A6M33BIO - Biometrie

Biometrické metody založené na rozpoznávání hlasu II

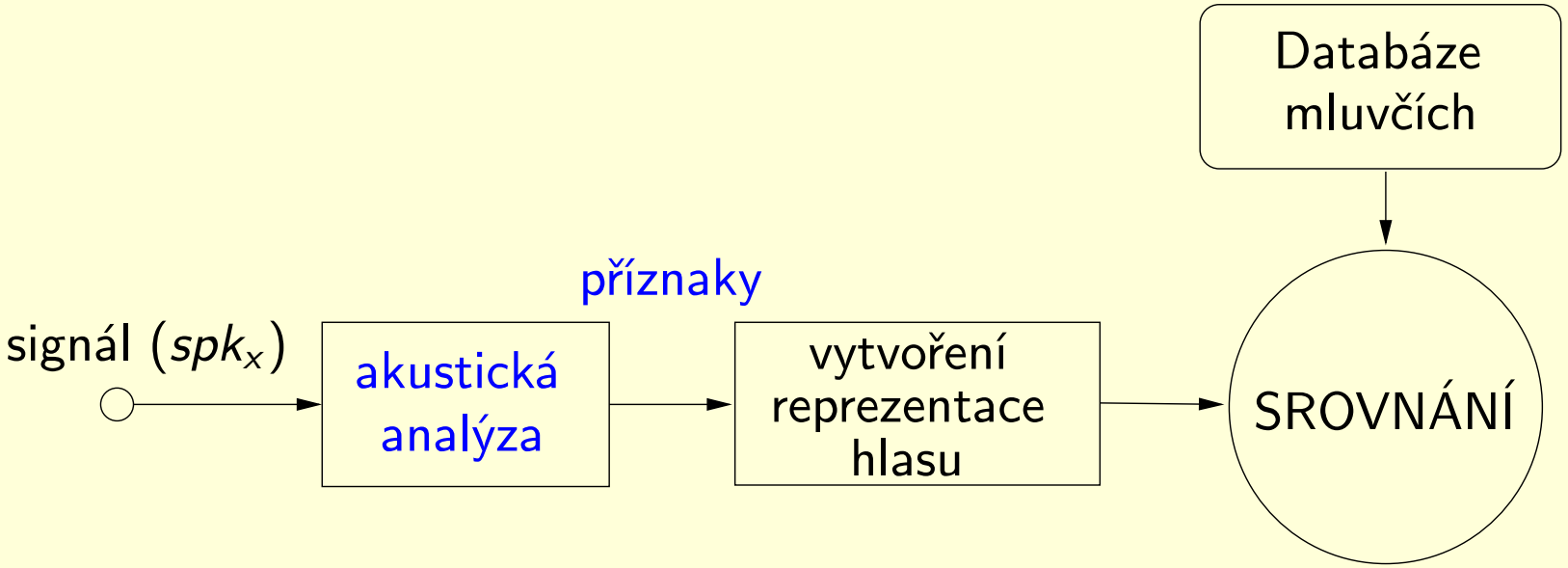
Doc. Ing. Petr Pollák, CSc.

