A6M33BIO - Biometrie

Biometrické metody založené na rozpoznávání hlasu II

Doc. Ing. Petr Pollák, CSc.

13. prosince 2018 - 16:13

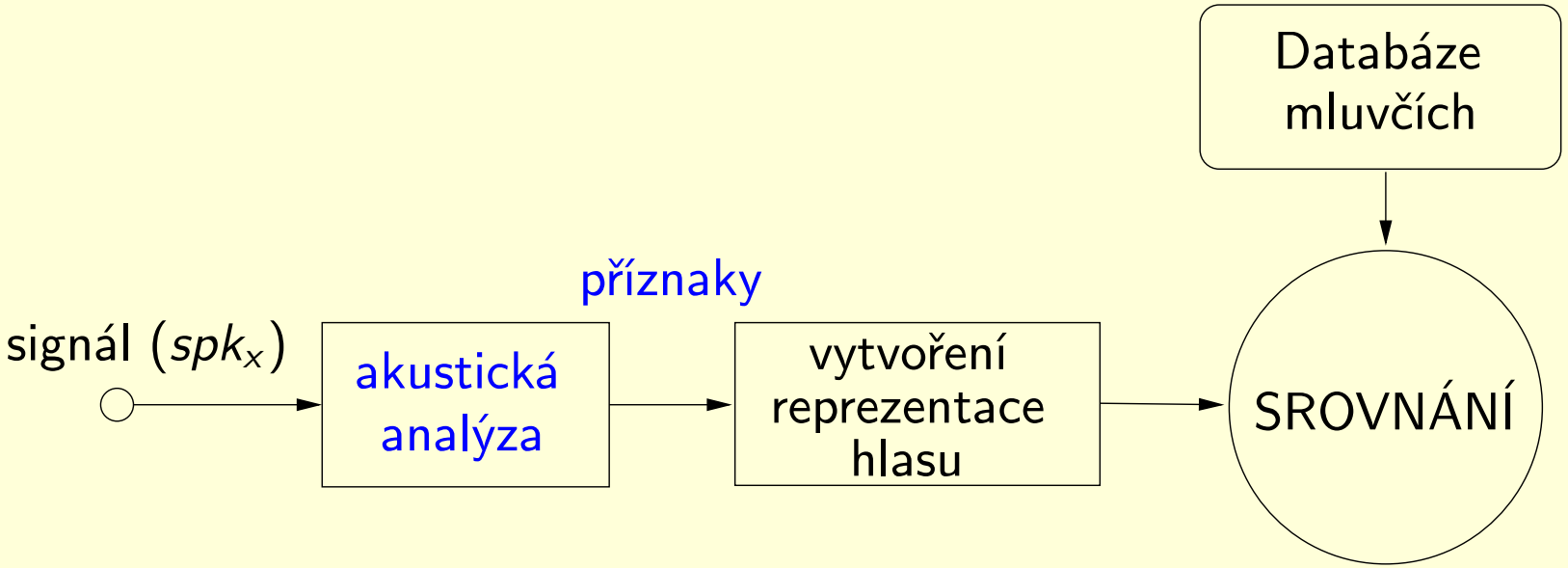
Obsah přednášky

- **Úlohy automatického rozpoznávání řečníka**
  - Verifikace vs. identifikace
- **Reprezentace řečníka a algoritmy klasifikace**
  - Časové funkce (DTW)
  - Kódová kniha (VQ)
  - Statistický model (GMM, HMM)
  - Moderní metody na bázi i-vektorů
  - Rozpoznávání s umělými neuronovými sítěmi
- **Příklady systémů verifikace**

I. část

Úlohy automatického rozpoznávání řečníka

Metody automatického rozpoznávání mluvčího

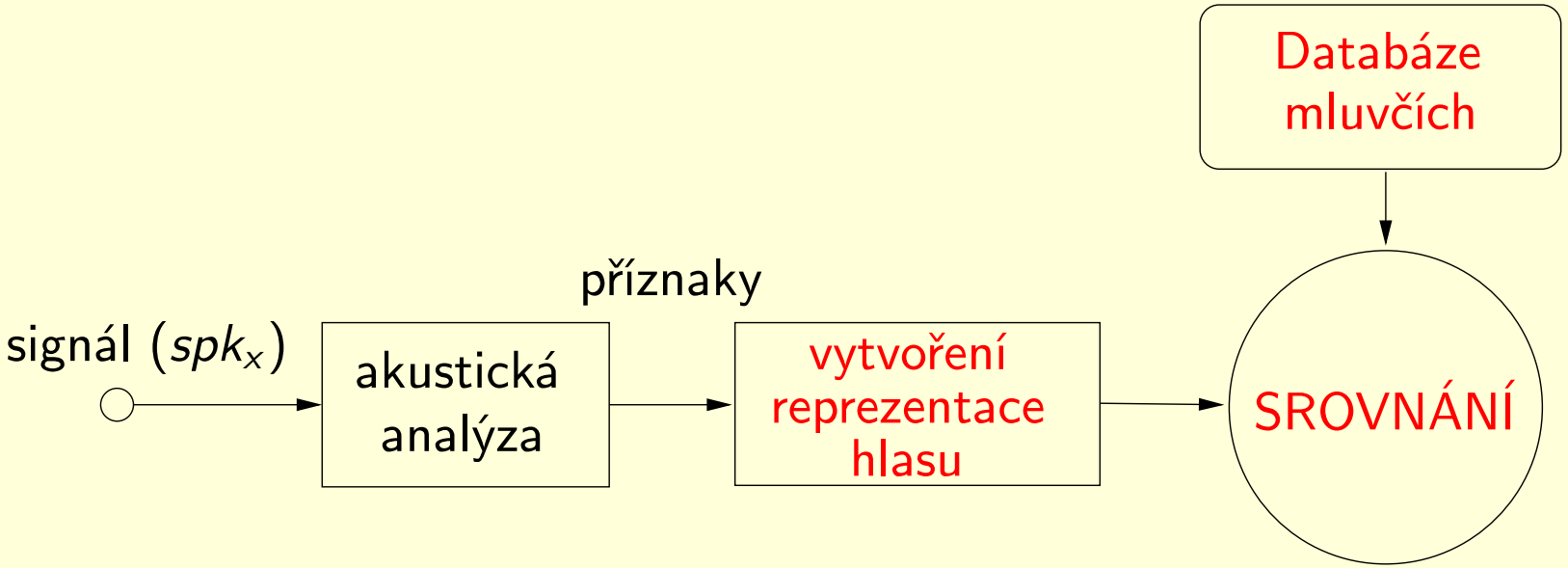


SHRNUTÍ - Používané příznaky:

- **MFCC** - obecně používané - textově nezávislé
  (možnost vyhlazení variability mezi mluvčími)
- LPC kepstrální příznaky
  (variabilita mezi mluvčími, přímá souvislost formanty
  malá robustnost vůči šumu)
- parametry AR modelu
  (menší robustnost, Itakurova vzdálenost)

- **kombinované vektory příznaků** pro komplexnější rozhodování (spíše textově závislé úlohy)
  - f_o - základní frekvence (charakteristika hlasu)
  - formanty (přímá souvislost s délkou vokálního traktu)

Metody automatického rozpoznávání mluvčího



Úlohy rozpoznávání mluvčího

Používaná řešení :

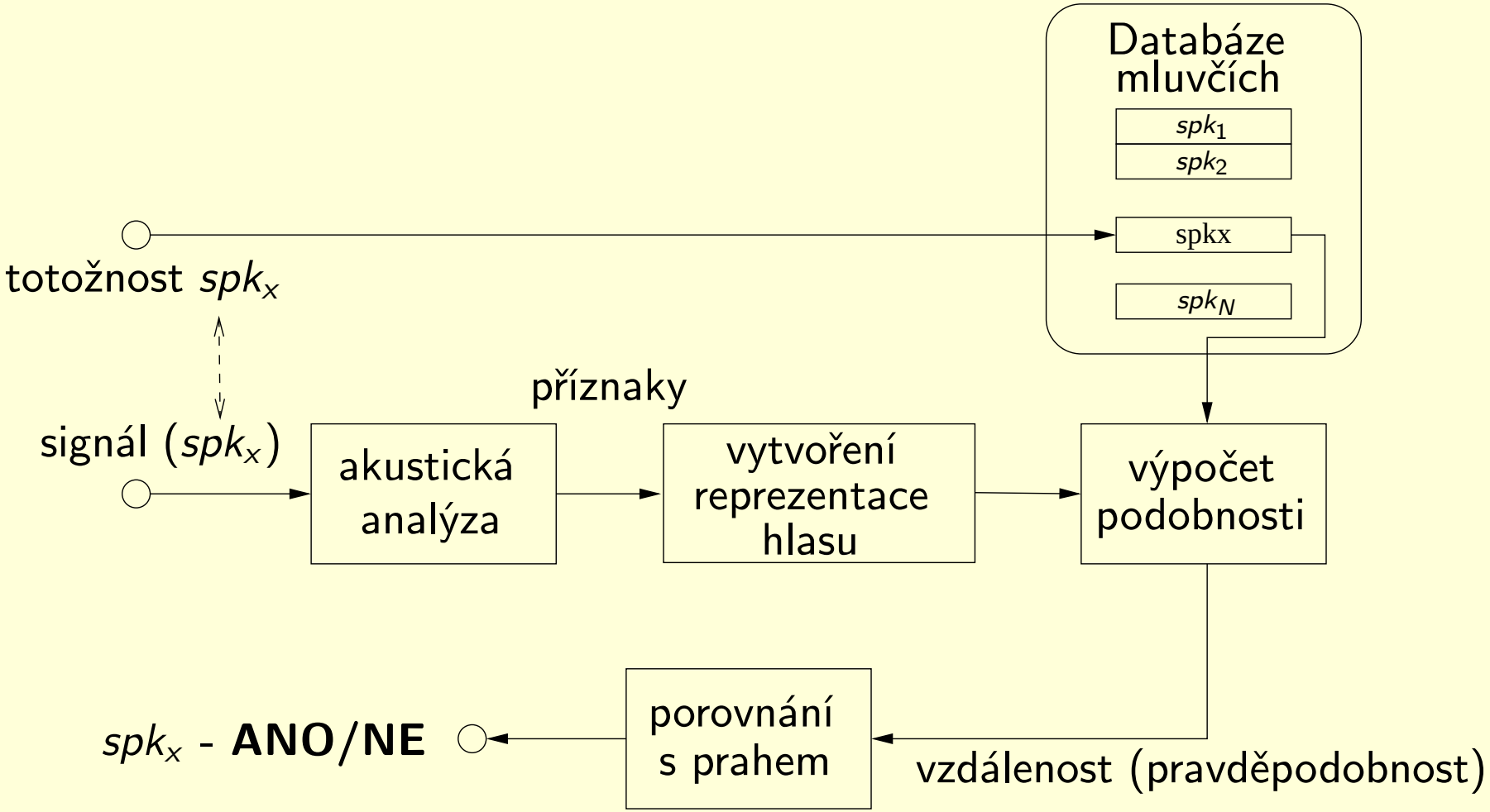
- expertní rozhodování (fonetici, lingvisté)
- automatizované vyhodnocování

Základní úlohy automatického rozpoznávání mluvčího :

