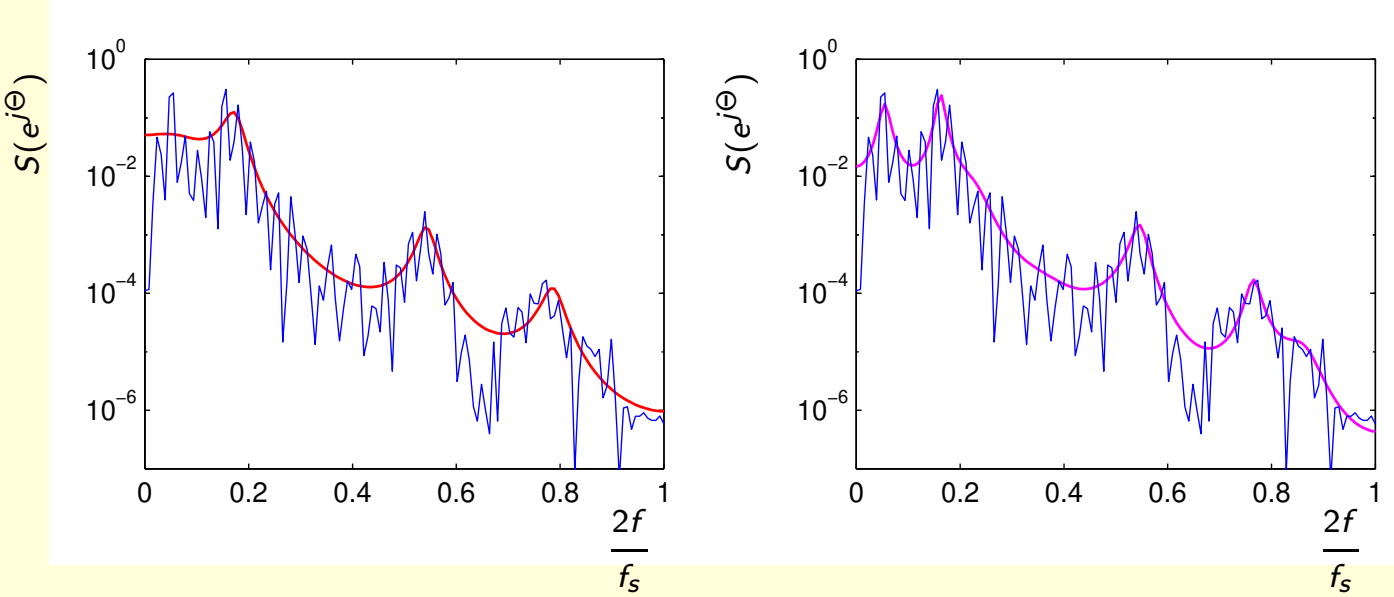


Syntéza s umělým signálem s jednotkovým výkonem (AR model)
- G závisí na úrovni analyzovaného signálu ($G = \sqrt{P_p}$)



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**

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** $H(z) = \frac{G}{A(z)}$
- F_i - formantová frekvence (centrální kmitočet rezonátoru)

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

- B_i - šířka pásma formantu

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

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

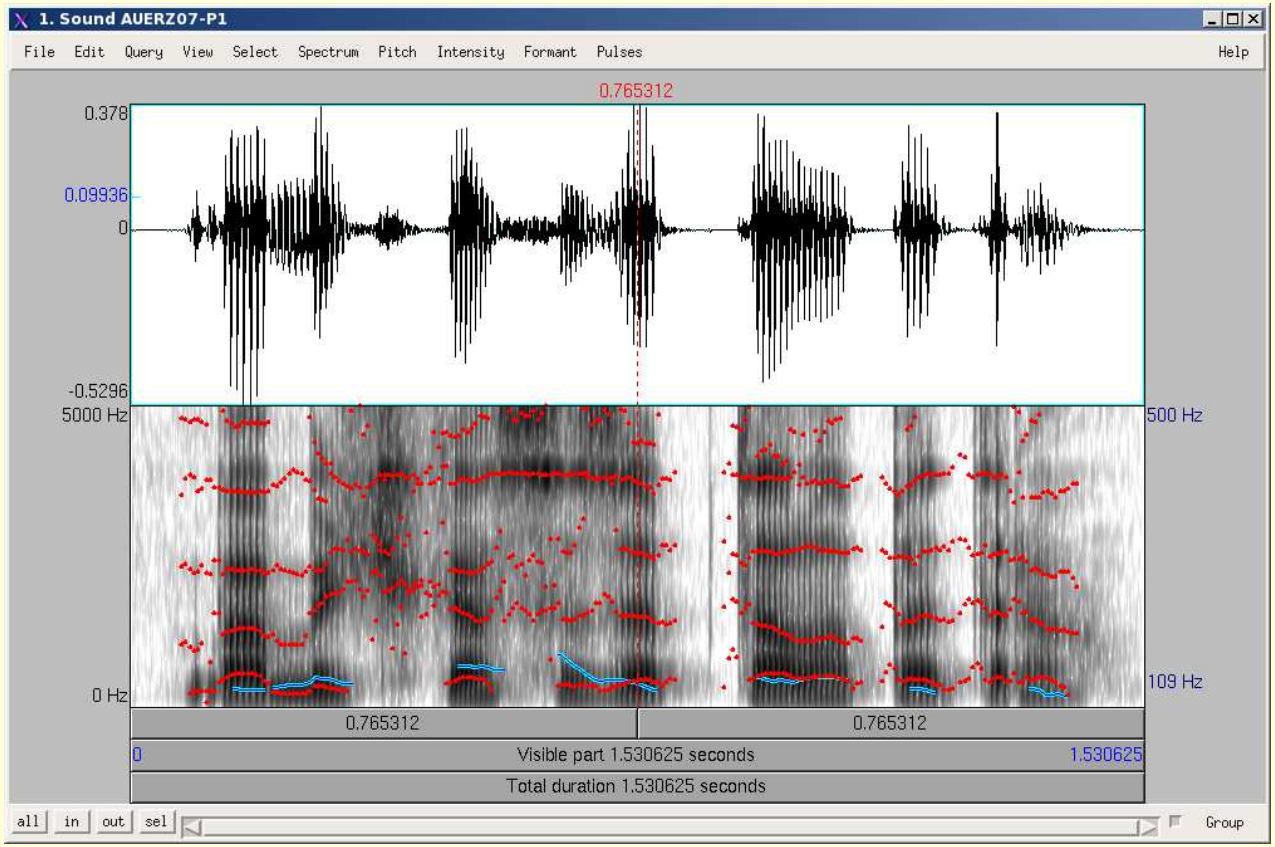
– **textově závislé příznaky** (klasifikace na bázi DTW či GMM, často expertní, ne zcela automatická)

Expertní metody při rozpoznávání mluvčího

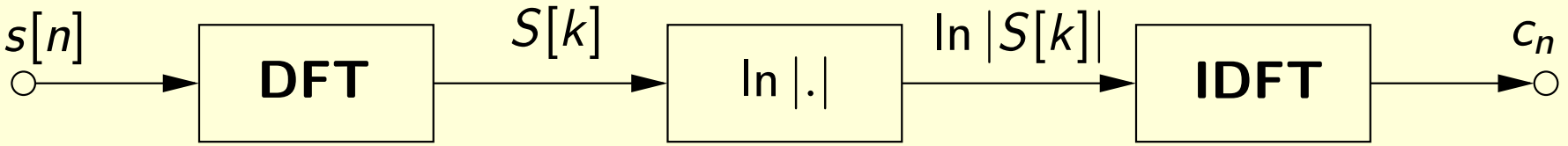
Forenzní lingvistika a fonetika - spektrografické metody

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

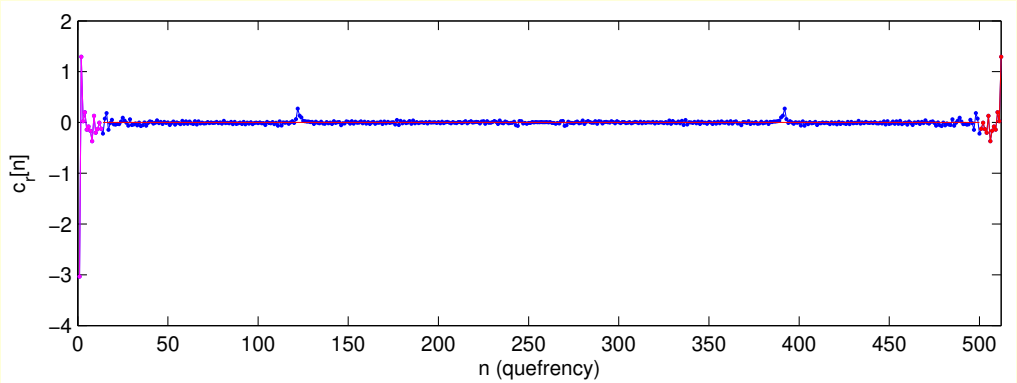
- možnost zobrazení diskutovaných hlasových charakteristik

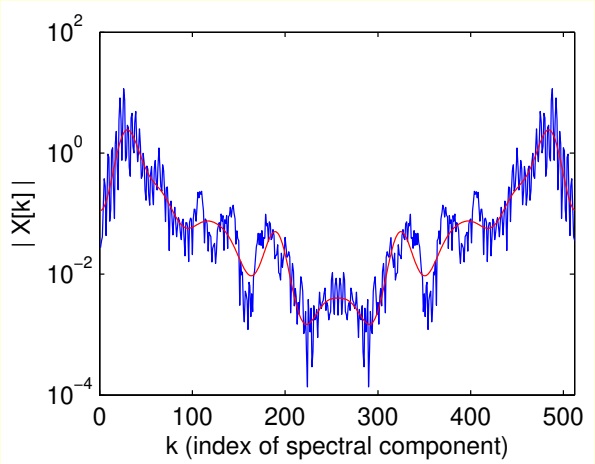


Kepstrum - definice na bázi DFT



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

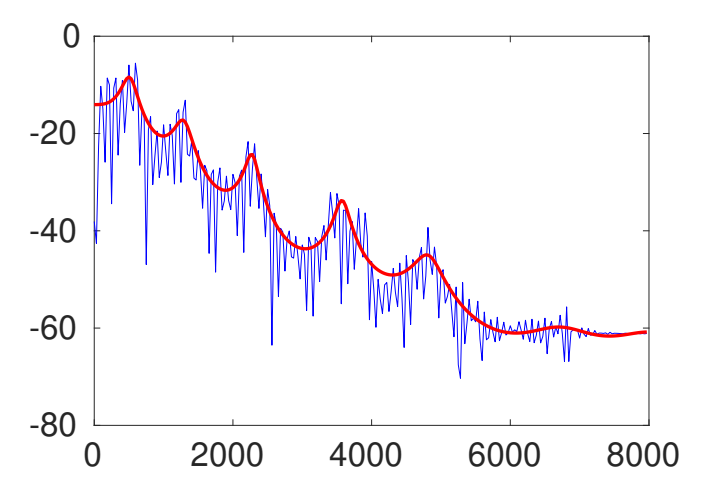




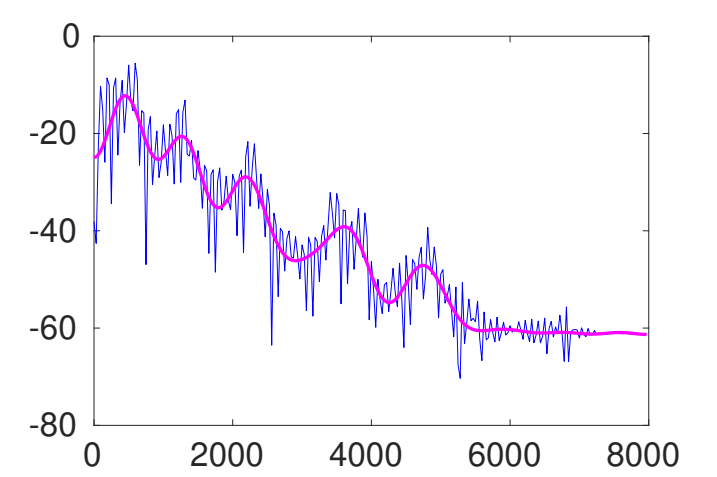
Názvosloví = slovní přesmyčky: **spektrum** vs. **kepstrum**, kvefrence vs. frekvence, liftrace vs. filtrace (modifikace kepstra)