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

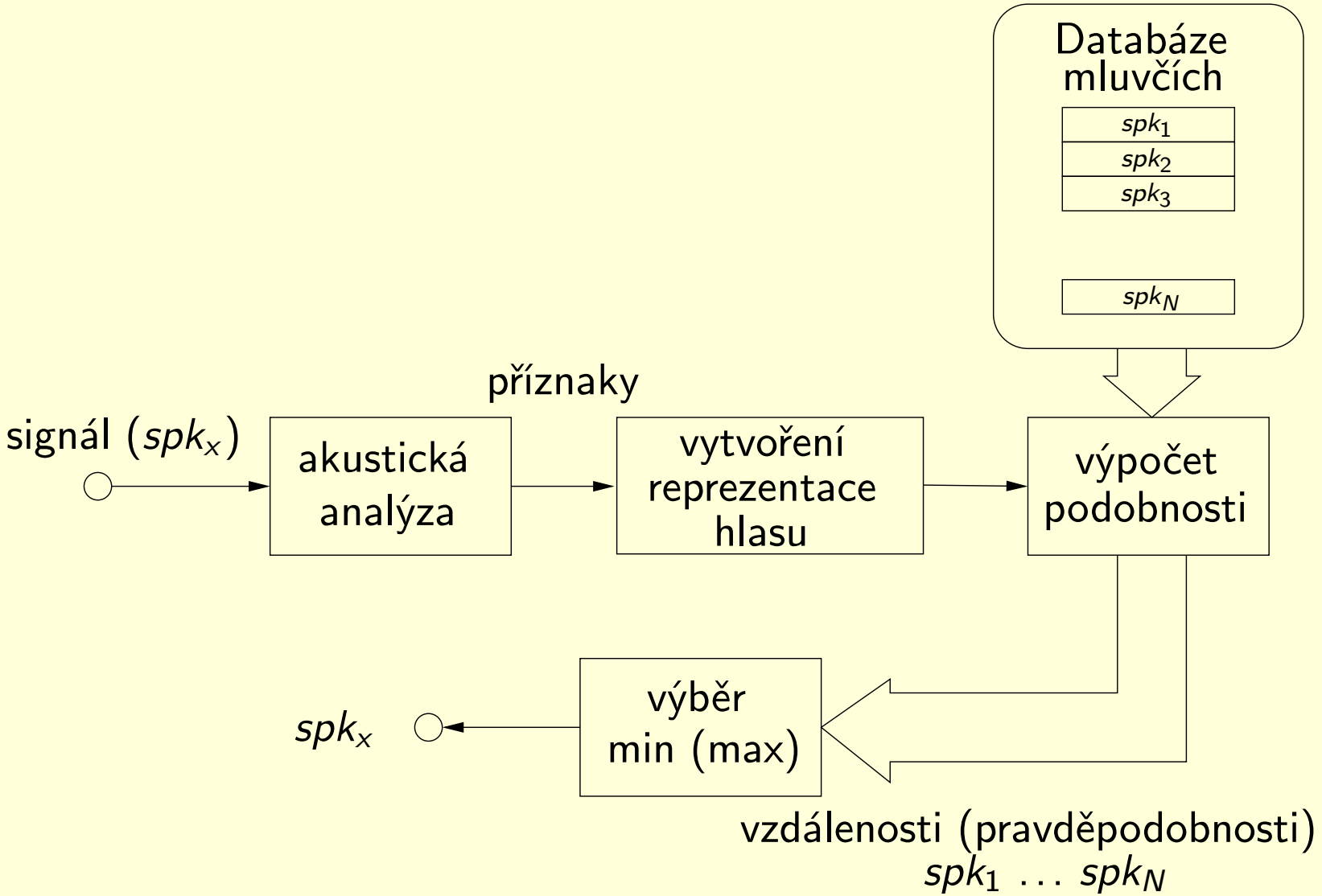
2. **Identifikace mluvčího**
   - **Identifikace v uzavřené množině**
     rozpoznání neznámého mluvčího z dané množiny mluvčích
   - **Identifikace v otevřené množině**
     rozpoznání neznámého mluvčího z neomezené množiny mluvčích $\longrightarrow$ identifikace & verifikace

Verifikace mluvčího



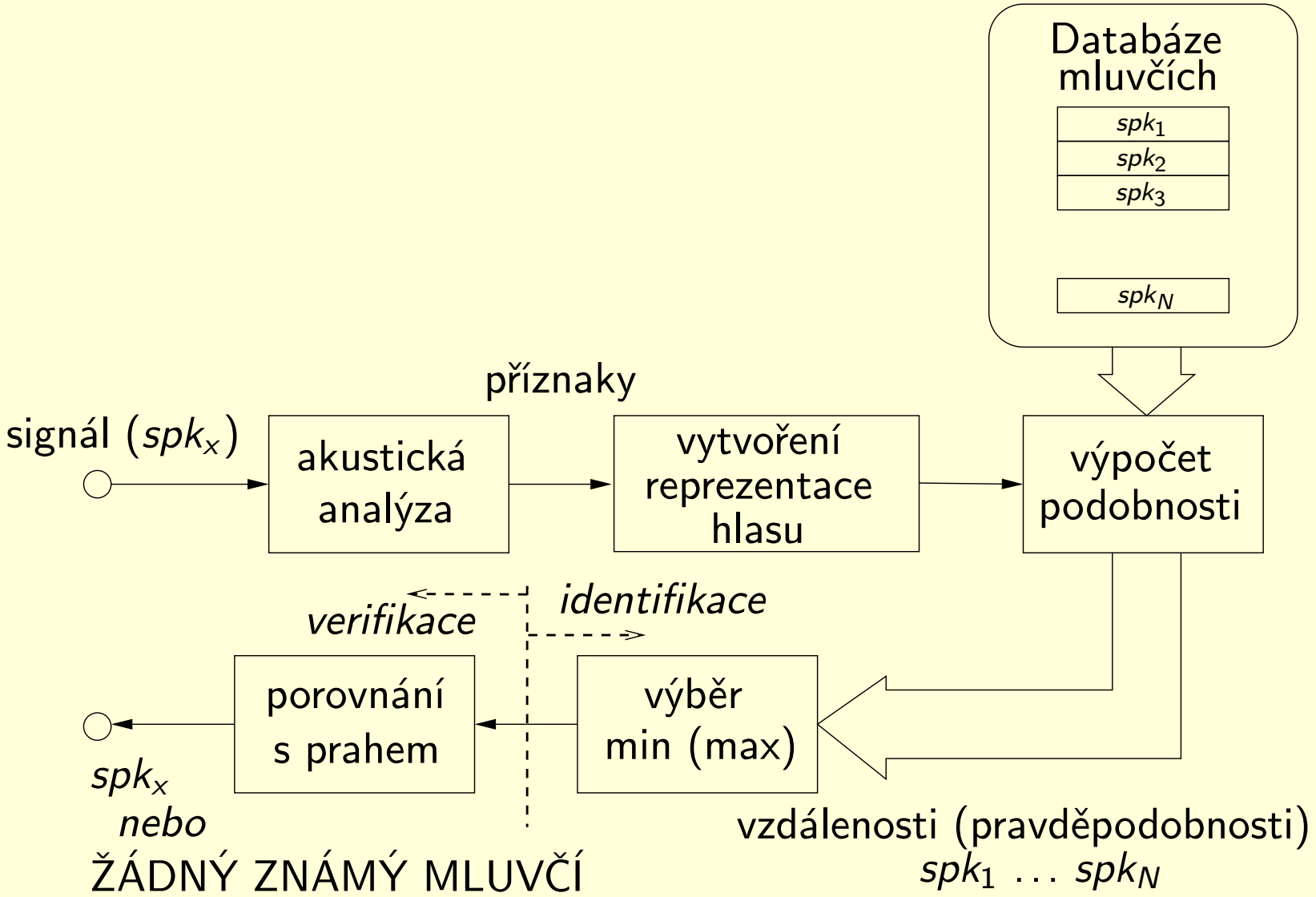
- ověření předpokládané totožnosti mluvčího
- **VÝSLEDEK** = **přijetí / odmítnutí** předpokl. totožnosti

Identifikace mluvčího (v uzavřené množině)



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

Identifikace mluvčího (v otevřené množině)



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

II. část

Klasifikační metody v úloze rozpoznávání řečníka

Reprezentace hlasu řečníka a klasifikační techniky

Reprezentace mluvčího na bázi vzorů

- Vzorová promluva či jiná časová funkce :
  míra = DTW vzdálenost mezi vzorovou a verifikovanou reprezentací
- 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
- HMM modely : modelují též průchod stavy, tj. časovou funkci
- i-vektory : modelování prostoru středních hodnot GMM modelů

Klasifikace na bázi neuronových sítí

- přímá identifikace mluvčího
- výpočet pravděpodobnosti místo GMM

Rozpoznávání na základě časových funkcí

- vzorová promluva
- průběh energie - rytmizace promluvy
- průběh f_o - intonace v promluvě

Vzdálenost mezi promluvami :

- normalizace délky promluvy - prosté prodloužení, zkrácení (problém pro různou rychlost v rámci promluvy)
- normalizovaná kumulovaná vzdálenost na bázi DTW

$$dist_s = \mathrm{dk}\ (\mathbf{O}, \mathbf{O}^s)$$

Vzdálenost mezi všemi segmenty (normovaná na délku):

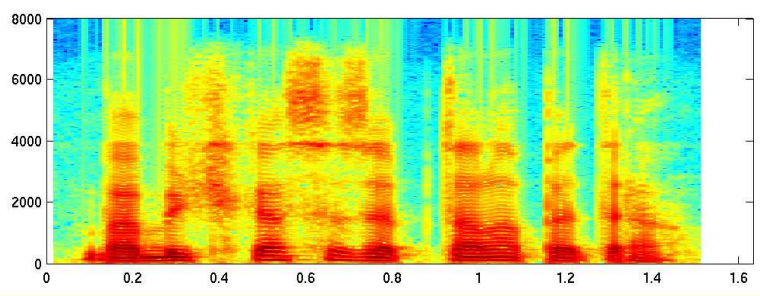
$$d_{i,j} = \frac{1}{NM} \sum_{k=1}^{p} (c_k[i] - \overline{c}_k[j])^2 \quad \text{pro } i = 1, \ldots, N\,;\, j = 1, \ldots, M$$

Kumulovaná vzdálenost algoritmem dynamického programování:

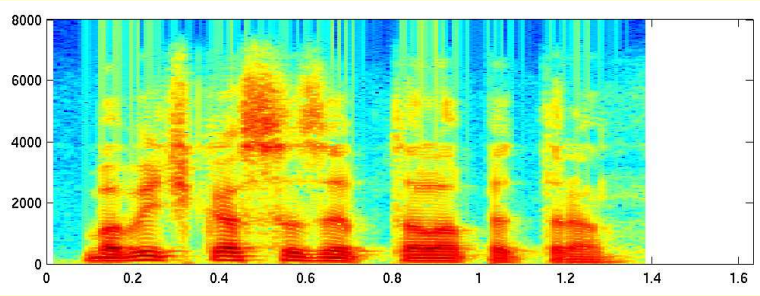
$$dk_{i,j} = \min\left(dk_{i-1,j} + d_{i,j}, dk_{i,j-1} + d_{i,j}, dk_{i-1,j-1} + d_{i,j}\right)$$
$$\text{pro } i = 1, \ldots, N\,;\, j = 1, \ldots, M$$

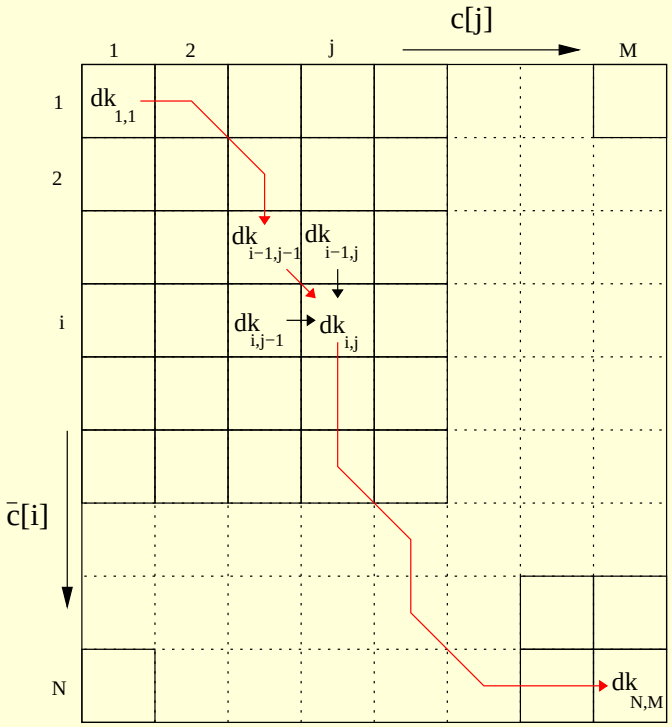
DTW - shrnující popis metody

Referenční promluva délky N



Rozpoznávaná promluva délky M





Identifikace řečníka $\Rightarrow$ hledání minima $dk_{N,M}$ pro různé vzory

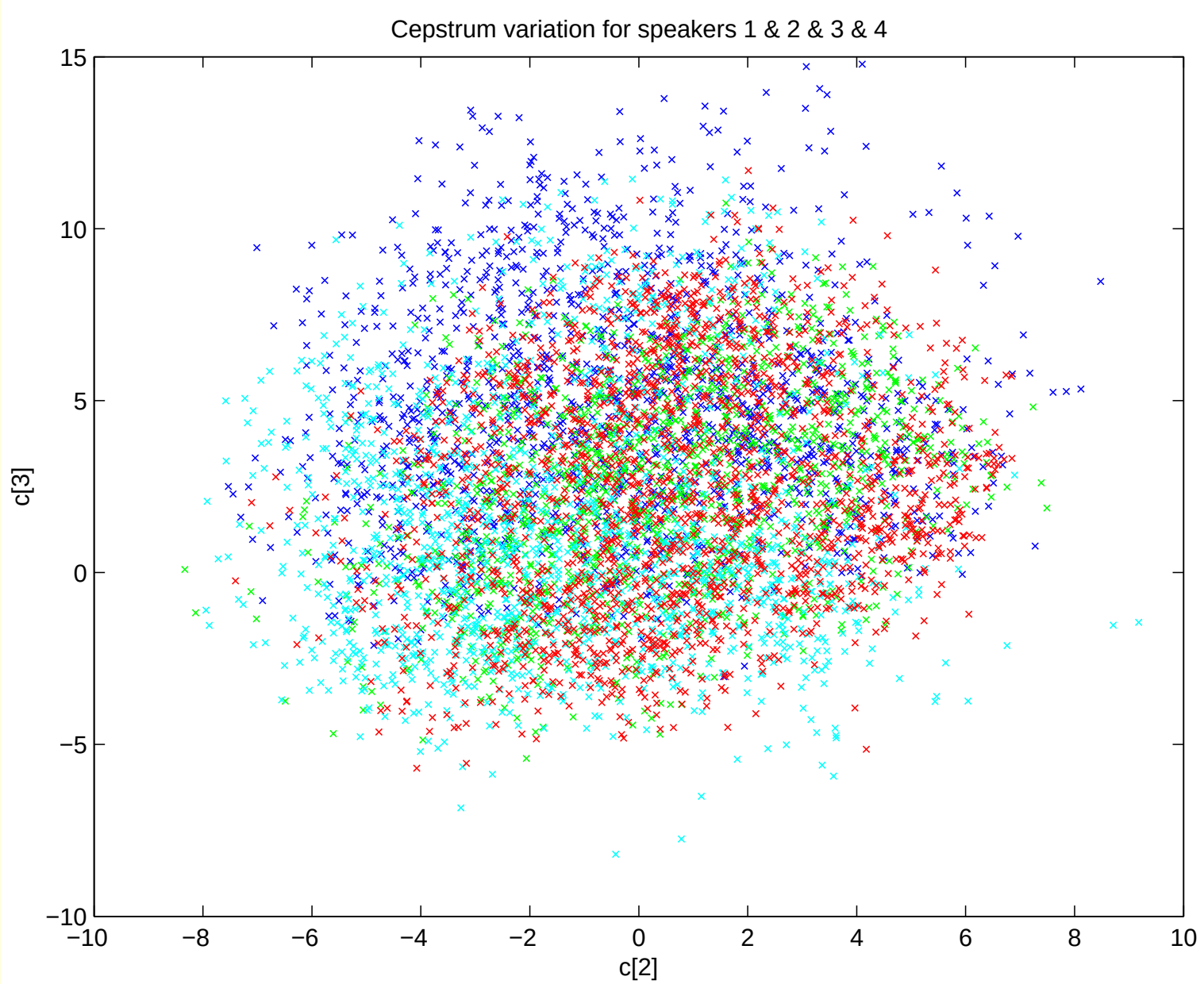
Verifikace řečníka $\Rightarrow$ srovnání $dk_{N,M}$ s prahem

SHRNUTÍ:

- VÝHODY - jednoduchá koncepce, potřeba málo dat
- NEVÝHODA - konkrétní časová funkce

  $\rightarrow$ textově závislá metoda

Klasifikace na bázi VQ
Rozložení kepstra řečníka - 1 & 2 & 3 & 4



Srovnání obecného rozložení příznaků

VQ - vektorová kvantizace

- redukce počtu uchovávaných příznakových vektorů
  $\rightarrow$ **kvantování všech příznakových vektorů**
- rozložení příznaku je potom popsáno kódovou knihou

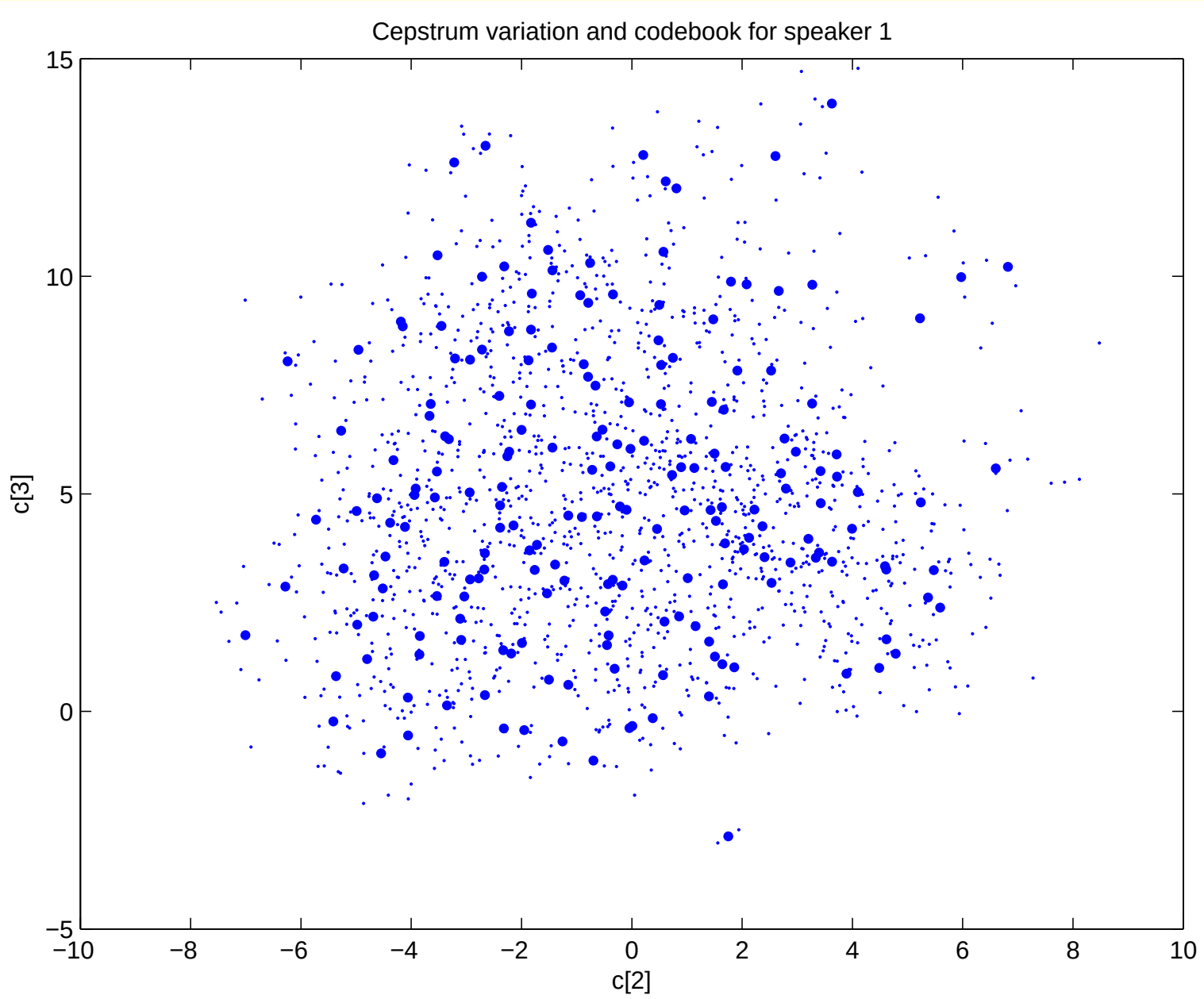
Kódová kniha - c^s_{VQ} ... **konečný počet reprezentantů** popisující variabilitu příznaků

- výpočet na bázi K-means algoritmu

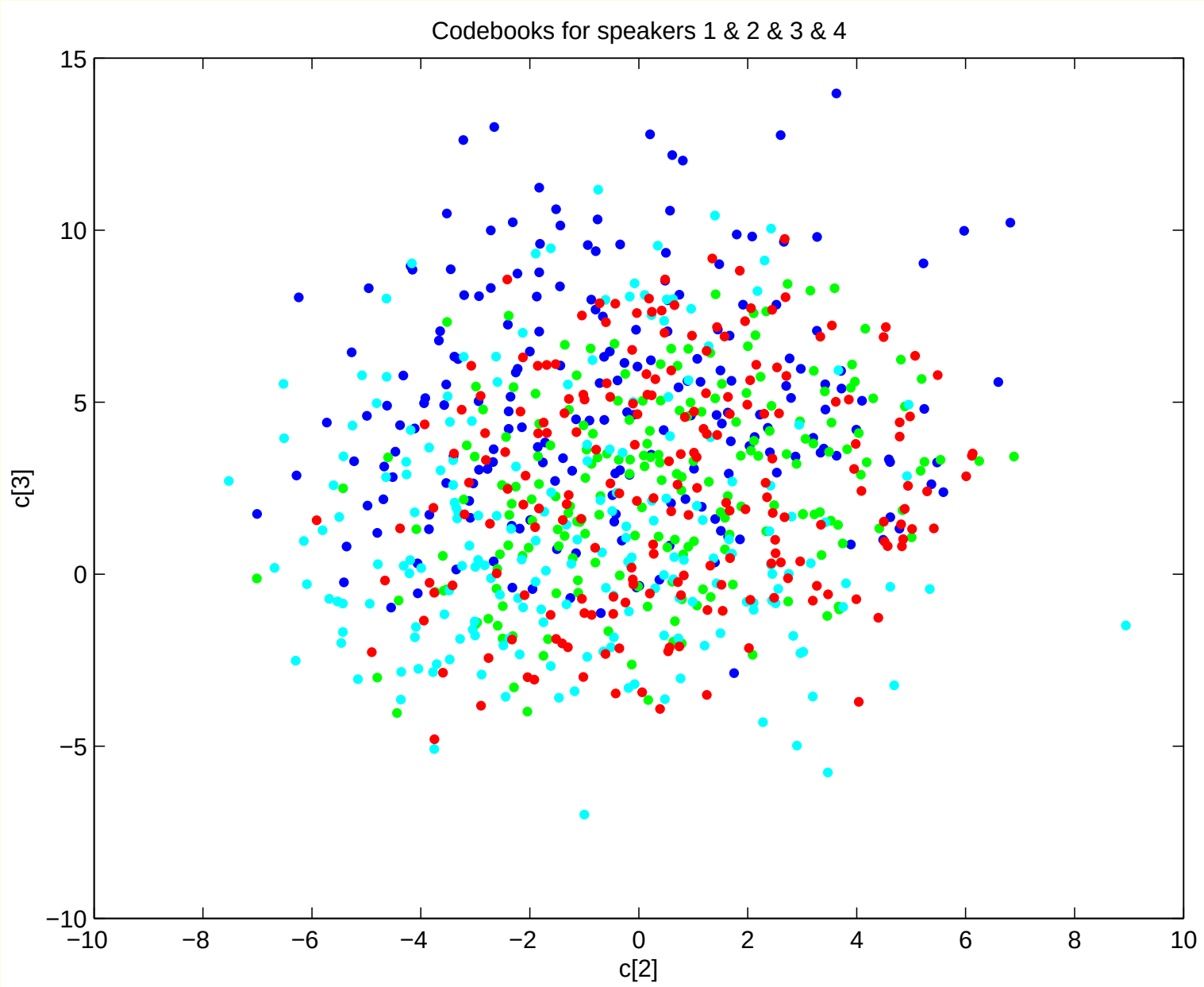
VAD (Voice Activity Detector) při výpočtu příznaků