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

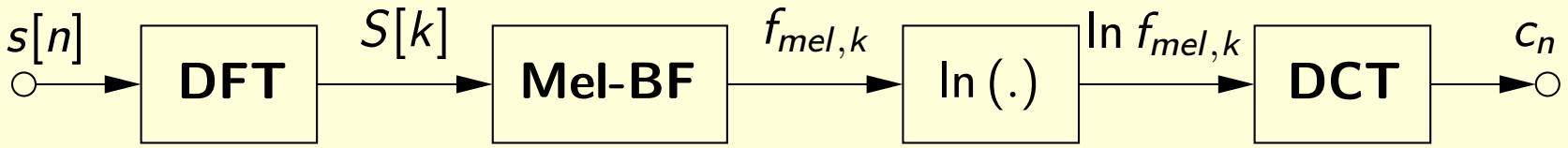
- **kepstrum** (amplit.spektrum) reprezentuje obecně informaci o **vokálním traktu**

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

  (**textově nezávislá** reprezentace)

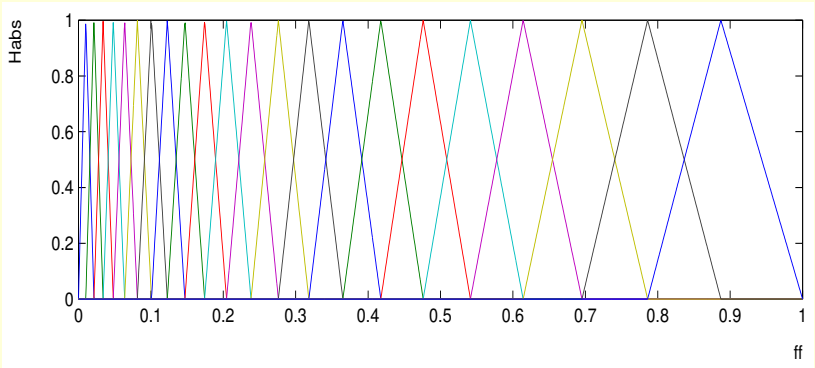
MFCC - Melovské kepstrální koeficienty

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

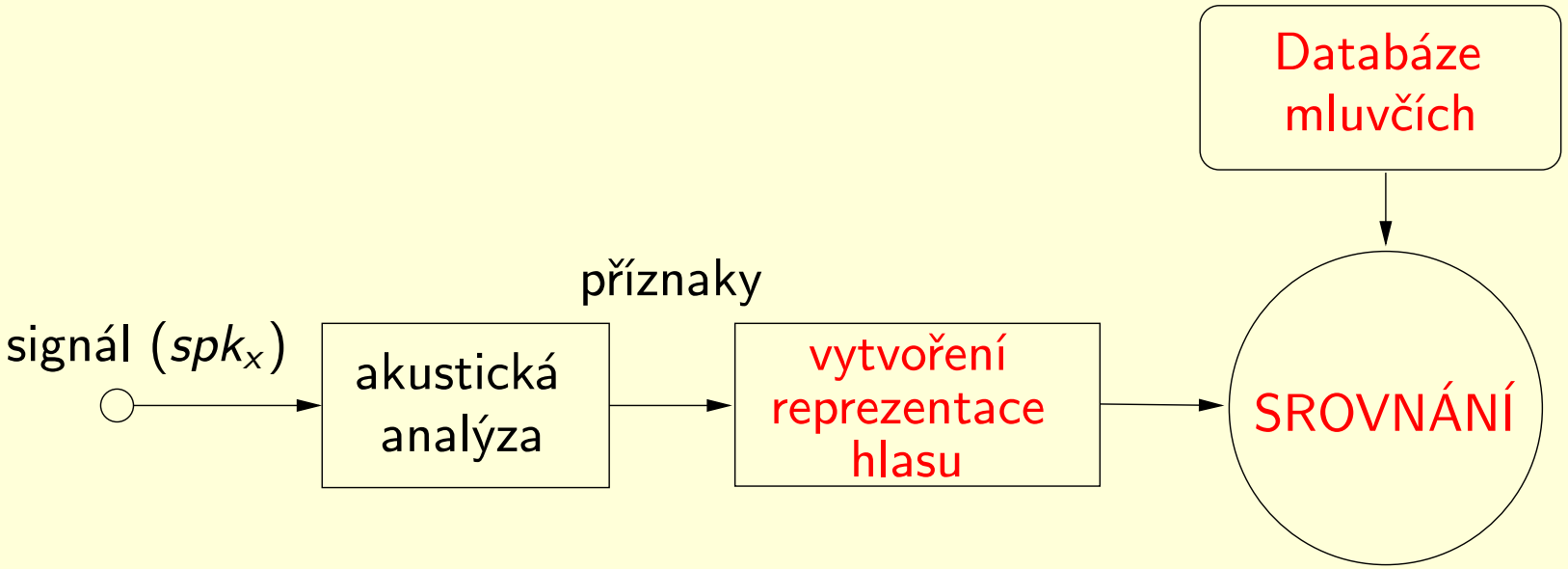
- *rozpoznánání řeči (ASR - Automated Speech Recognition)*
- rozpoznávání řečníka (SRE - *Speaker Recognition*) na bázi GMM (dekorelované příznaky)
- používané obvykle **textově nezávislou** hlasovou identifikaci

III. část

Úlohy automatického rozpoznávání řečníka

(textově nezávislé)

Reprezentace hlasu pro textově nezávislé SRE



Možné reprezentace **hlasu/řečníka/promluvy:**

- Statistický model (GMM)
- Embedding na bázi GMM (i-vektory)
- Embedding na bázi DNN (x-vektory)

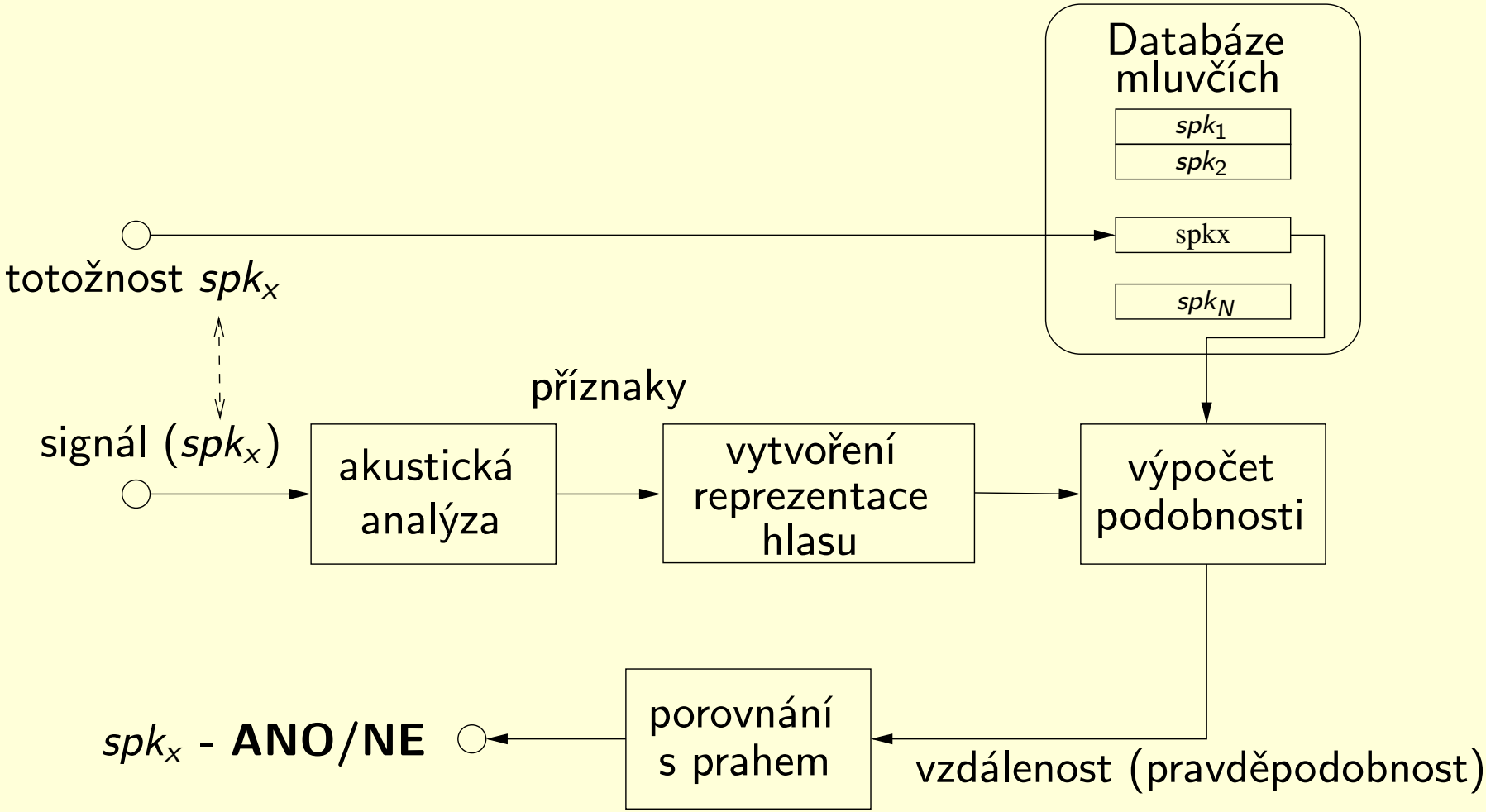
Úlohy rozpoznávání mluvčího

Základní úlohy automatického rozpoznávání mluvčího :