- vhodné pro **odstranění neřečových segmentů** (jednoduchý energetický detektor může být postačující pro kvalitní signál)

Variace kepstra a kódová kniha vybraného řečníka



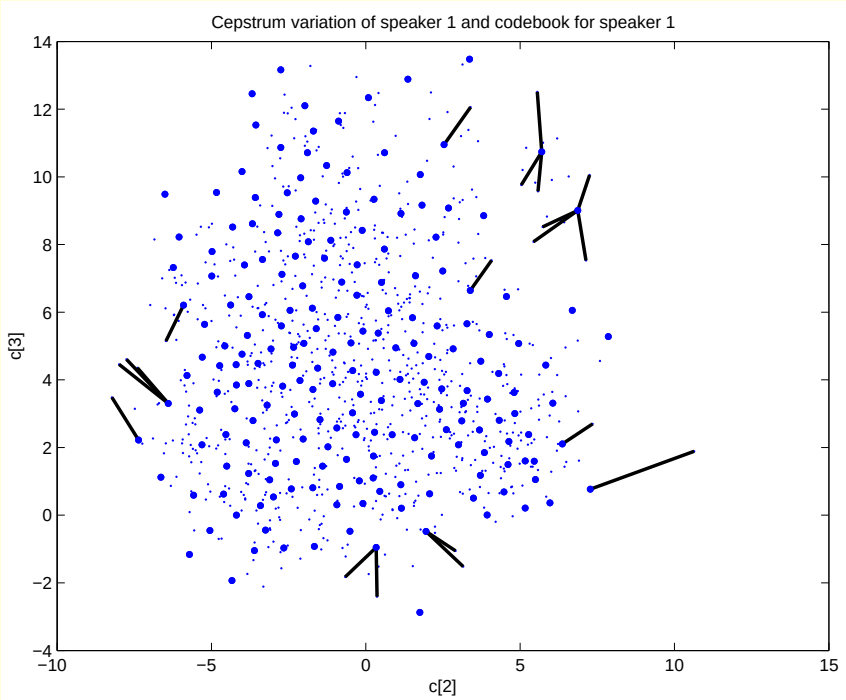
Kódová kniha řečníka - 1 & 2 & 3 & 4

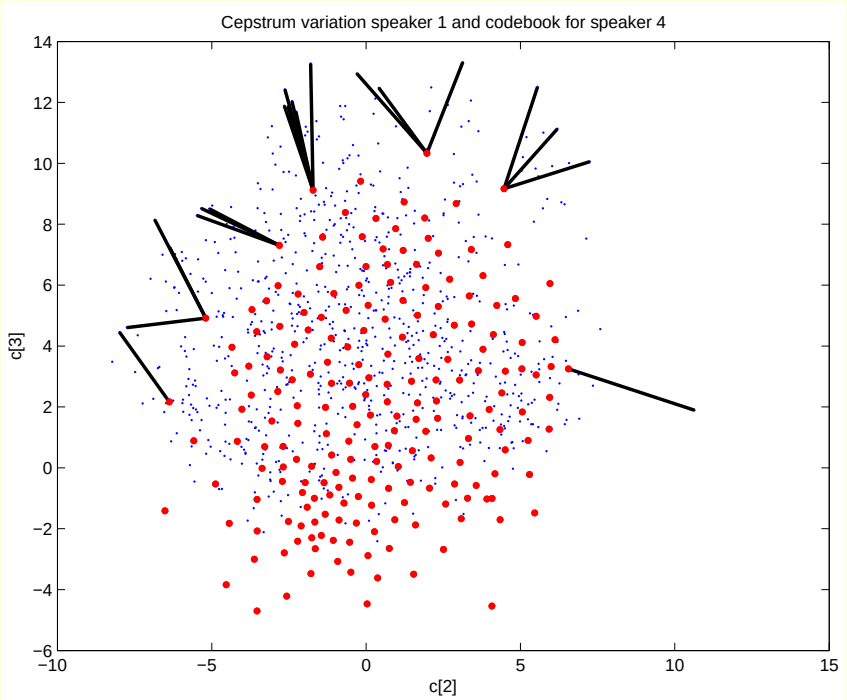


Realizace klasifikace na bázi VQ

Srovnání dvou mluvčích na bázi VQ

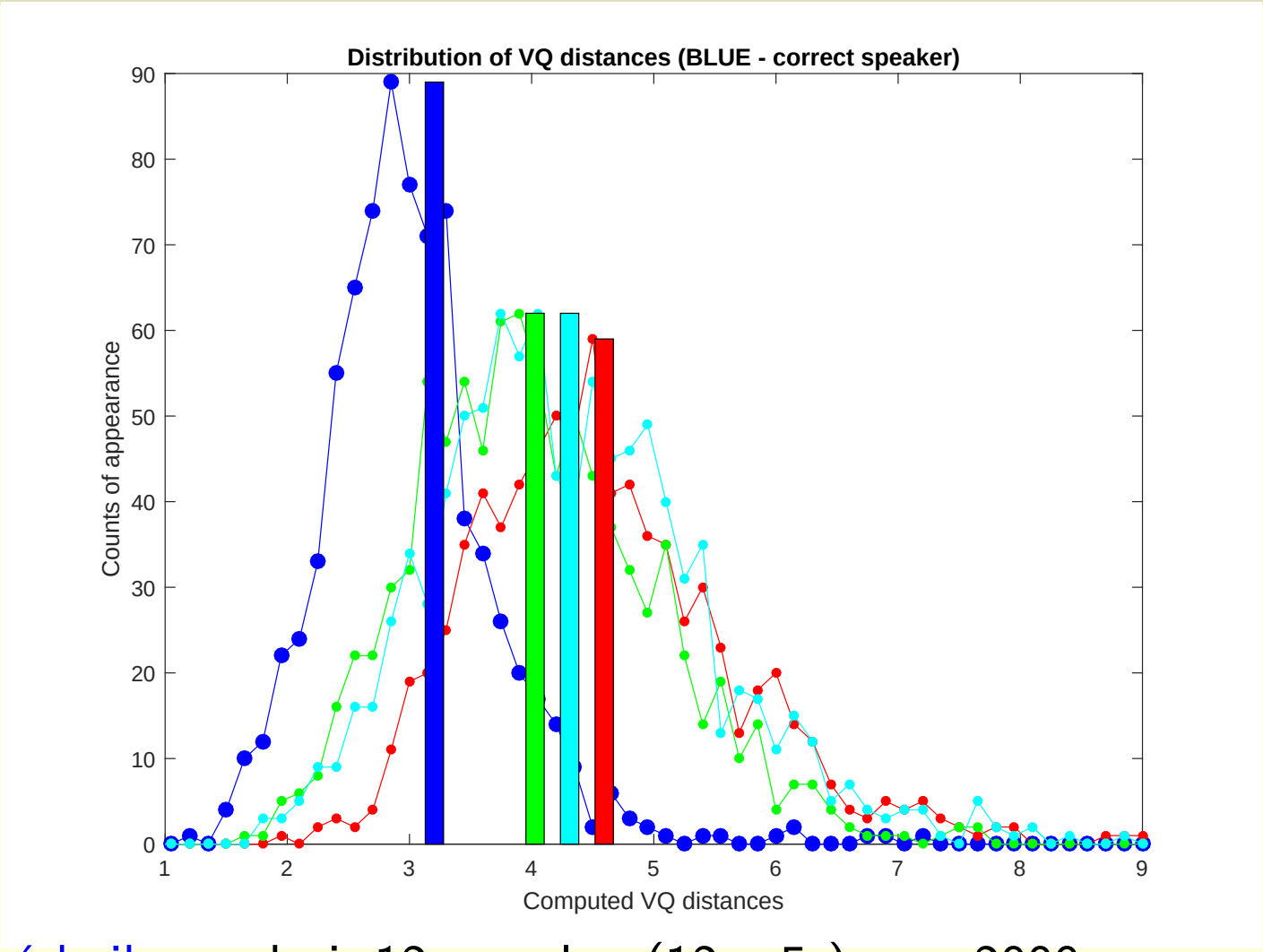
- měření průměrné vzdálenosti aktuálních příznakových vektorů od uložených typických reprezentantů





$$dist_s = \text{mean}\left(\text{cd}\left(c_i, \underset{c_{VQ,s}}{\text{argmin}}\, \text{cd}\left(c_i, c^s_{VQ}\right)\right)\right)$$

Statistiky výsledků pro 4 řečníky a 1 kódovou knihu



Kódová kniha - zdroj: 12 promluv (12 x 5s), cca 2000 segmentů
- velikost kódové knihy: 200

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

Průměrné hodnoty vzdálenosti

Rozpoznávání řečníka na bázi GMM

Hlavní idea - rozložení kepstra je popsáno **statistickým modelem na bázi GMM**

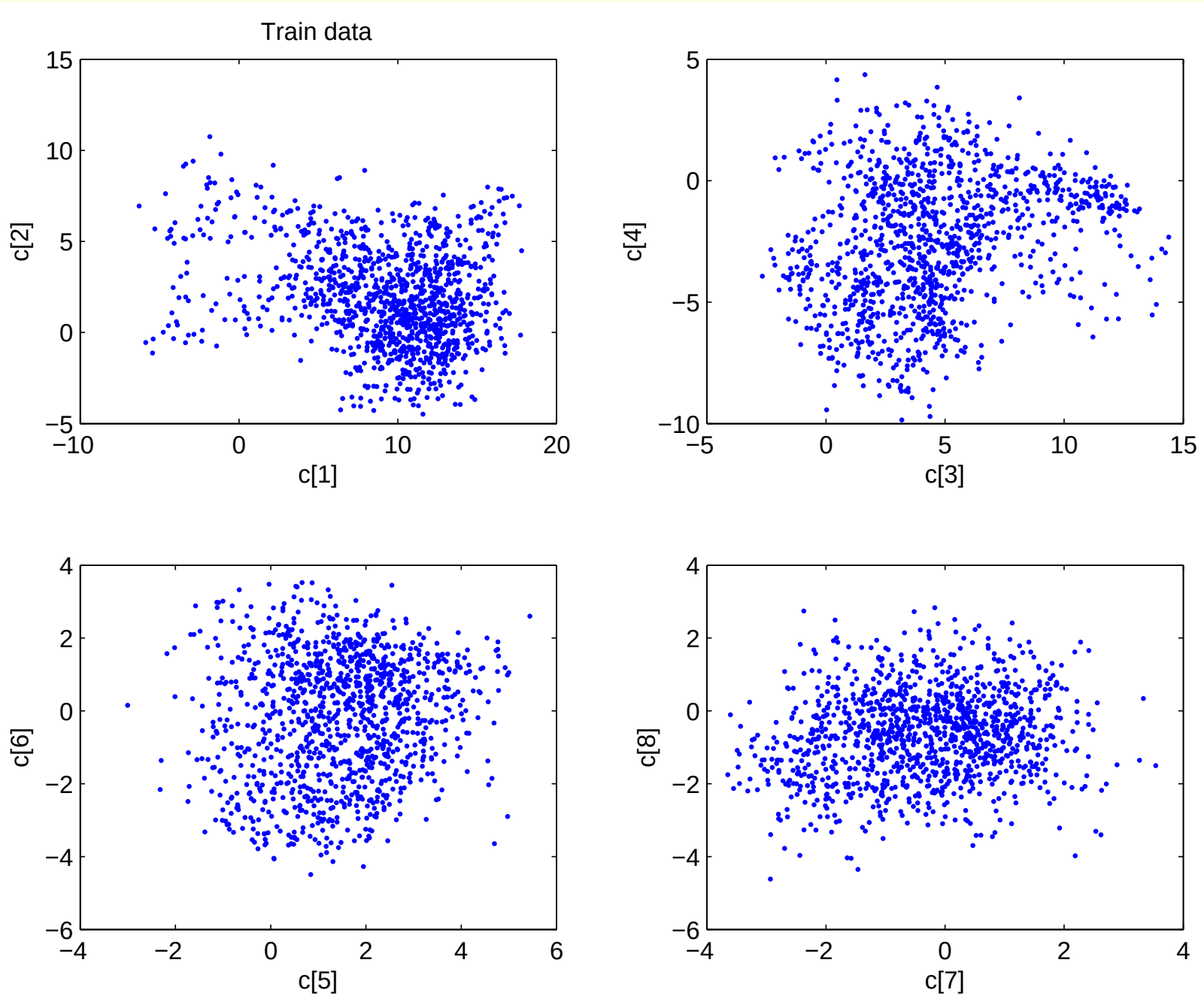
$$p(\mathbf{o}|\lambda^s) = \sum_{i=1}^{M_s} c_i^s \cdot \mathcal{N}(\boldsymbol{o}, \boldsymbol{\mu}_i^s, \mathbf{C}_i^s)$$

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

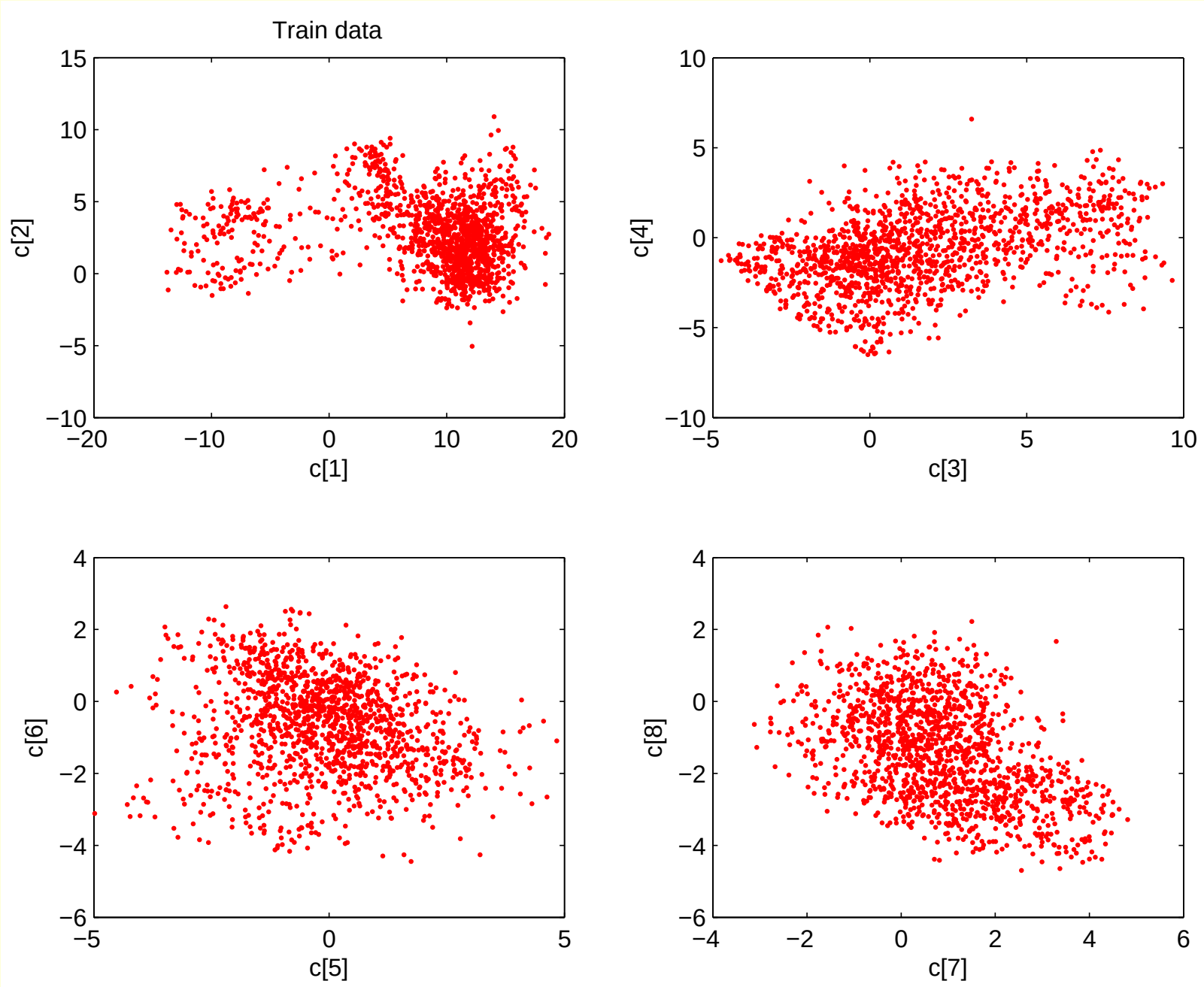
- více směsí modeluje lépe variabilitu příznaků pro daného řečníka

- typické počty směsí: 8-256 (model řečníka), 512-2048 (univerzální model) - *počty směsí zavisí na množství trénovacích dat*

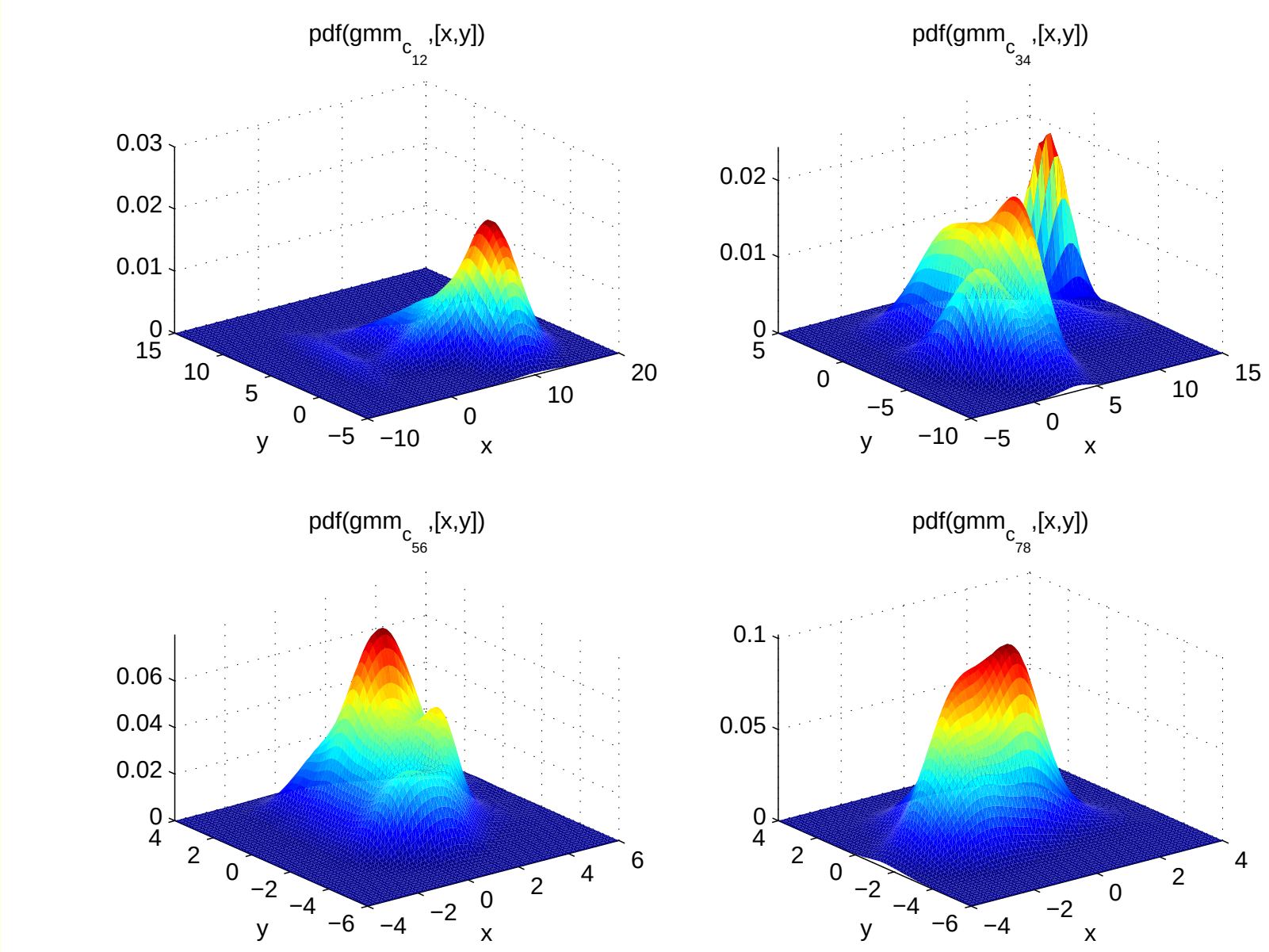
Rozložení prvků kódové knihy kepstra mluvčího A