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

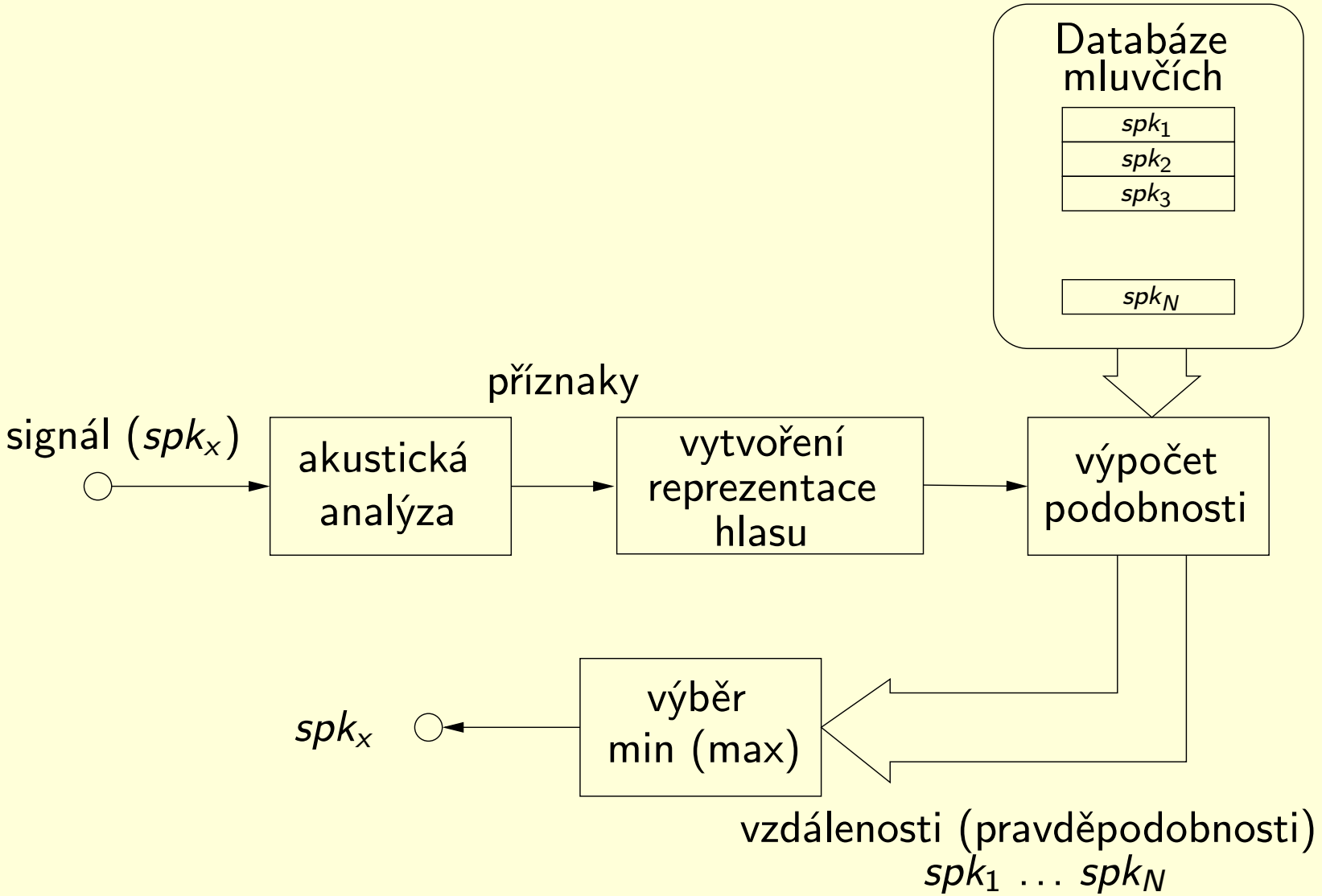
2. **Identifikace mluvčího**
   - **Identifikace v uzavřené množině**
     rozpoznání neznámého mluvčího z dané množiny mluvčích
   - **Identifikace v otevřené množině**
     rozpoznání neznámého mluvčího z neomezené množiny mluvčích $\longrightarrow$ identifikace & verifikace

Verifikace mluvčího



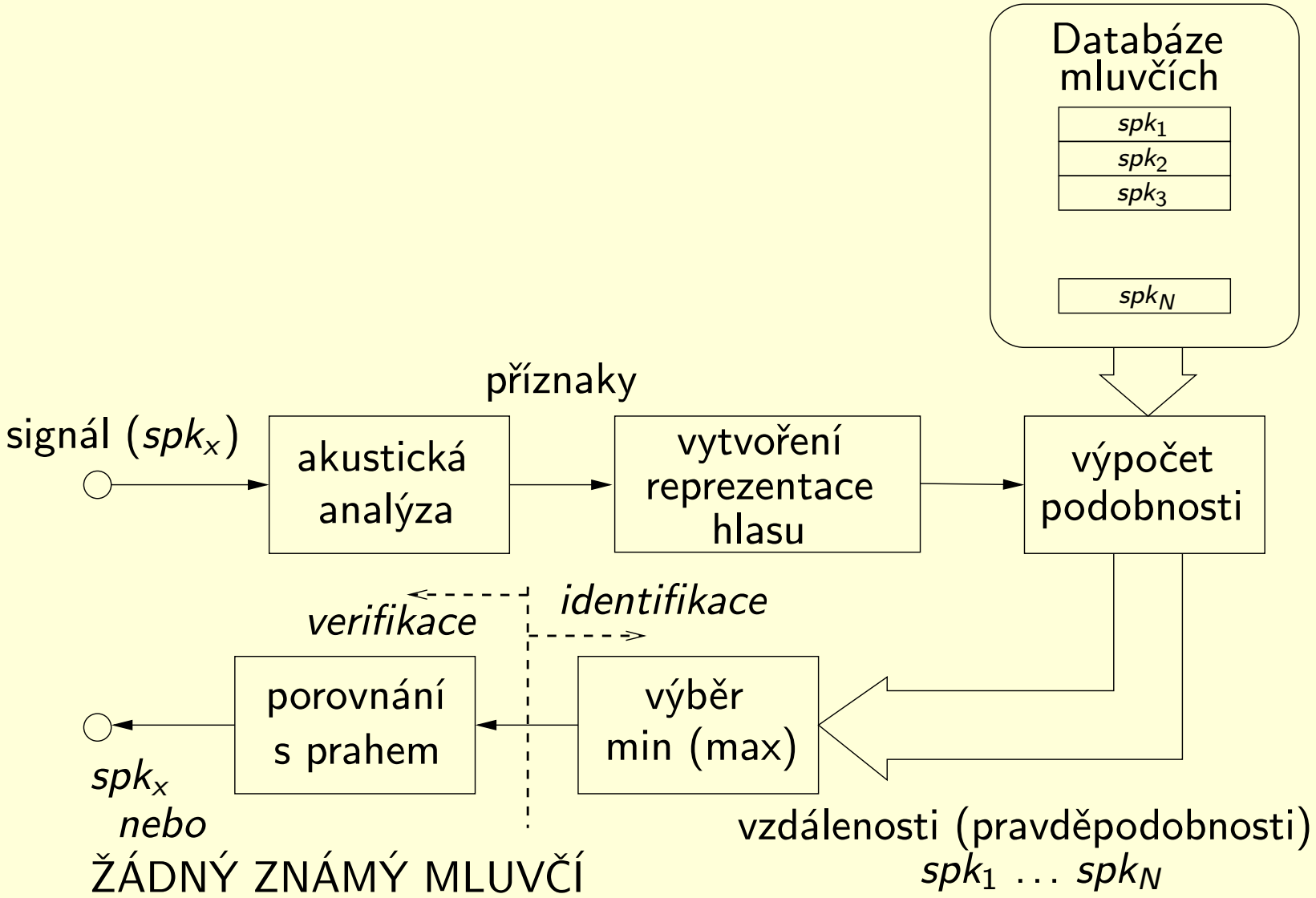
- ověření předpokládané totožnosti mluvčího
- **VÝSLEDEK** = **přijetí / odmítnutí** předpokl. totožnosti

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**

Reprezentace hlasu řečníka a klasifikační techniky

Reprezentace mluvčího na bázi vzorů

- *Kódová kniha používaných parametrů* :
  míra = střední vzdálenost příznaků od typických reprezentantů

Reprezentace mluvčího na bázi statistických modelů

- GMM modely : modelují rozložení příznaků pro daného řečníka
  míra = věrohodnost spočítaná z emitovaných pravděpodobností
- i-vektory : modelování prostoru středních hodnot GMM modelů

Klasifikace na bázi neuronových sítí

- přímá identifikace mluvčího
- nepřímé použití DNN (výpočet příznaků)
- první End-to-End systémy

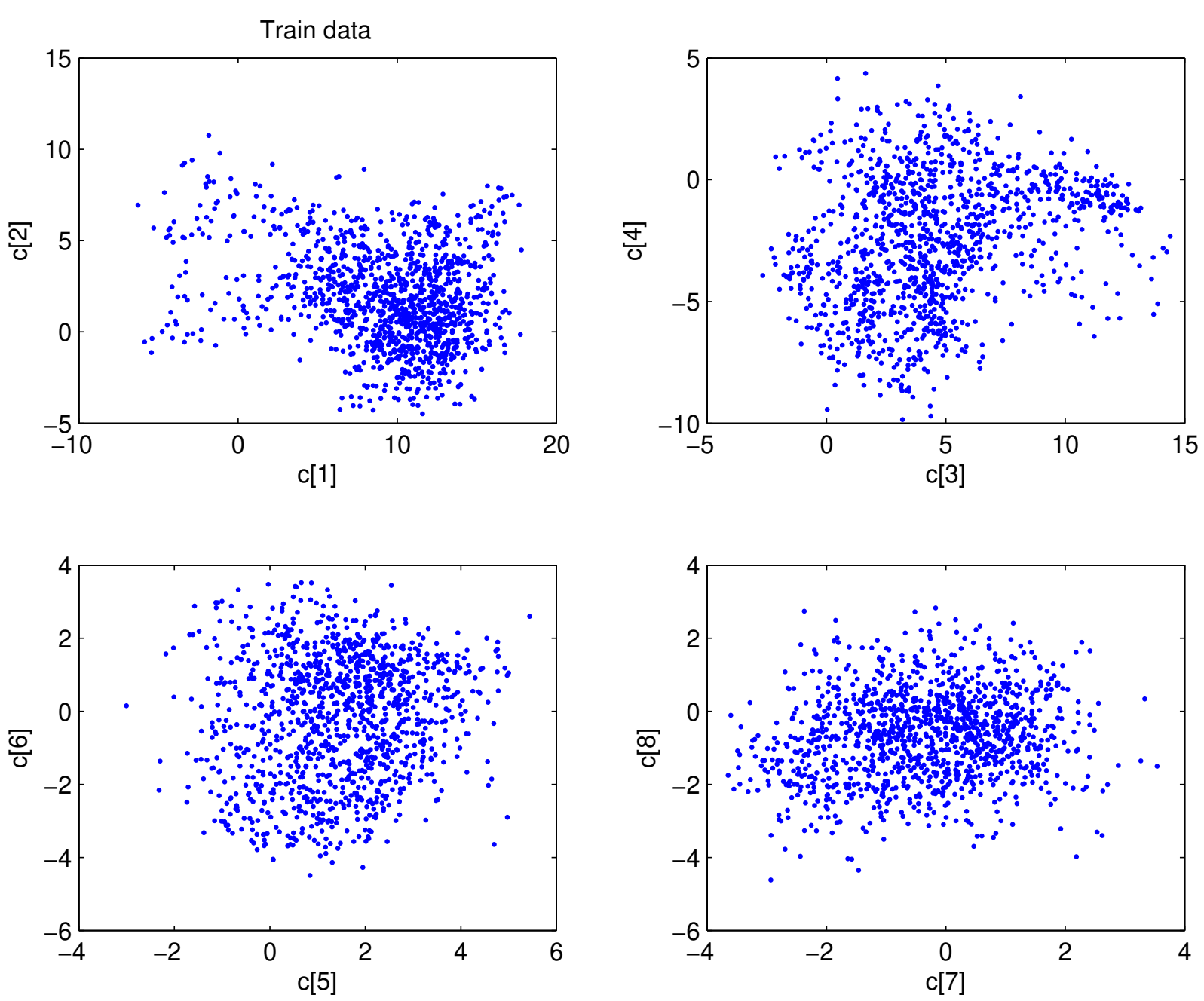
Rozpoznávání řečníka na bázi GMM

- příznaky - nejčastěji kepstrum (MFCC)
- sledují se rozdíly v rozložení kepstra pro různé mluvčí
- rozložení kepstra je popsáno **statistickým modelem na bázi GMM**

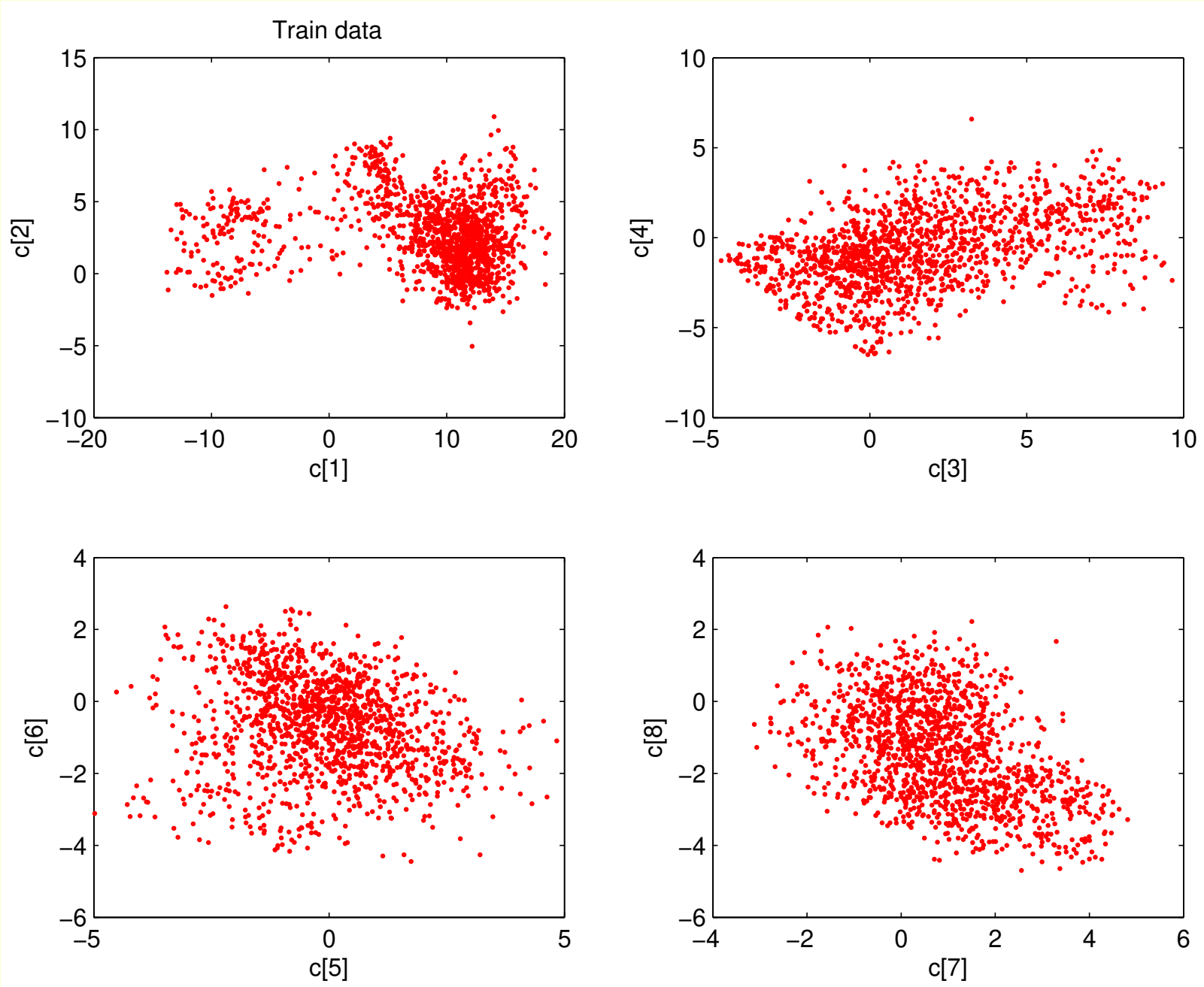
$$p(\mathbf{o}|\lambda^s) = \sum_{i=1}^{M_s} c_i^s \cdot \mathcal{N}(\mathbf{o}, \boldsymbol{\mu}_i^s, \mathbf{C}_i^s)$$

- $\mathcal{N}(\mathbf{o}, \boldsymbol{\mu}, \mathbf{C})$ ... N-rozměrná gaussovská funkce daná vektorem středních hodnot $\boldsymbol{\mu}$ a kovarianční maticí $\mathbf{C}$

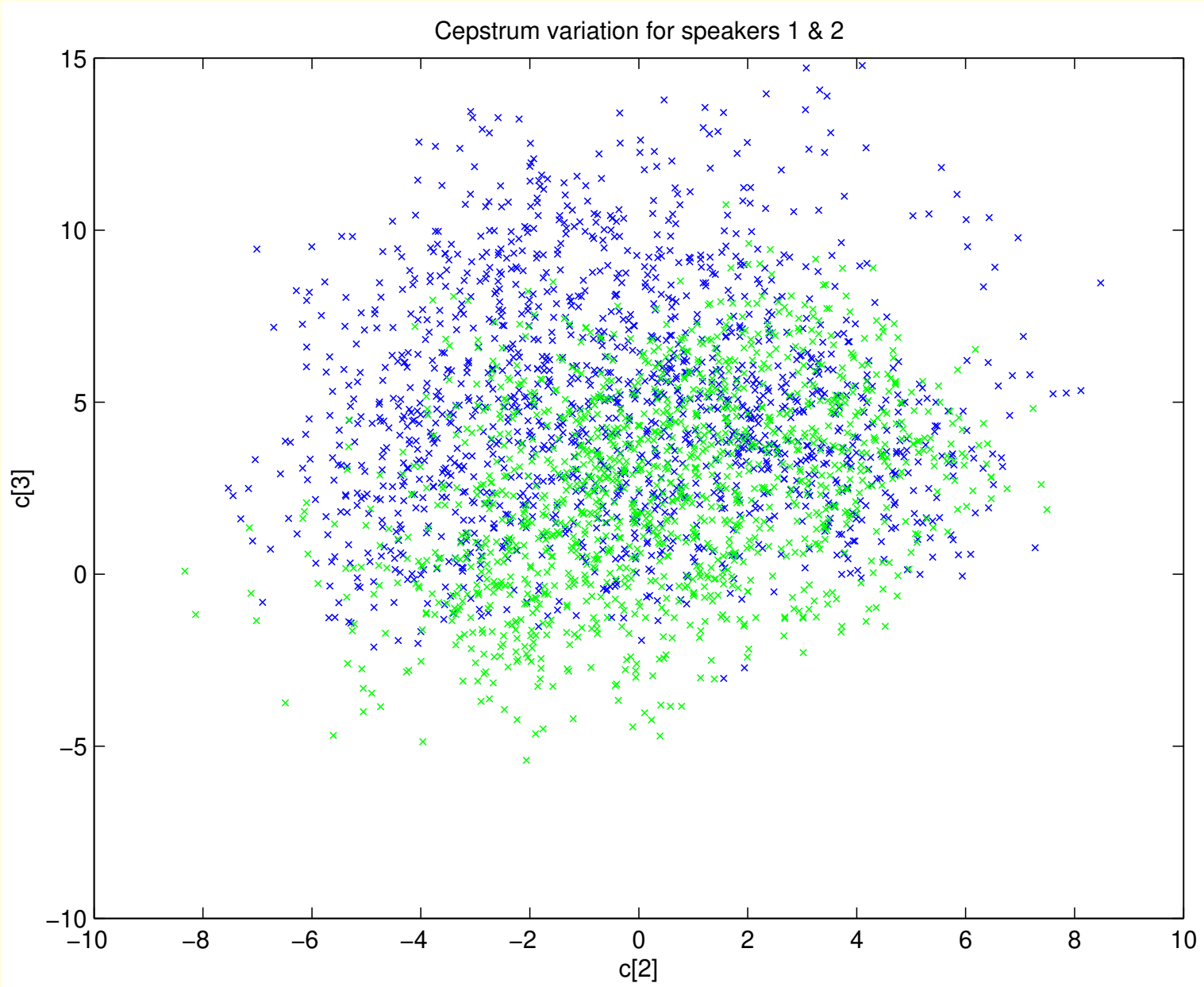
Rozložení prvků kódové knihy kepstra mluvčího A



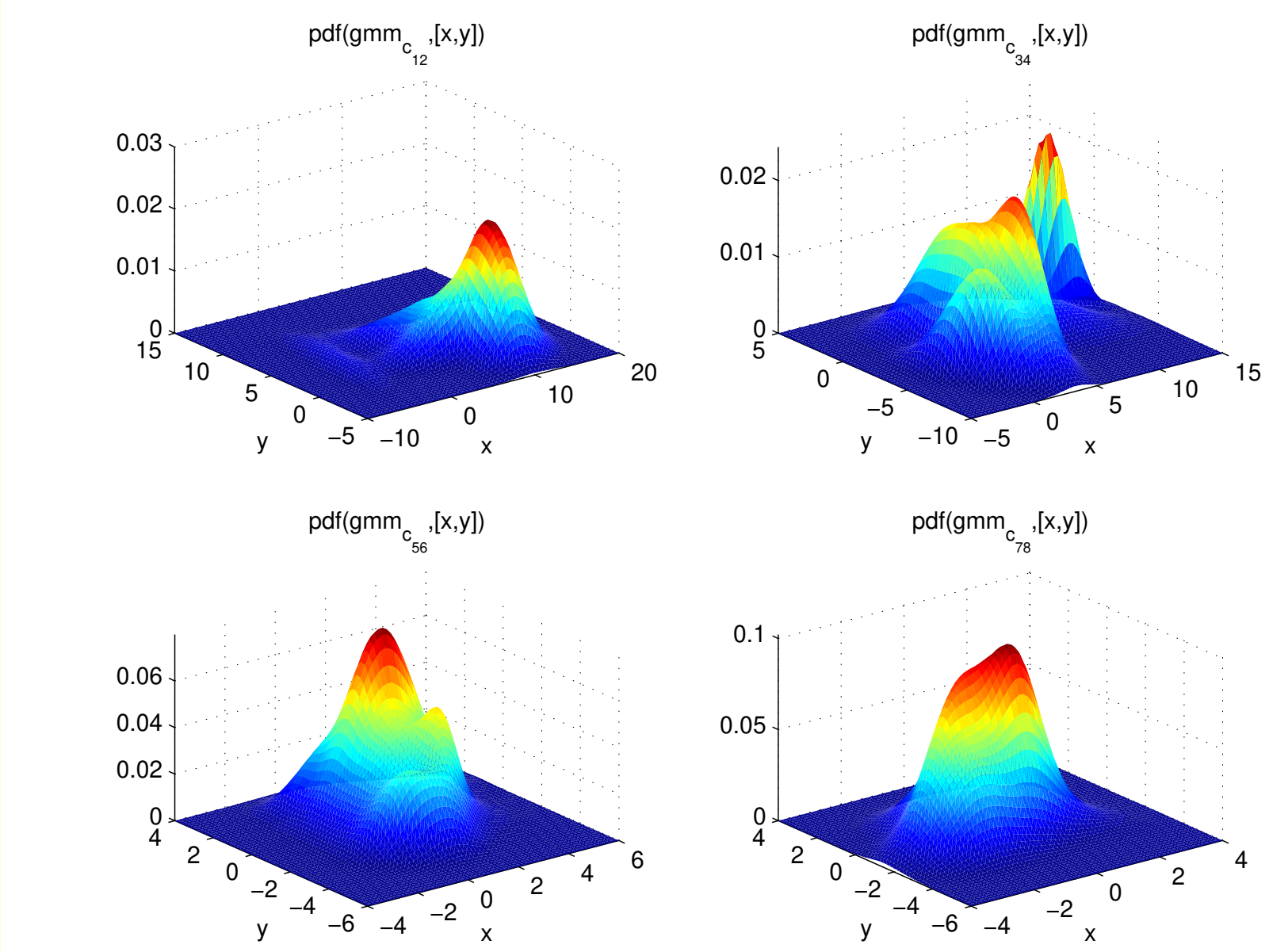
Rozložení prvků kódové knihy kepstra mluvčího B