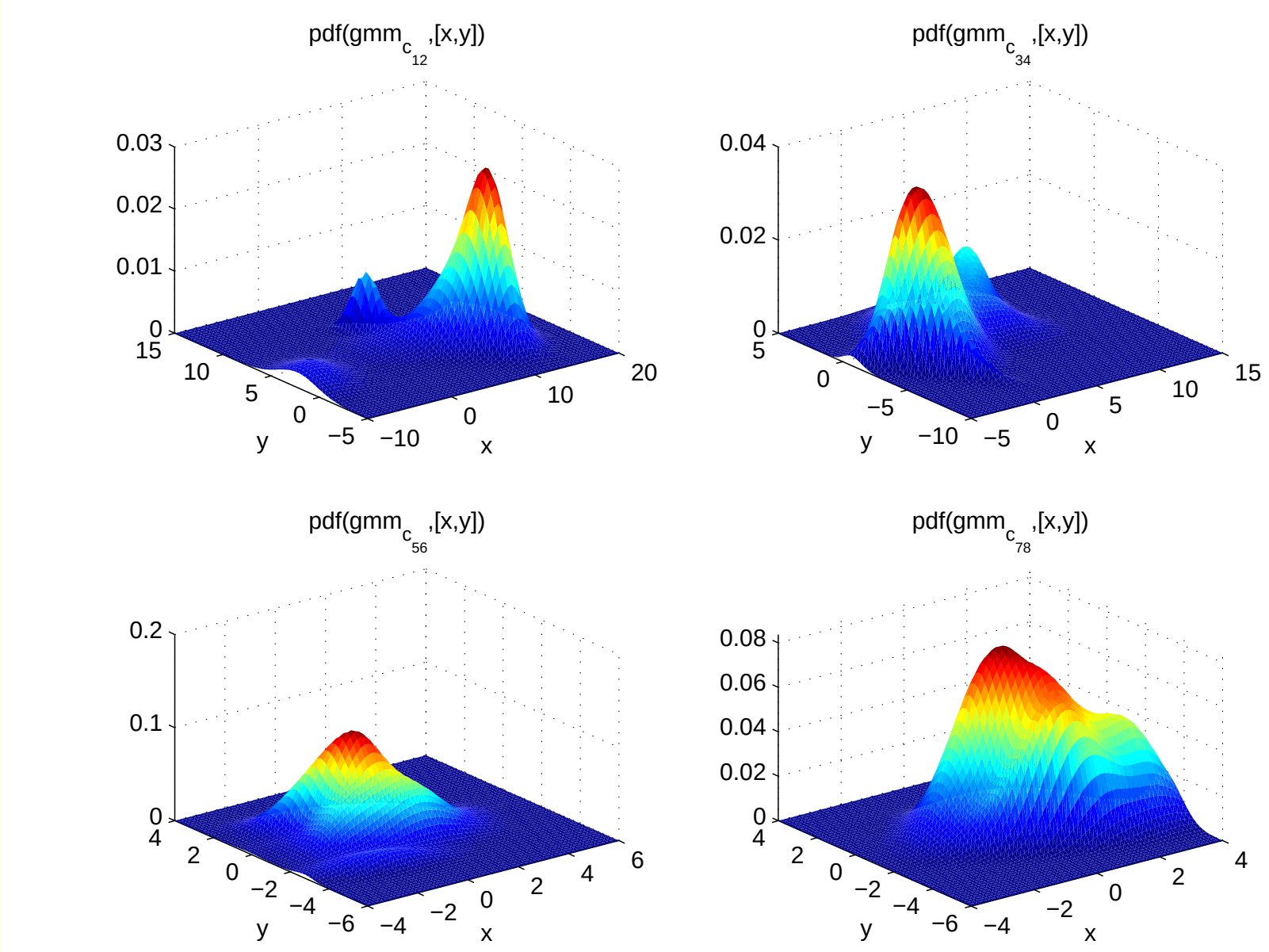
Rozložení prvků kódové knihy kepstra mluvčího B



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

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

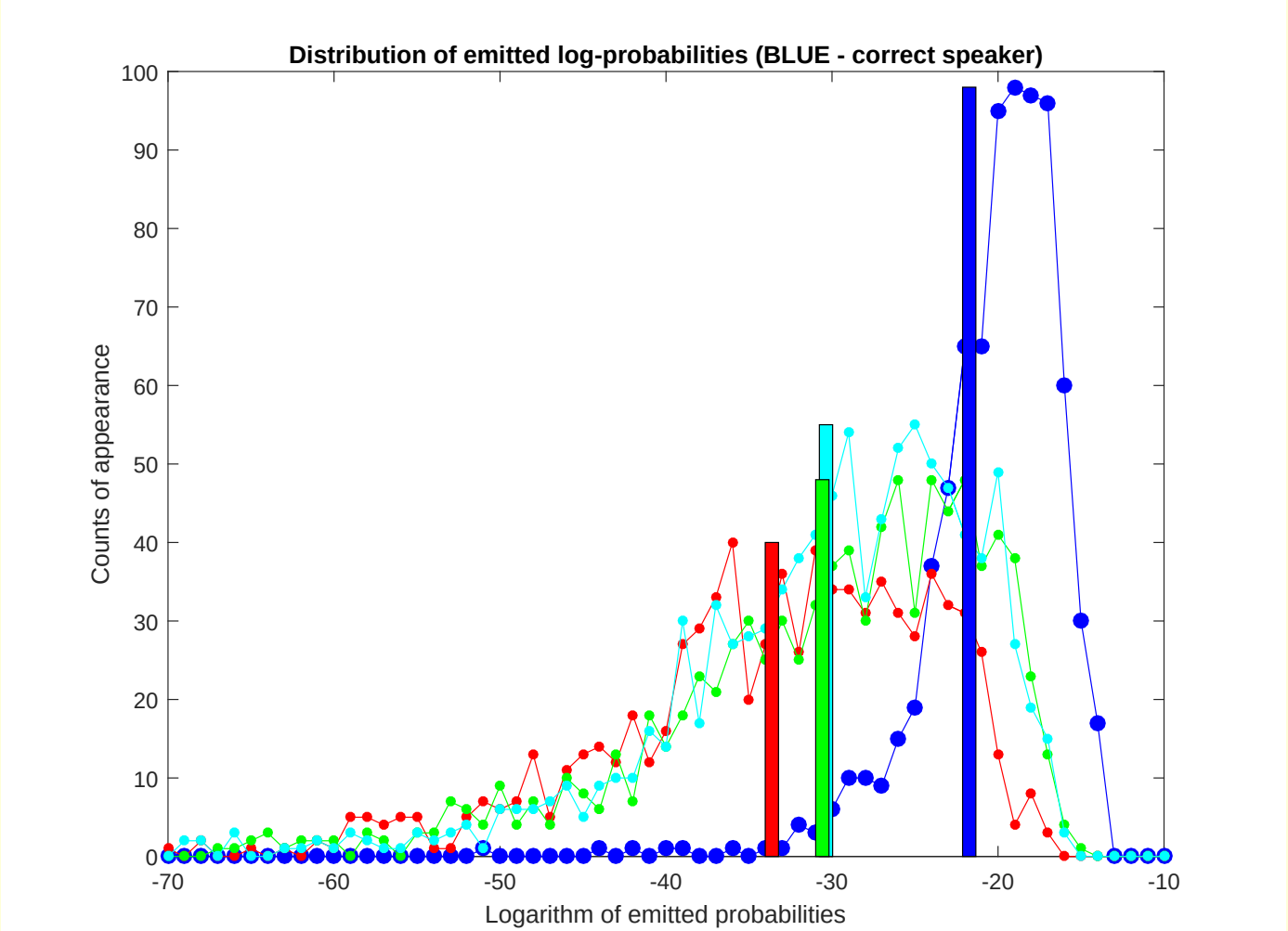
$$P(\mathbf{O}|\lambda^s) = \prod_{j=1}^{N} p(\mathbf{o_j}|\lambda^s)$$

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

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

- opět 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

Rozpoznávání mluvčího na bázi UBM-GMM

Trénování GMM modelu pro jednotlivého mluvčího je problematické:

- málo dat $\rightarrow$ malá schopnost generalizace

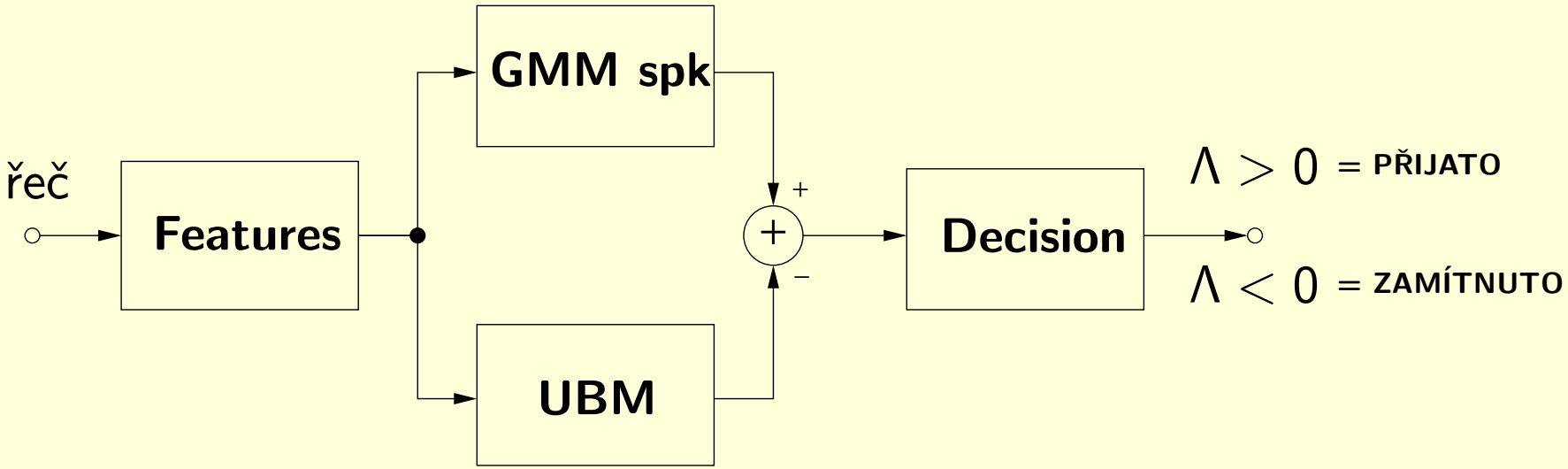
$\Downarrow$

UBM-GMM modelování

- UBM - Universal Background Model
  - generalizující model popisující společný prostor parametrů pro všechny mluvčí
- GMM model mluvčího - získáný adaptací UBM
  - nejčastěji MAP (Maximum Aposteriori Probability)
- UBM-GMM systém = **základ současných systémů**

Rozhodování UBM-GMM verifikace

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



Rozpoznávání mluvčích na bázi i-vektorů

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

- vektor délky $C \cdot F$ všech středních hodnot
- lze použít i pro reprezentaci nahrávek (různé délky)
- lze pak použít i jiné klasifikátory (SVM)
- lze provést dekompozici na složku mluvčího resp. prostředí

i-vektor : jednoduchý model na bázi faktorové analýzy (JFA)

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

- základní idea - redukce dimenze supervektoru $\boldsymbol{m}_{r,s}$, $CF \rightarrow D_{ivec}$
- $\boldsymbol{\mu}$ - supervektor nezávislý na mluvčím a nahrávce
- $\boldsymbol{T}$ - transformační matice dimenze $CF \times D_{ivec}$ (odhad - iterativní EM algoritmus)
- $\boldsymbol{x}_{r,s}$ - **i-vektor** popisující variabilitu mluvčího či prostředí (identity vector, itermediate vector)

Klasifikace na bázi i-vektorů

i-vector : reprezentace mluvčích (trénování) či nahrávek (provoz)
SVM klasifikátor s jádrovou funkcí na bázi kosinové vzdálenosti

$$score = \frac{\boldsymbol{x}_1^T \boldsymbol{x}_2}{||\boldsymbol{x}_1|| \cdot ||\boldsymbol{x}_2||}$$

- verifikace hypotézy, že mluvčí v nahrávce 1 je shodný s mluvčím v nahrávce 2
- srovnání skóre s verifikačním prahem

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

Zobecnění rozhodování v úlohách rozpoznávání řečníka

- **identifikace**

$$spk = \underset{s}{\text{argmin}}\, dist_s \quad \text{pro} \quad s = 1, 2, ..., L$$
$$spk = \underset{s}{\text{argmax}}\, P(\mathbf{O}|\lambda^s) \quad \text{pro} \quad s = 1, 2, ..., L$$

- **verifikace**

$dist_s < dist_{thr}$ .... předpokládaná identita PŘIJATA
$dist_s > dist_{thr}$ .... předpokládaná identita ZAMÍTNUTA

$P(\mathbf{O}|\lambda^s) > P_{thr}$ .... předpokládaná identita PŘIJATA
$P(\mathbf{O}|\lambda^s) < P_{thr}$ .... předpokládaná identita ZAMÍTNUTA

ANN/DNN - Artificial Neural Networks/Deep Neural Networks
Použití: výpočet pravděpodobnosti místo GMM či přímo klasifikace

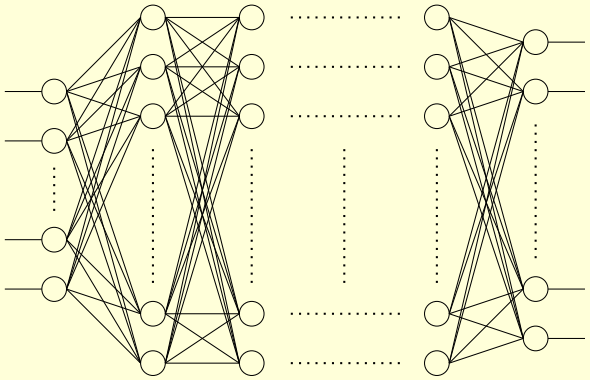
VÝHODY:

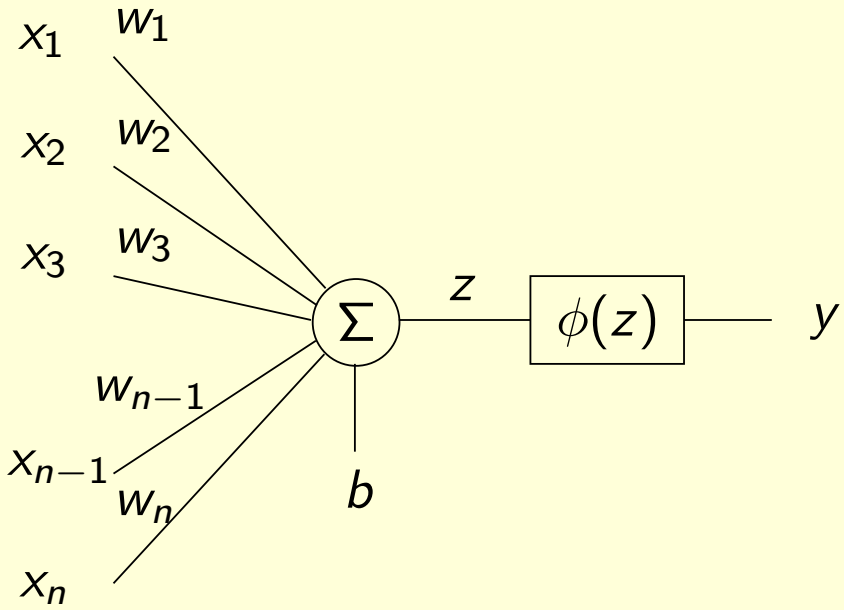
- možnost natrénování složitější funkce
- možné rozšíření příznakového vektoru
(řetězení příznaků s širším kontextem)
- přesnější výsledky lze dosáhnout s DNN
(vícevrstvé sítě s hlubokým učením - deep learning)
- použití konvolučních sítí (CNN) umožňuje End-to-End processing
(vstup: přímo signál/spektrogram, výstup: přímo ID mluvčího)

NEVÝHODY:

- obecně náročnější trénování (algoritmy hlubokého učení)
- potřeba většího množství trénovacích dat
(nastavení mnoha vnitřních parametrů sítě)

ANN, DNN - dopředný výpočet





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

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

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)

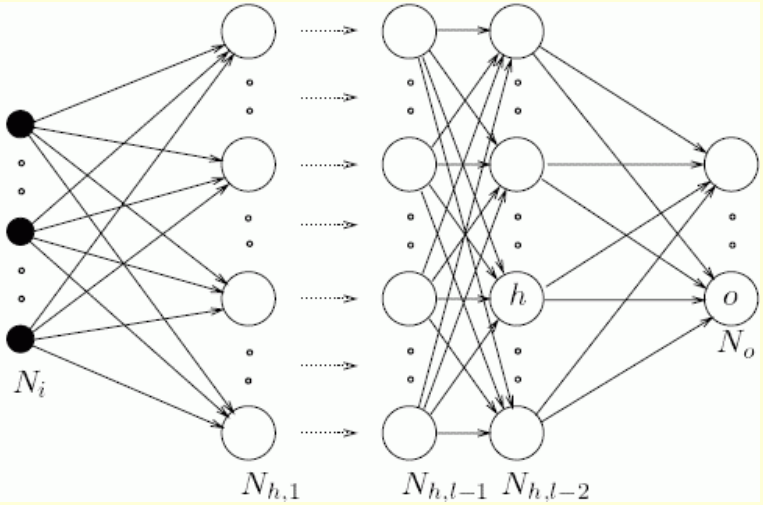
Inicializace sítě před trénování:

- 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

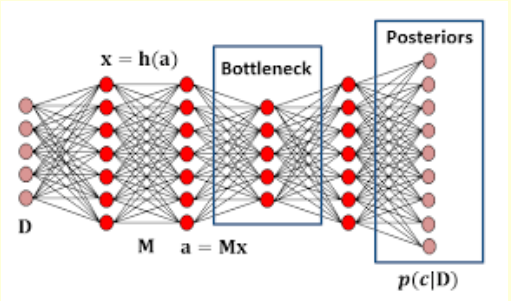


VSTUP: BF, kepstrum (MFCC), možný kontext několik oken

SKRYTÉ VRSTVY: 4-10

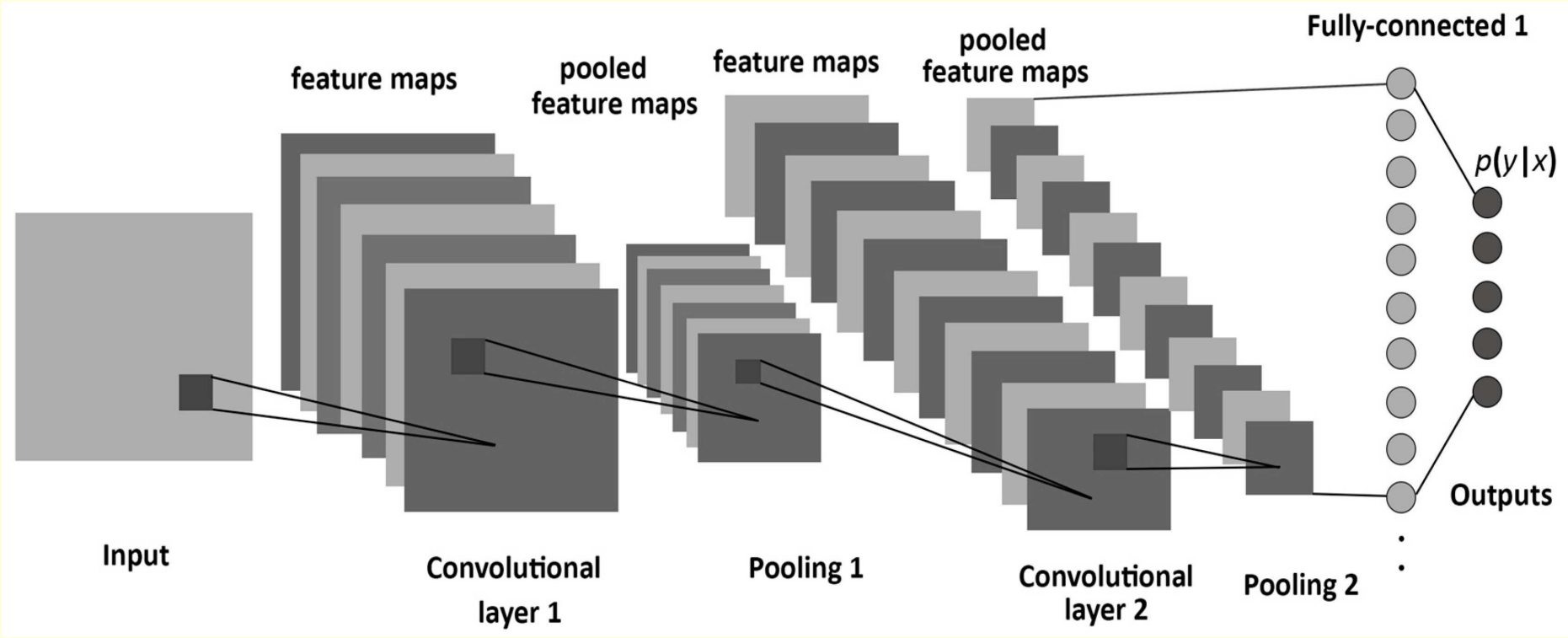
VÝSTUP: Softmax (aposteriors)

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



CNN (Convolution Networks): End-to-End Recognition

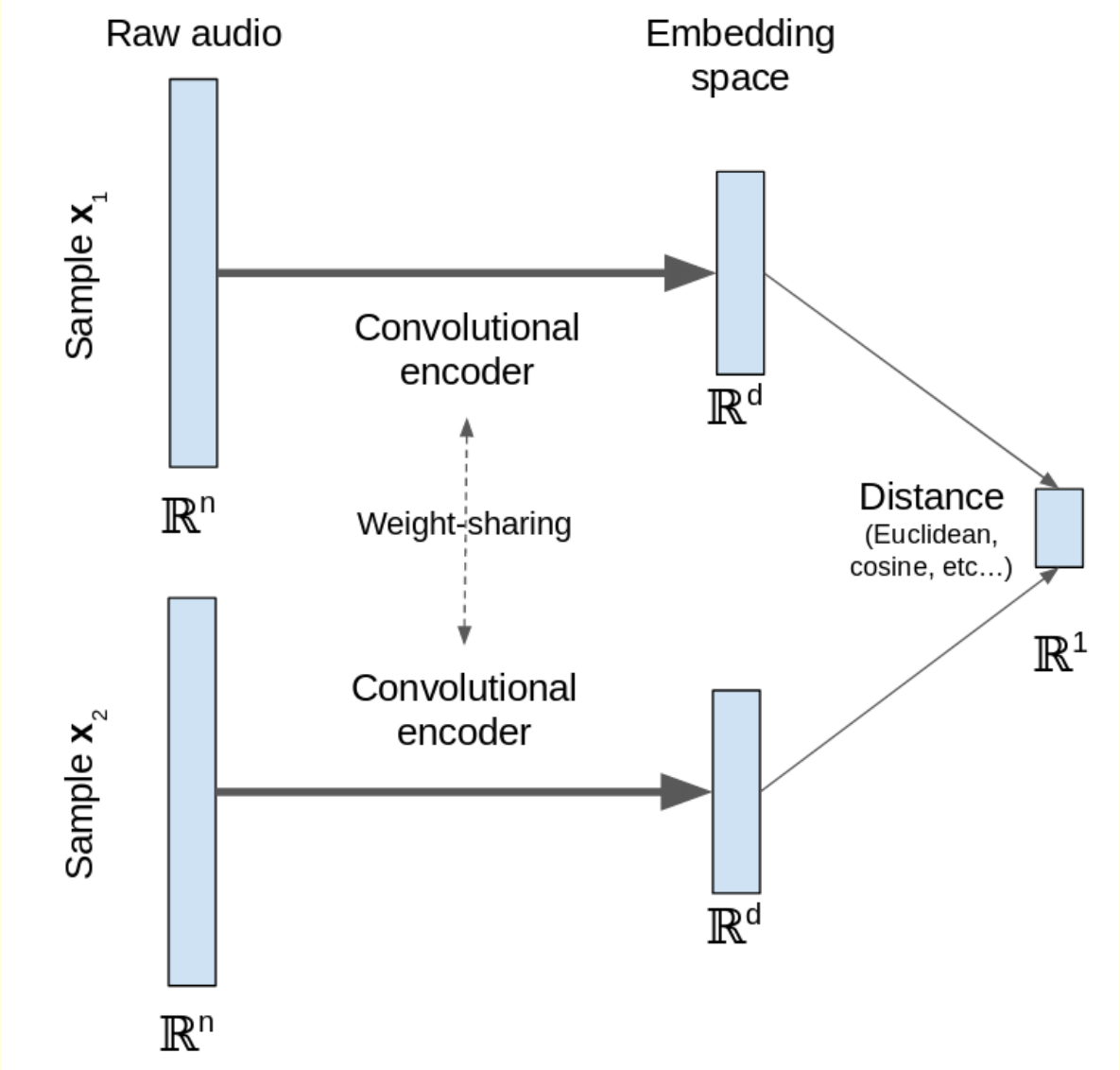
Principiální schéma CNN:



Nejčastější aplikace CNN ve zpracování obrázků

Aplikace pro řeč: vstupem je spektrogram (obrázek) či signál
$\rightarrow$ End-to-End Recognition (tj. bez výpočtu příznaků)

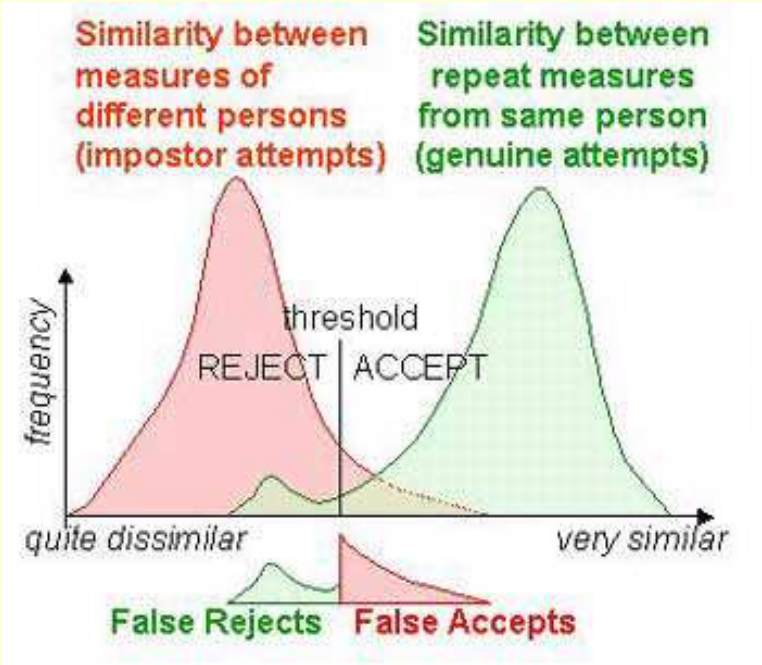
CNN Siamese Speaker Verification

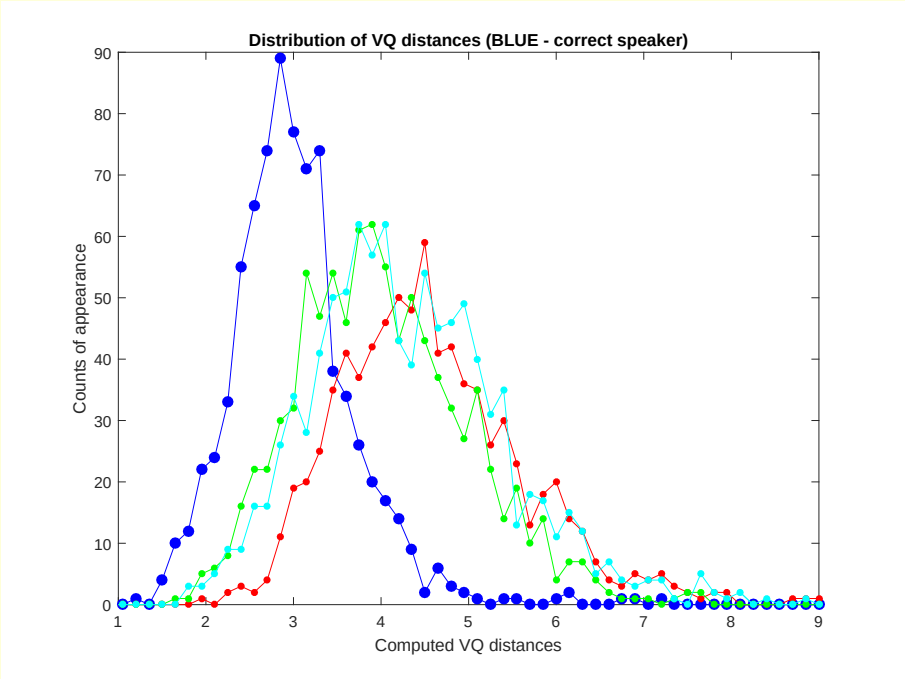


III. část

Příklady systémů rozpoznávání řečníka

Hodnotící kritéria při verifikaci mluvčího - Míra stejné chyby



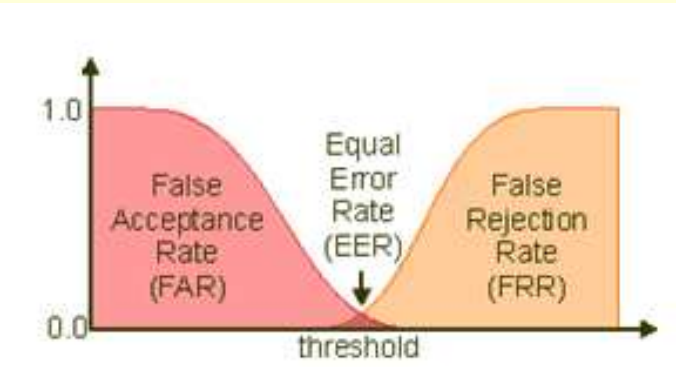


TA - True acceptance

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 :

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

Příklady systémů

J. R. Campbell: Speaker Recognition. Department of Defense Fort Meade, MD, at http://scgwww.epfl.ch

- přehled různých systémů verifikace

Autor	Příznaky	Metoda	Text	Vstup	Error
Atal'74	kepstr.	Pattern	depend.	LAB	2% (1s)
Fururi'81	nor. kepstr.	Pattern	depend.	TEL	0,2% (3s)
Doddington'85	FB	DTW	depend.	LAB	0,8% (6s)
Tishby'91	kepstr.	HMM	10 digits	TEL	2,8% (1,5s) 0,8% (3,5s)
Reynolds'96	MFCC+Δ	GMM	indep.	TEL match.	11% (3s) 6% (10s) 3% (30s)
Reynolds'96	MFCC+Δ	GMM	indep.	TEL mism.	16% (3s) 8% (10s) 5% (30s)

Příklady systémů

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

Příklady systémů

Současné systémy na bázi GMM a ANN/DNN

Tian et al, Interspeech 2015 - NIST 2010 task
EER 2.91% - GMM UBM - 2048
EER 2.28% - DNN (English)
EER 2.61% - DNN (Multiling)

Garcia-Romero & McCree, Interspeech 2015
EER 1.82% UBM-GMM - 8kHz, 25/10 ms, preemfáze, 20 MFCC+Δ, short-time CVMN, 2048 mixtures, UBM, PLDA speaker subspace
EER 1.23% - DNN - 4-5 hidden layers, příznaky 9x40 dim (20 MFCC+Δ), LDA+MLLT

Chen, Moreno, etal, Interspeech 2015
EER 3.88-5.53% fully connected DNN - 48 mel filter-banks, 12 frame context, 4 x 256 hidden layers, softmax output function, 3200 output speakers

Příklady systémů - Interspeech 2016

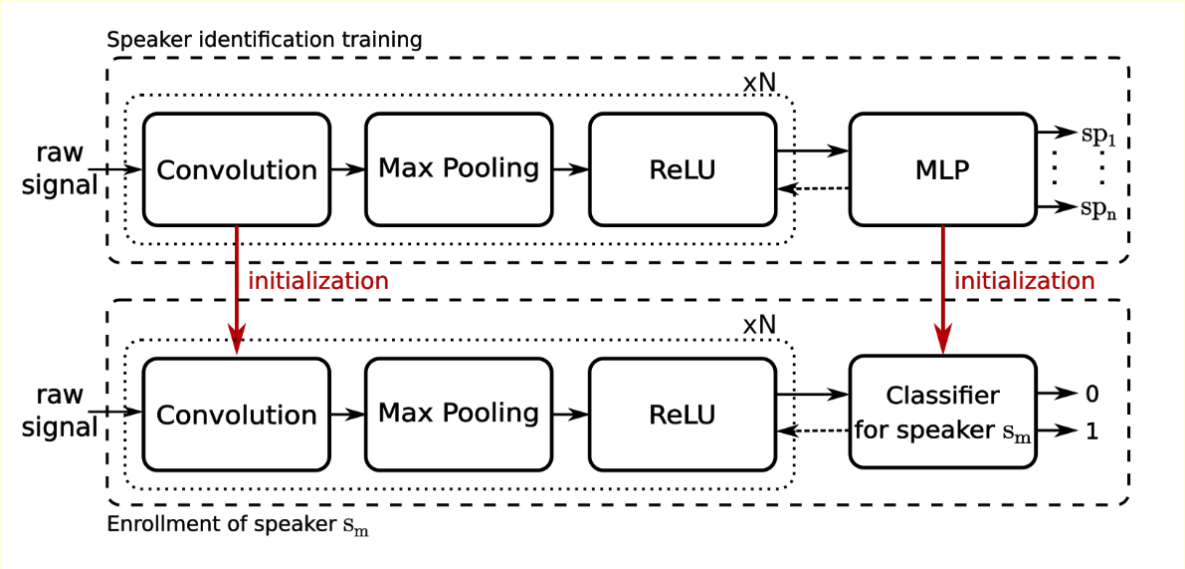
The IBM Speaker Recognition System
EER 2.11% - GMM-UBM (i-vector) - MFCC - LDA
EER 1.49% - GMM-UBM (i-vector) - MFCC - NDA
(NDA - Nearest-neighbour discriminant analysis)
EER 0.59% - DNN-fMLLR-NDA (English)

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 - 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)

Příklady systémů

Vlasta Radová: Rozpoznávání řečníka. ZČU Plzeň. Prosinec 2004.

Komplexní systém využívající kombinované příznaky a víceúrovňové rozhodování

- textově závislá verifikace
  - EER 0.5 % - vstup z mikrofonu (16 kHz)
  - EER 2 % - vstup z telefonní linky (8 kHz)

- textově nezávislá verifikace
  - EER 2 % - vstup z mikrofonu (16 kHz)
  - EER 10 % - vstup z telefonní linky (8 kHz)

Děkuji vám za pozornost !