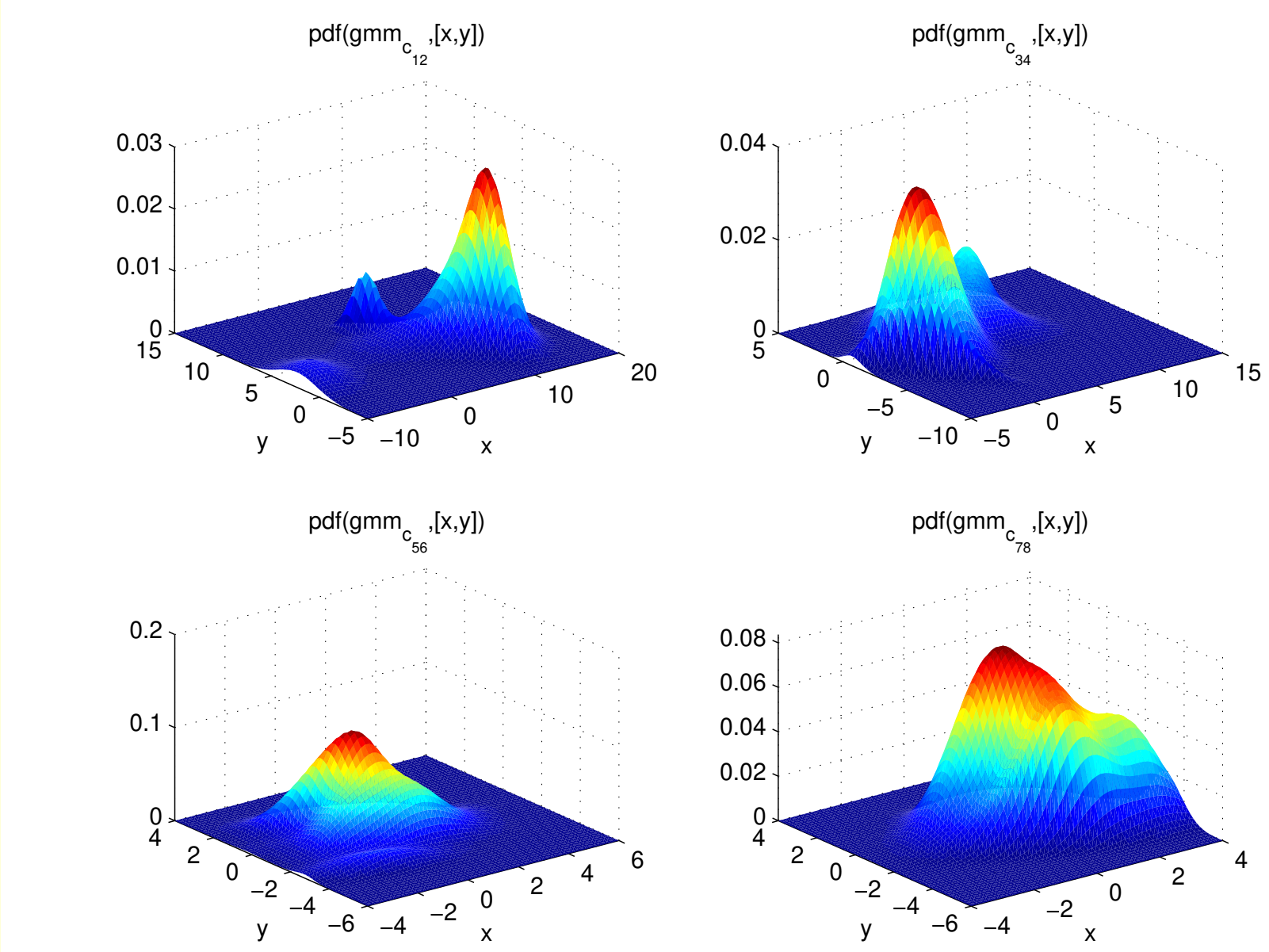
Rozložení kepstra ($c[2]$ vs. $c[3]$) - řečník 1 & 2



GMM model rozložení kepstra mluvčího A



GMM model rozložení kepstra mluvčího B



Klasifikace na bázi GMM

Klasifikační míra:
$\rightarrow$ **věrohodnost příznaku pro daný model** - $p(\mathbf{o_j}|\lambda^s)$

- hodnota věrohodnosti se počítá z celé promluvy

$$\mathbf{O} = (\mathbf{o_1}, \mathbf{o_2}, ..., \mathbf{o_n})$$

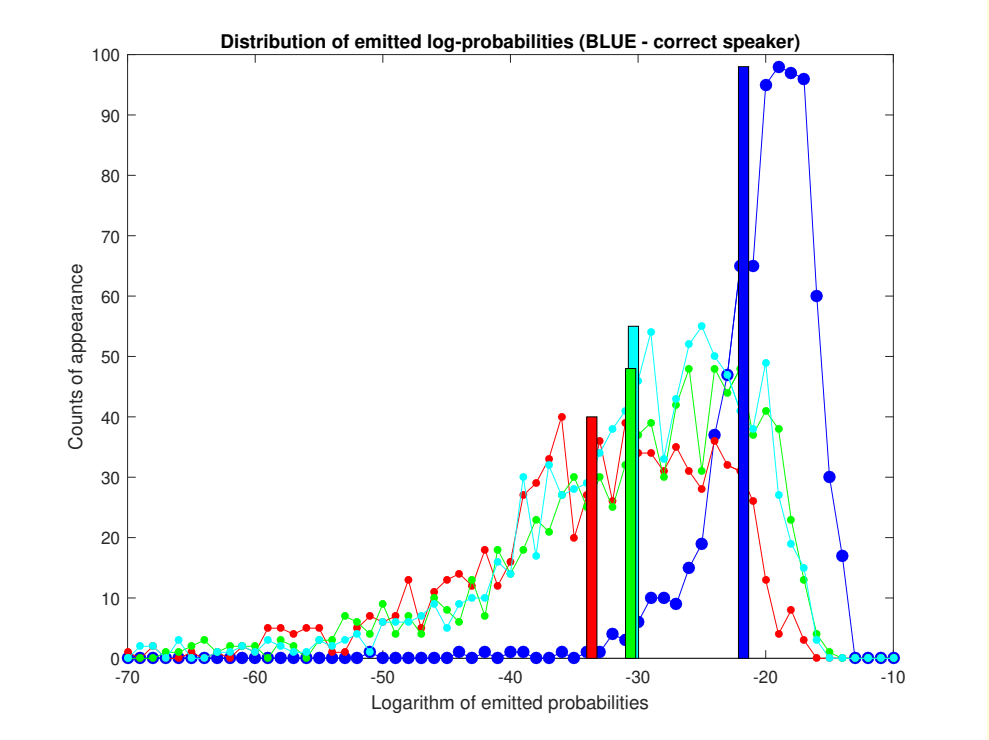
$$P(\mathbf{O}|\lambda^s) = \prod_{j=1}^{N} p(\mathbf{o_j}|\lambda^s)$$

- logaritmická věrohodnost - průměrování (sčítání) logaritmů emitovaných pravděpodobností pro všechny krátkodobé realizace (segmenty) a GMM model daného mluvčího
    (omezení možnosti podtečení)

$$\log P(\mathbf{O}|\lambda^s) = \sum_{j=1}^{N} \log p(\mathbf{o_j}|\lambda^s)$$

Je vhodné aplikovat detektor řečové aktivity!

Statistiky výsledků pro 4 řečníky a 1 GMM model



GMM model - zdroj: 12 promluv (12 x 5s), cca 2000 segmentů
- počet vážených směsí v GMM: 6

Identifikace - 20 promluv (20 x cca 1s), cca 1200 segmentů

Průměrné hodnoty logaritmické pravděpodobnosti

- více směsí modeluje lépe variabilitu příznaků pro daného řečníka

- typické počty směsí: 8-256 (model řečníka), *počty směsí zavisí na množství trénovacích dat*

Rozpoznávání mluvčího na bázi UBM-GMM

Problém s trénování GMM modelu pro jednotlivého mluvčího

málo dat $\rightarrow$ malá schopnost generalizace

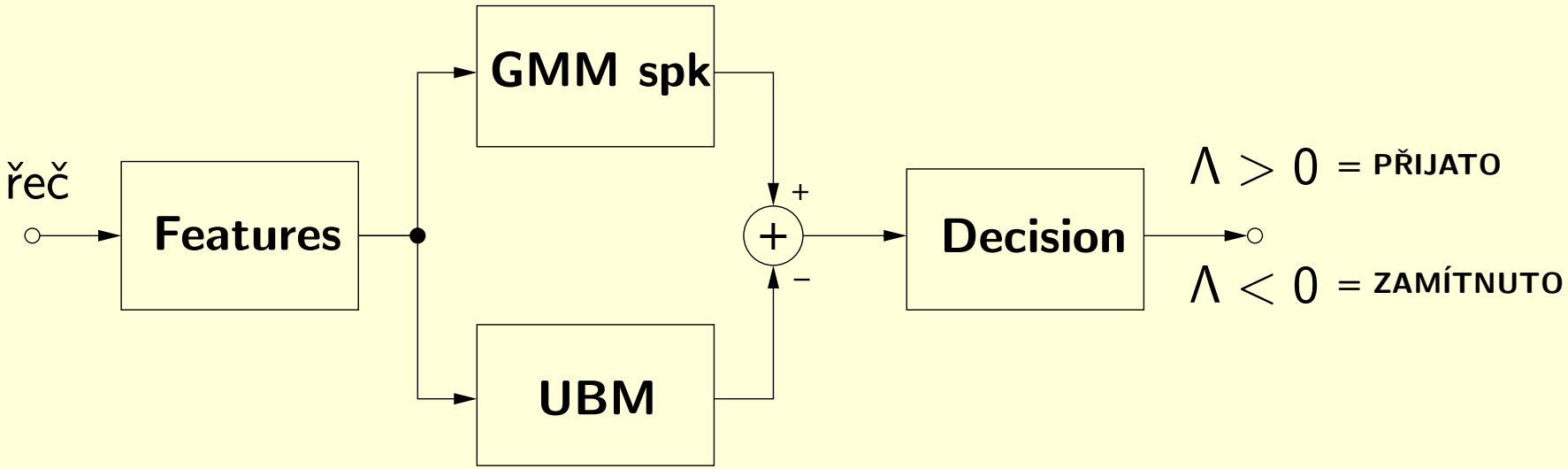
$\Downarrow$

UBM-GMM modelování

- UBM - Universal Background Model
  - generalizující model popisující společný prostor parametrů
  - vytvořený terénováním GMM pro velkou množinu mluvčích
- GMM model mluvčího - získáný adaptací UBM
  - nejčastěji MAP (Maximum Aposteriori Probability)
  - zápis mluvčího
    (neexistujícího v množině pro trénování UBM)
  - typicky velmi malé množství dat pro zápis
    (několik jednotlivých promluv, obvykle cca $3 \div 30s$)

Rozhodování UBM-GMM verifikace

$$\Lambda = log \frac{P(\mathbf{O}|\lambda^{spk})}{P(\mathbf{O}|\lambda^{UBM})}$$



UBM-GMM → základ pokročilejších systémů na bázi i-vektorů

Typické počty směsí:

- cca 512 (model řečníka)
- 512-2048 (univerzální model)

III. část

Textově nezávislá verifikace/identifikace na bázi i-vektorů

Definice a význam i-vektoru

GMM-UBM : adaptace UBM $\rightarrow$ GMM (pouze střední hodnoty)
Mluvčího charakterizují hodnoty vektoru středních hodnot
$\longrightarrow$ supervektor :

- lze použít i pro reprezentaci nahrávek (různé délky)
- vektor délky $C \cdot F$ všech středních hodnot
  (C počet složek GMM, F počet použitých příznaků)
- NEVÝHODA - velká dimenze supervektoru

i-vektor - $\mathbf{x}_{r,s}$ - dimenze $D_{ivec} < CF$

- model supervektoru $\mathbf{m}_{r,s}$ na bázi faktorové analýzy (JFA)

$$\mathbf{m}_{r,s} = \boldsymbol{\mu} + \mathbf{T}\mathbf{x}_{r,s}$$

- společná složka pro všechny řečníky - supervektor $\boldsymbol{\mu}$
- složka jednoho řečníka - $\mathbf{T}\mathbf{x}_{r,s}$
  (generovaná transformací z vektoru menší dimenze $\mathbf{x}_{r,s}$)
- $\mathbf{x}_{r,s}$ (i-vektor) - popisuje specfické charakteristiky řečníka
- $\mathbf{T}$ - transformační matice dimenze $CF \times D_{ivec}$

Klasifikace na bázi i-vektorů

- trénování UBM - společný supervektor $\boldsymbol{\mu}$ (obecný korpus)
- EM odhad matice $\mathbf{T}$ (ze stejných dat jako UBM)
- i-vector extractor : $\mathbf{x}_{r,s} = \mathbf{T}^{-1} \cdot (\boldsymbol{\mu} - \mathbf{m}_{r,s})$
  kde $\mathbf{m}_{r,s}$ je supervektor řečníka resp. promluvy
  (získaný z GMM na bázi MAP adaptace UBM)
- i-vector $\mathbf{x}_{r,s}$ = **reprezentace mluvčího/promluvy**

- klasifikace = srovnání dvou i-vektorů
  (SVM s jádrovou funkcí na bázi kosinové vzdálenosti)

$$score_{ivec} = \frac{\mathbf{x}_1^T \mathbf{x}_2}{||\mathbf{x}_1|| \cdot ||\mathbf{x}_2||}$$

- variabilita akustických podmínek není explicitně modelována
- možnosti potlačení variability akustických podmínek :
  - LDA (nalezení podprostoru s optimální rozlišitelností tříd)
  - WCCN (normalizace kovariance uvnitř tříd)
  - PLDA - $\mathbf{x}_{r,s} = \boldsymbol{\mu} + \mathbf{V}\mathbf{y}_s + \mathbf{U}\mathbf{w}_{r,s} + \boldsymbol{\epsilon}_{r,s}$
    (modelování variability akustických podmínek)

IV. část

Textově nezávislá verifikace/identifikace s hlubokými neuronovými sítěmi (DNN)

Systémy rozpoznávání řečníka s neuronovými sítěmi

ANN/DNN - Artificial Neural Networks/Deep Neural Networks

Použití v SRE: - přímé, tj. výpočet pravděpodobnosti (klasifikace)
- **nepřímé**, tj. výpočet příznaků/reprezentace řečníka (embeddings)

VÝHODY:

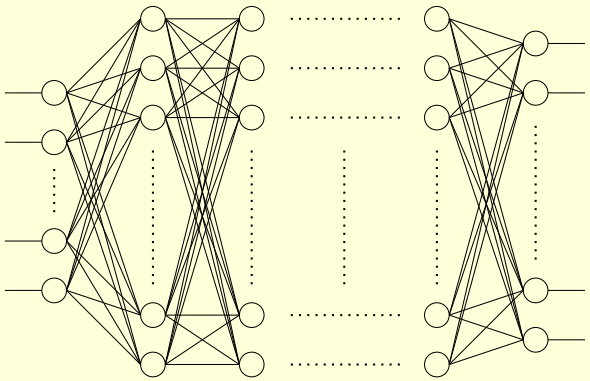
- možnost natrénování složitější funkce
- možné rozšíření příznakového vektoru
  (řetězení příznaků se širším kontextem)
- přesnější výsledky lze dosáhnout s DNN
  (vícevrstvé sítě s hlubokým učením - deep learning)
- speciální struktury sítí (RNN, TDNN, CNN, LSTM)

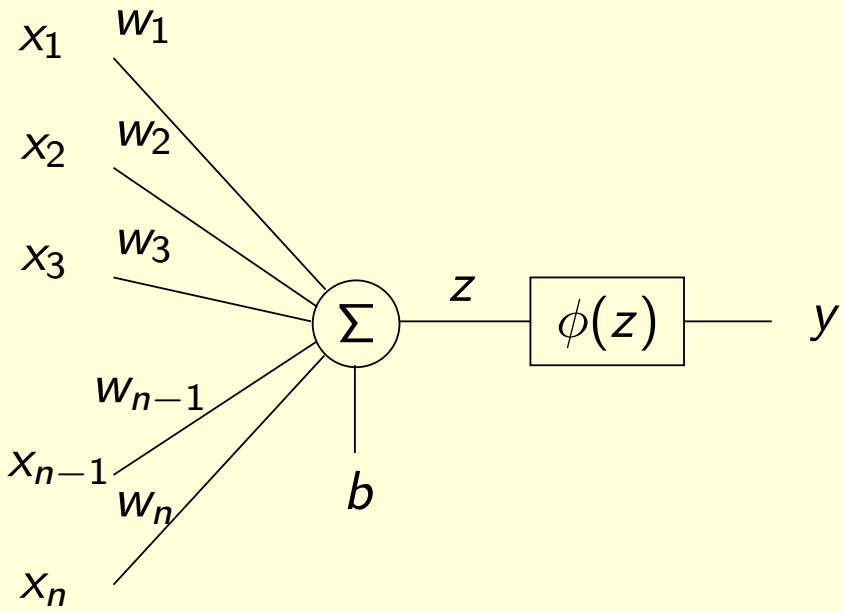
NEVÝHODY:

- obecně náročnější trénování (algoritmy hlubokého učení)
- potřeba většího množství trénovacích dat

(nastavení mnoha vnitřních parametrů sítě)

Základní dopředné neuronové sítě





Obecný výstup neuronu: $y = \phi\left(b + \sum_{i=1}^{m} w_i x_i\right) = \phi(z)$

Sigmoidní přenosová fce ve skryté vrstvě: $\phi(z) = \dfrac{1}{1+e^{-z}}$

ReLU, usměrněná lineární fce (Rectified Linear) $\phi(z) = \max(0, z)$

Softmax přenosová fce ve výstupní vrstvě (pravděpod. C tříd, součet 1):

$$\phi_k(z) = p_k = \frac{e^{z_k}}{\sum_{j=1}^{C} e^{z_k}}$$

Lineární přenosová fce ve výstupní vrstvě - regresní síť (obecné mapování)

Trénování ANN, DNN

Základní algoritmy trénování (učení) sítě:

- kritérium na bázi MSE (střední kvadr. chyba) - regresní síť
- kritérium na bázi CE (vzájemné entropie) - klasifikační síť
- algoritmus zpětného šíření chyby (gradient kritéria)

- dávkový odhad gradientu pro danou trénovací sadu
- gradientní stochastický algoritmus
  (odhad gradientu s každým vzorkem)
- "minibatch training" (odhad gradientu s menším souborem náhodně vybraných dat)

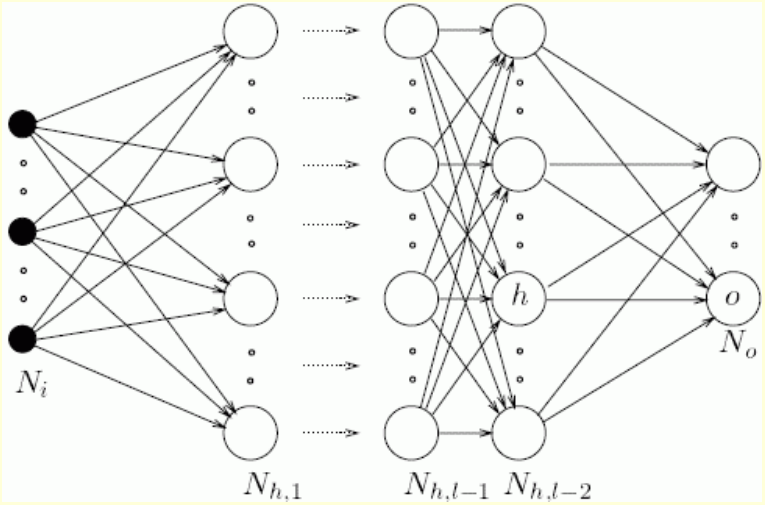
Inicializace sítě před trénování:

- náhodná - OK pro 3-vrstvé sítě, problém pro DNN
- předtrénování pro DNN
  - RBM (Restricted Boltzmann Machines)
  - DPT - diskriminativní předtrénování

Klasifikace na bázi ANN/DNN

Přímá klasifikace pomocí DNN (výpočet pravděpodobnosti)

- DNN ve funkci odhadu aposteriorní pravděpodobnosti řečníka

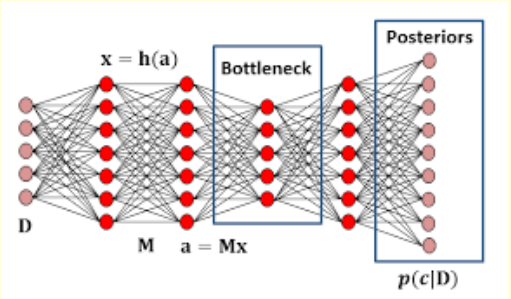


VSTUP: BF, kepstrum (MFCC), možný kontext několik oken

SKRYTÉ VRSTVY: 4-10

VÝSTUP: Softmax (aposteriors)

VARIANTA - DNN síť s bottleneck vrstvou (zúžení = komprese)

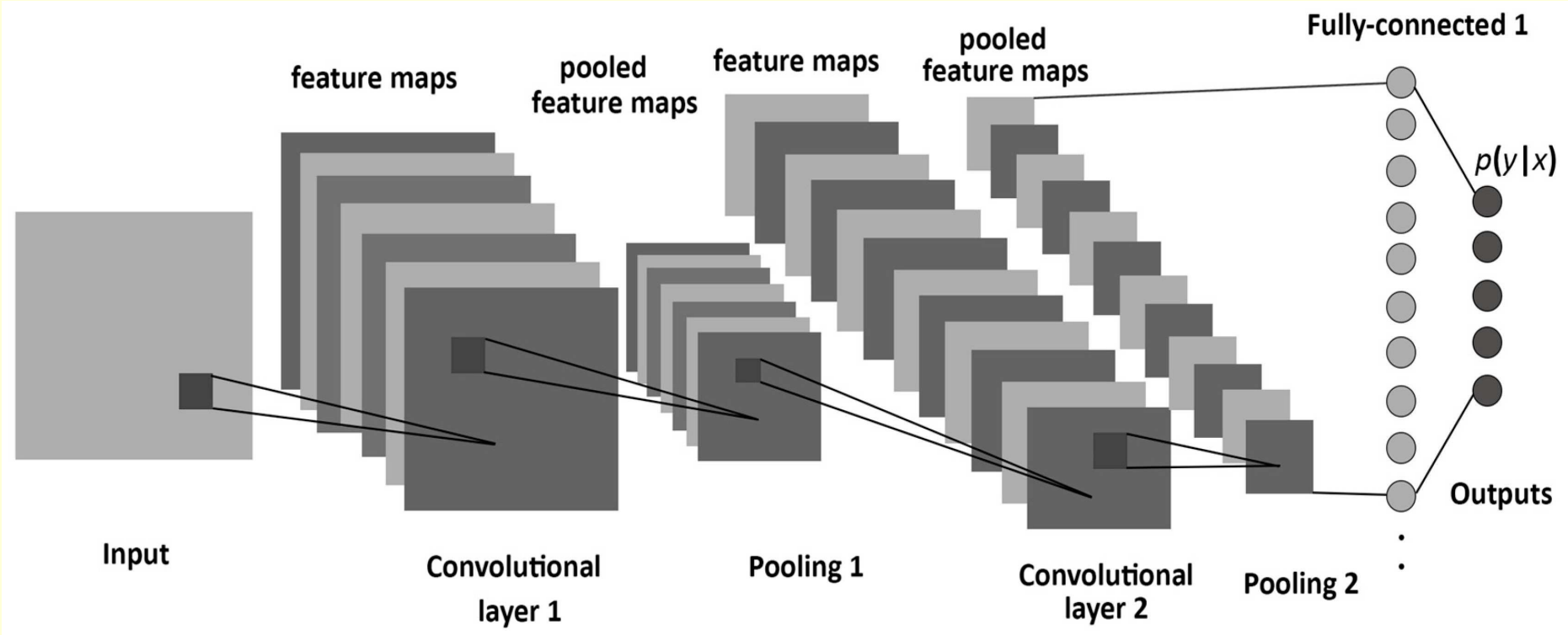


Nepřímé použití DNN: bottleneck výstup + MFCC (LDA, PCA)

→ příznaky pro i-vektorový systém

Přímá klasifikace na bázi CNN (Convolution Networks)

Principiální schéma CNN:

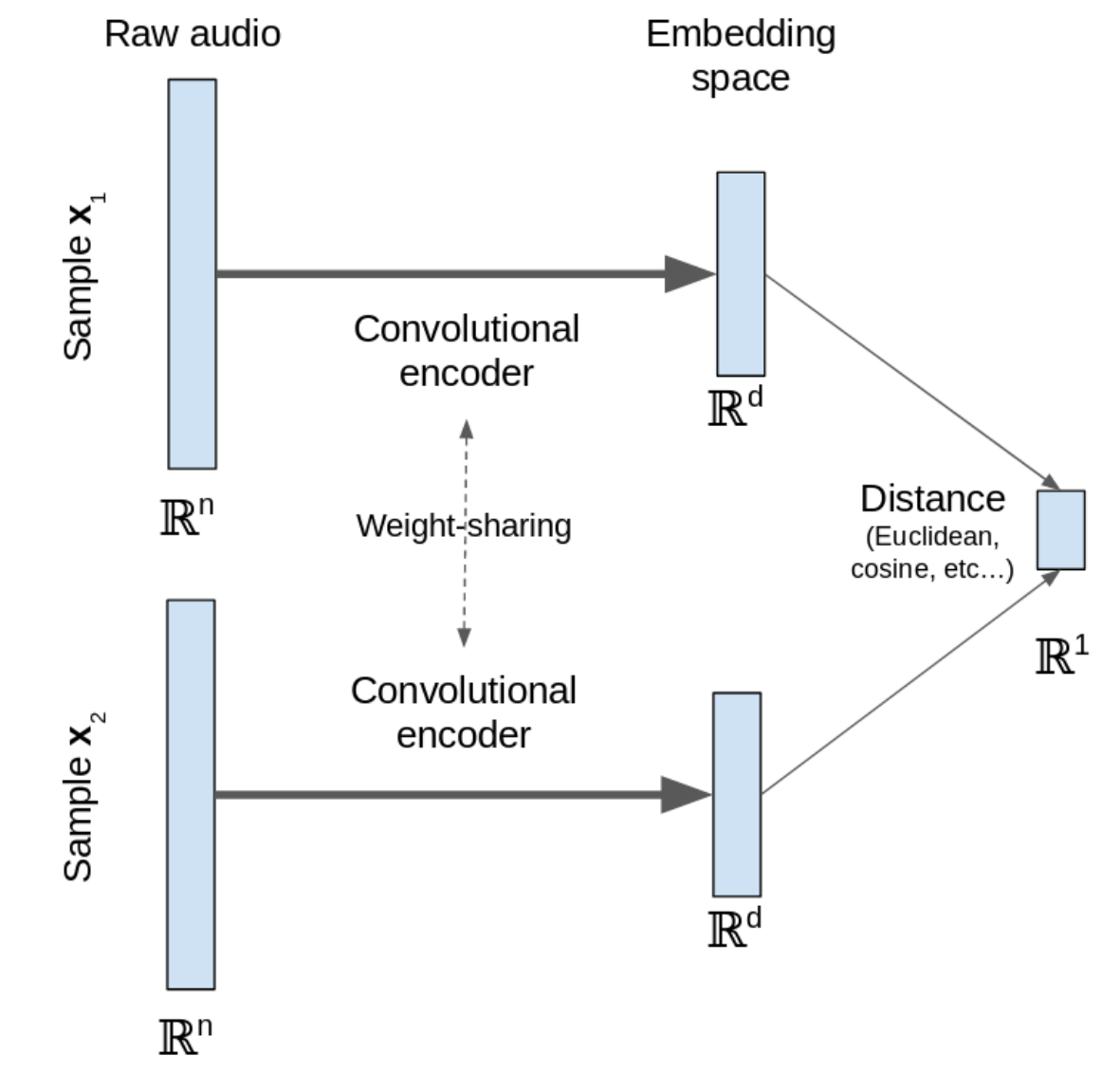


Nejčastější aplikace CNN ve zpracování obrázků

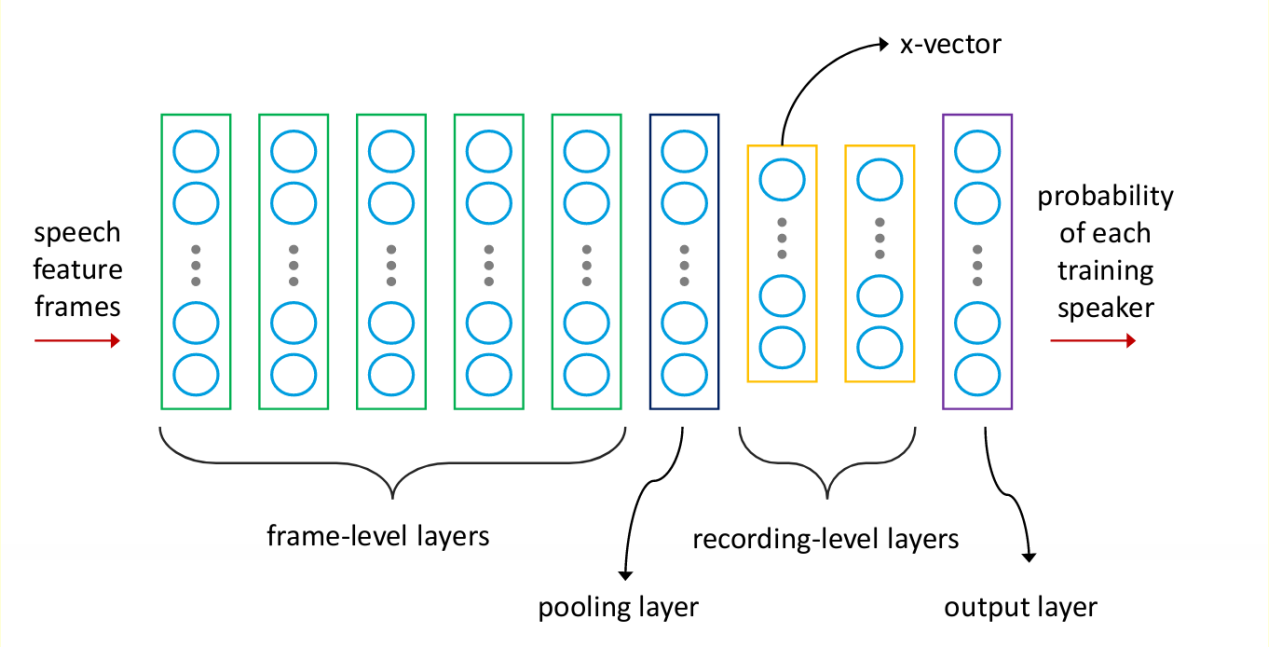
Aplikace pro SRE: vstupem je spektrogram (obrázek) či signál

$\rightarrow$ End-to-End Recognition (klasifikace bez výpočtu příznaků)

CNN Siamese Speaker Verification - nepřímá klasifikace



Aktuální DNN standard: systém na bázi x-vektorů



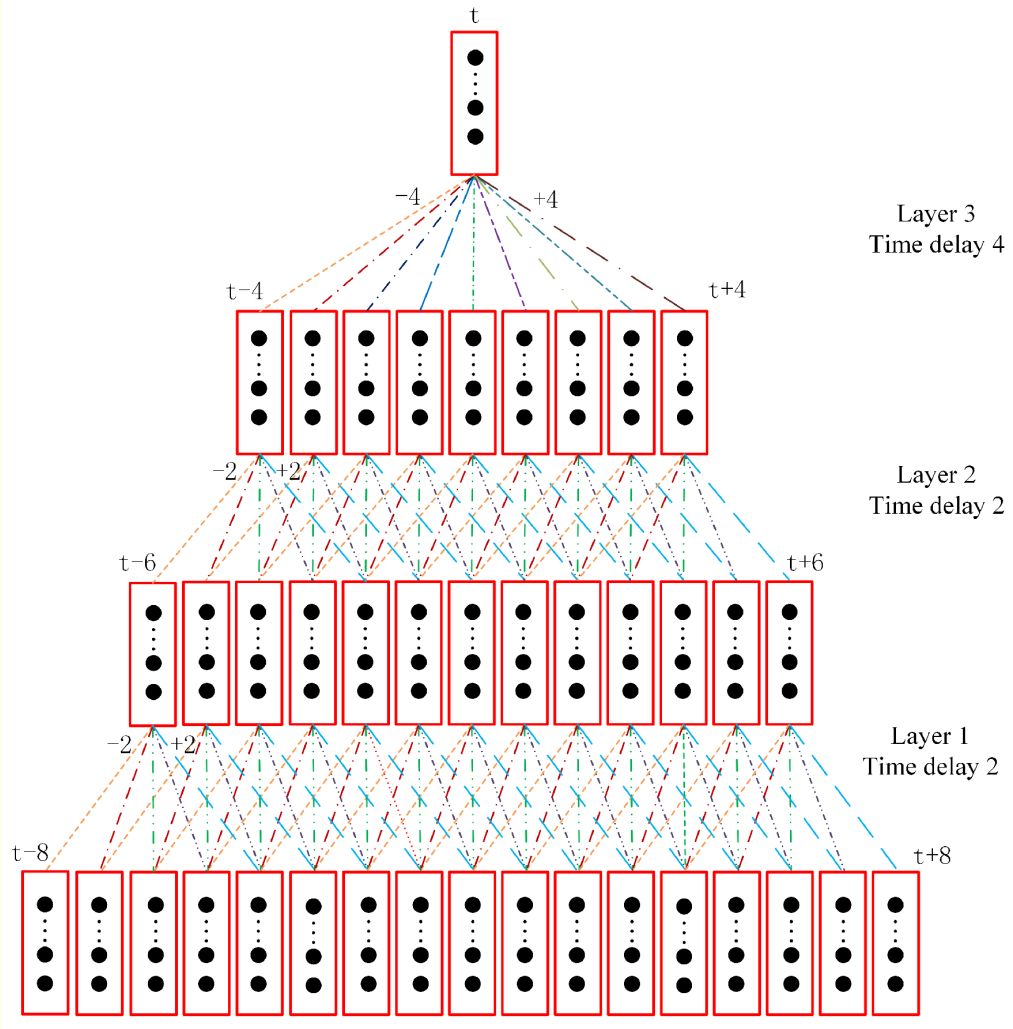
- vstupní příznaky jsou zpracovány v 5 TDNN vrstvách, zvyšující se zpoždění zahrnuje potřebnou kontextovou informaci (Δ příznaky nejsou používány)
- 6. poolingová vrstva počítá střední hodnoty a standardní odchylky výstupu 5. vrstvy přes všechny segmenty nahrávky
- dopředné (bottleneck) vrstvy 7 a 8 zahrnují reprezentaci mluvčího-nahrávky $\rightarrow$ x-vektor (příznaky pro SRE)
- 9. výstupní softmax vrstva realizuje identifikaci řečníka (využíváno v trénovací fázi)

TDNN - Time-Delay Neural Networks

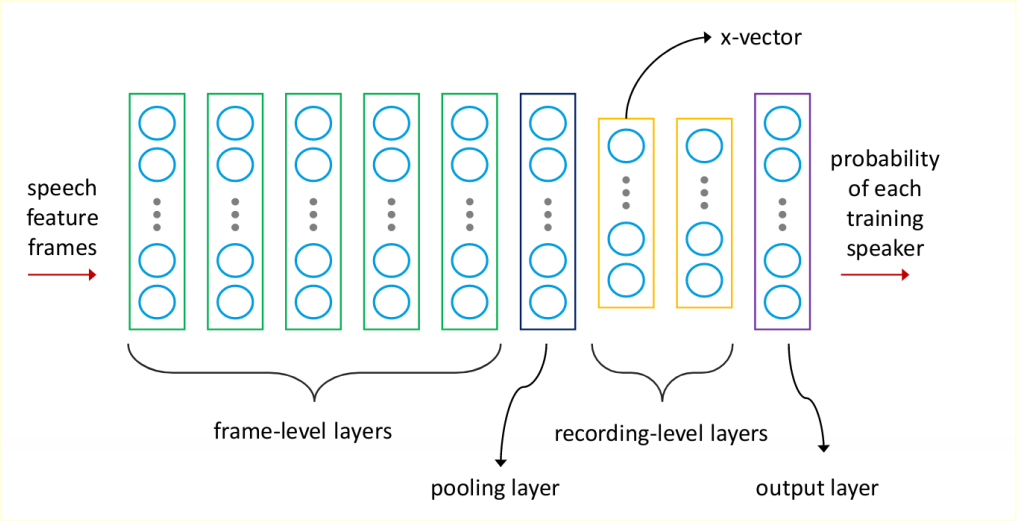
TDNN vrstvy - zahrnují kontextovou informaci

tj. výstupy z předchozí vrstvy pro více časů (řetězení)

- používá se plný či částečný kontext
- pro více TDNN vrstev $\rightarrow$ zvyšující se kontext



Aktuální DNN standard: systém na bázi x-vektorů



Vrstva	Kontext vrstvy	Celkový kontext	Vstup x výstup
frame1	t-2 ÷ t+2	5	120 x 512
frame2	t-2, t, t+ 2	9	1536 x 512
frame3	t-3, t, t+ 3	15	1536 x 512
frame4	t	15	512 x 512
frame5	t	15	512 x 1500
stats pooling	0 ÷ T	T	1500T x 3000
segment6	0	T	3000x512
segment7	0	T	512x512
softmax	0	T	512xN

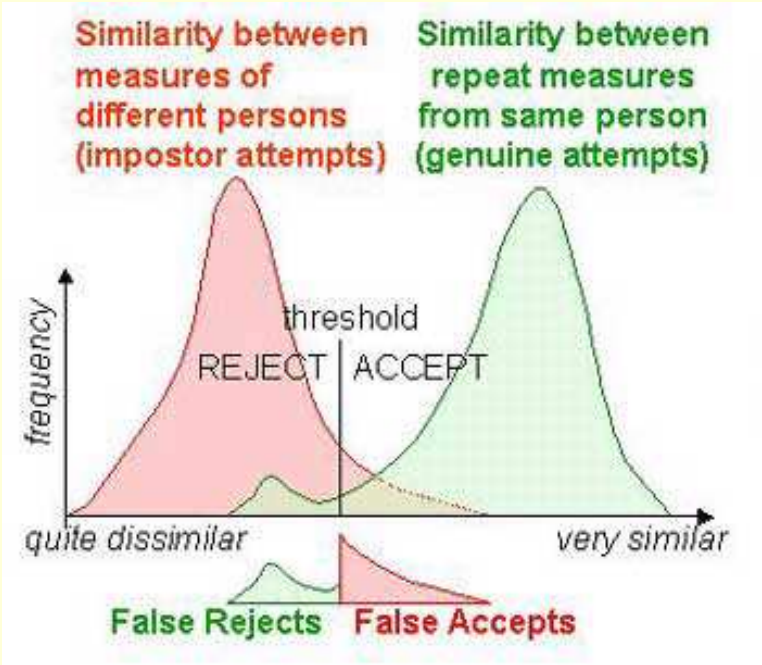
vstup 24 pásem melovské BF, pooling přes počet segmentů T, N řečníků

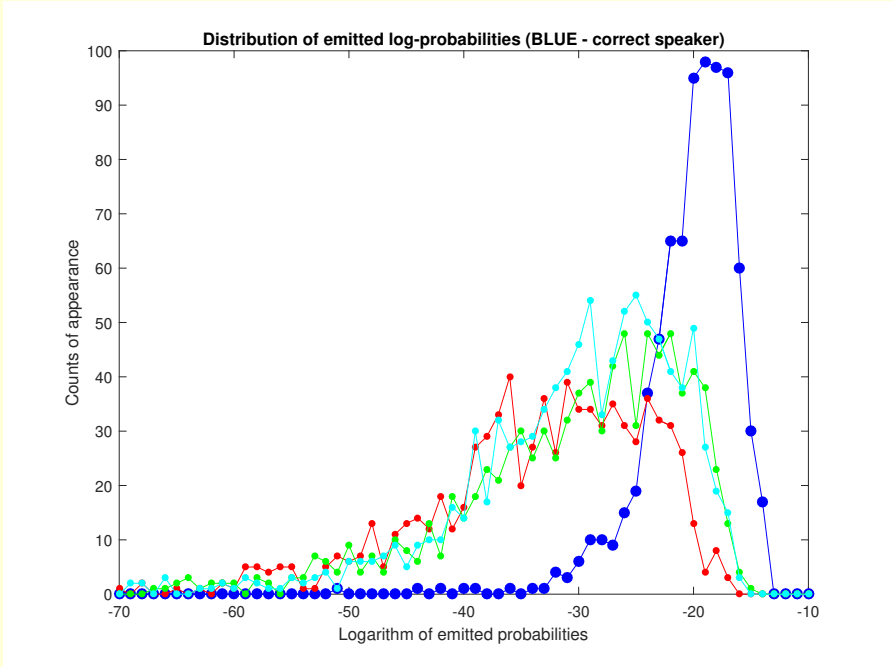
Více detailů viz: Snyder et al: X-Vectors: Robust DNN Embeddings for Speaker Recognition. In ICASSP 2018

V. část

Příklady systémů rozpoznávání řečníka

Hodnotící kritéria při verifikaci mluvčího - Míra stejné chyby



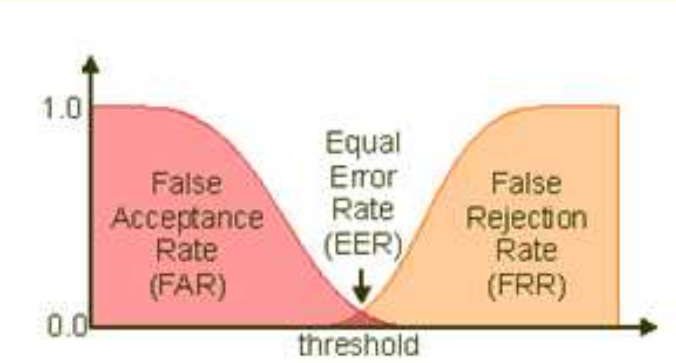


TA - True acceptance

FA - False acceptance: $R_{FA} = \frac{N_{FA}}{N_{podv}}$

TR - True rejection

FR - False rejection $R_{FR} = \frac{N_{FR}}{N_{spr\,ef}}$



EER - Equal Error Rate

Míra stejné chyby :

$$EER = R_{FR}(P_{thr}) = R_{FA}(P_{thr})$$

Standardy na bázi UBM-GMM (konec 90. let - cca 2012)

NIST 2010 - Speaker verification evaluations.

- výsledky verifikace pro rozdílné evaluační podmínky
- GMM-UBM systémy (UBM - Universal Background Model)
- EER - Equal Error Rate

	mic-mic	mic-mic2	mic-tel	tel-tel
Systém 1 - muži	8,39	17,29	16,24	15,68
Systém 1 - ženy	13,5	23,47	18,42	17,18
Systém 1 - AVG	10,94	20,38	17,54	16,52
Systém 2	6,00	8,64	5,32	5,11

Systém 1 - 8kHz, 25/10 ms, preemfáze, 16 MFCC ($+\Delta$, $+\Delta\Delta$), log energie, energetický VAD, normalizace příznakových vektorů, 512 směsí

Systém 2 - 8kHz, 25/10 ms, 19 MFCC & $c[0]$ ($+\Delta$), detektor řeči na bázi automatického přepisu (rozpoznávání), normalizace příznakových vektorů, adaptace akustických modelů, 512 směsí

Systémy na bázi i-vektorů - cca 2010-2020

Interpseech 2016:

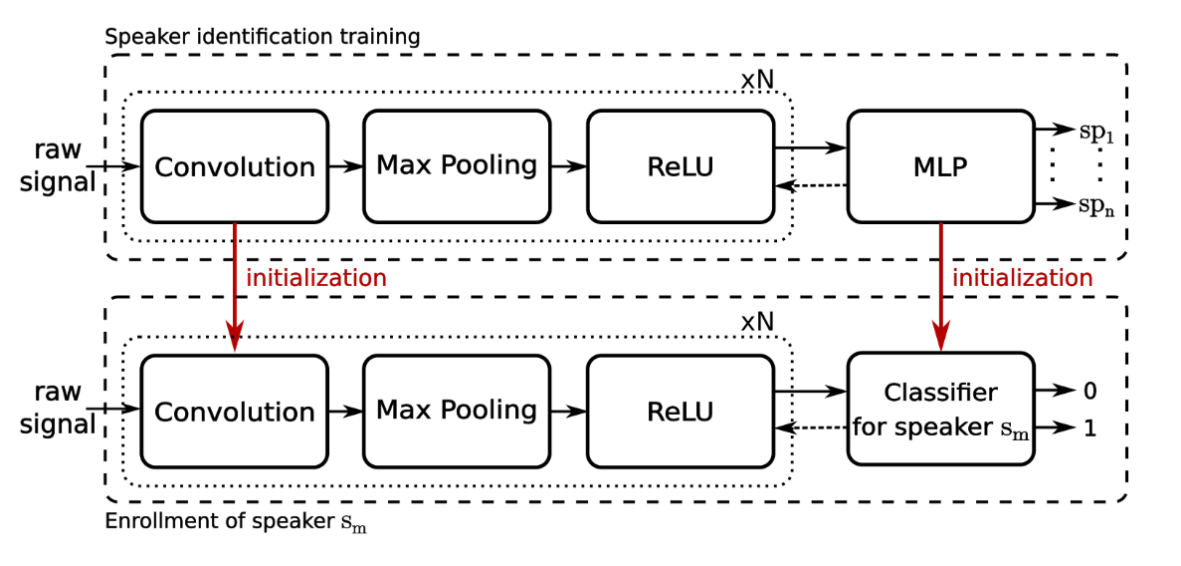
The IBM Speaker Recognition System
EER 2.11% - GMM-UBM (i-vector) - MFCC - LDA
EER 1.49% - GMM-UBM (i-vector) - MFCC - NDA
(NDA - Nearest-neighbour discriminant analysis)
EER 0.59% - DNN-fMLLR-NDA (English)

SITW - Speakers In The Wild - core-core
Speakers In The Wild Database (for speaker recognition)
- 299 speakers, 8 different sessions per speaker
- mismatch of acoustic conditions
- "core" conditions (data from one person of interest)
- training 6180 seconds
- test 6-180 s of speech per file

Brno University of Technology : EER = 5.85%
Quensland University of Technology, Australia : EER = 8.69%

Příklady systémů s DNN - Interspeech 2018

Muckenhirn, Magimai-Doss, Marcel: On Learning Vocal Tract System Related Speaker Discriminative Information from Raw Signal Using CNNs



EER 3.05% - GMM-UBM (standard baseline approach)
EER 2.40% - ISV (inter-session variability)
EER 2.82/5.87% - i-vector, cosine distance/PLDA
EER 5.00% - JFA (Joint Factor Analysis)
EER 0.80 / 1.15% - CNN (kW1=300 / kW1=30)
EER 0.75% - Fusion of 2 CNN systems (average score)

SRE na bázi x-vektorů - 2018-současnost

Hossein Zeinali, Kong Aik Lee, Jahangir Alam, Lukas Burget: **SdSV** Challenge 2020: Large-Scale Evaluation of **Short-duration Speaker Verification** (**Interspeech 2020:**)

- velmi krátké promluvy k verifikaci
- významná závislost na fonetickém kontextu (obsahu) (proto textově závislé i textově nezávislé úlohy)
- většina systémů na bázi x-vektorů

1. Textově závislá verifikace (*5 nejlepších týmů*)
   - promluvy pro zápis - avg 7.6s, testovací promluvy - avg 2.6s
   T56: EER 1.45 % , T14: EER 1.45 %, T10: EER 1.58 % ,
   T08: EER 1.62 %, T34: EER 2.09 %, T26: EER 2.10 %
2. Text independent verification (*6 nejlepších týmů*)
   - promluvy pro zápis - náhodně 4-180s (avg 49),
   - testovací promluvy - avg 2.6s
   T37: EER 1.45 % , T35: EER 1.51 %, T41: EER 1.77 %,
   T64: EER 1.84 % , T05: EER 2.00 %, T10: EER 2.32 %

Děkuji vám za pozornost !