10. prosince 2021 - 11:19

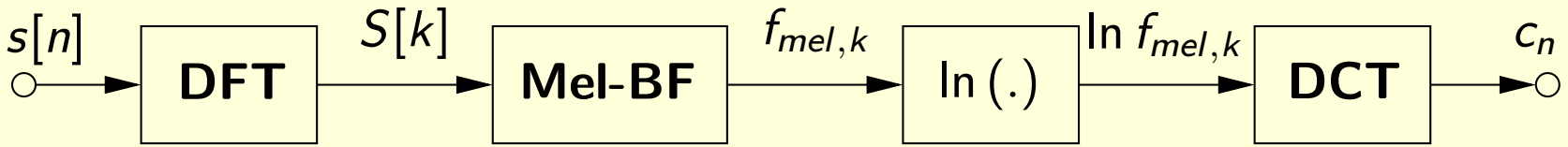
Obsah přednášky

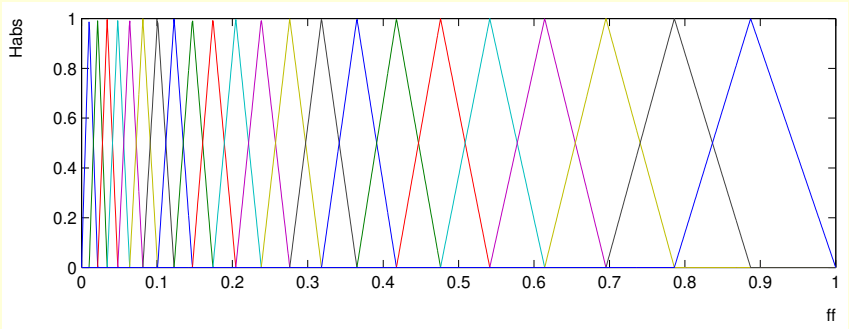
- **Textově nezávislá verifikace/identifikace**
  - Metody na bázi vektorové kvantizace (VQ)
  - Statistické metody na bázi (GMM, GMM-UBM)
  - Moderní metody na bázi i-vektorů
- **Rozpoznávání řečníka s neuronovými sítěmi**
  - Používané struktury neuronových sítí
  - Příklady speciálních architektur
  - Rozpoznávání řečníka na bázi x-vektorů
- **Příklady systémů verifikace**

SHNRUTÍ - příznaky pro textově nezávislý SRE systém

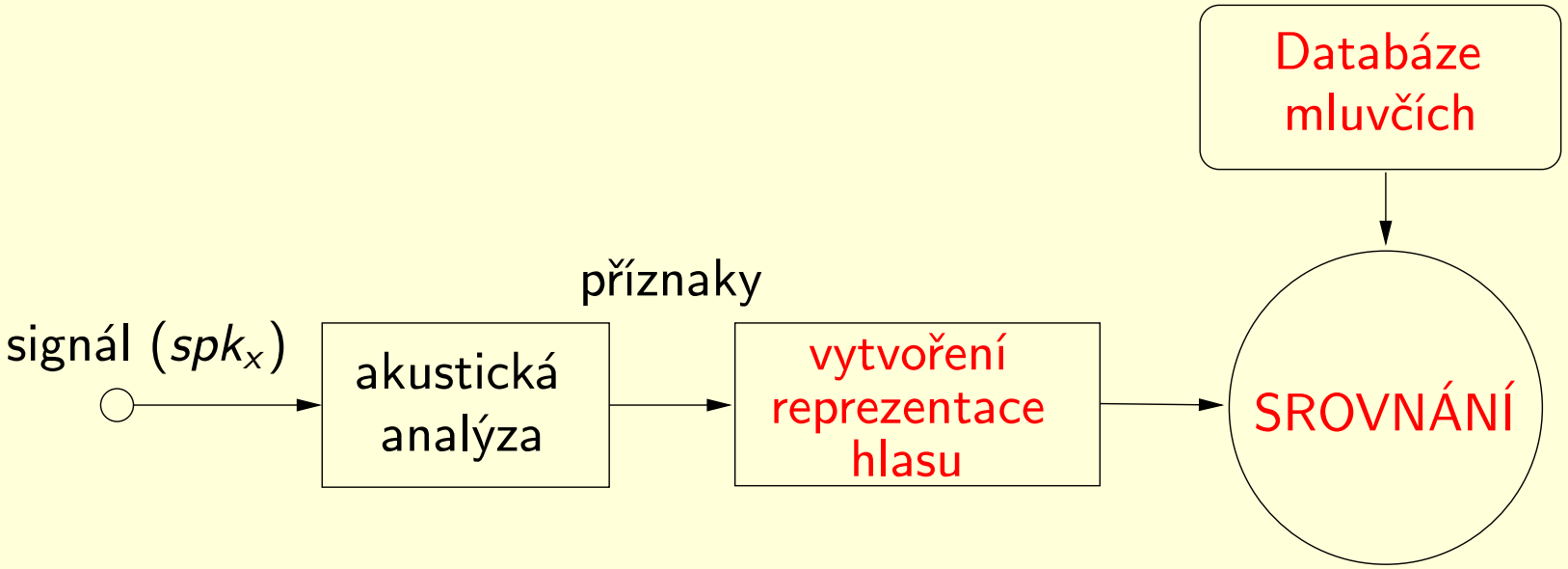


MFCC (mel-frequency cepstral coeficients)





Reprezentace hlasu pro textově nezávislé SRE



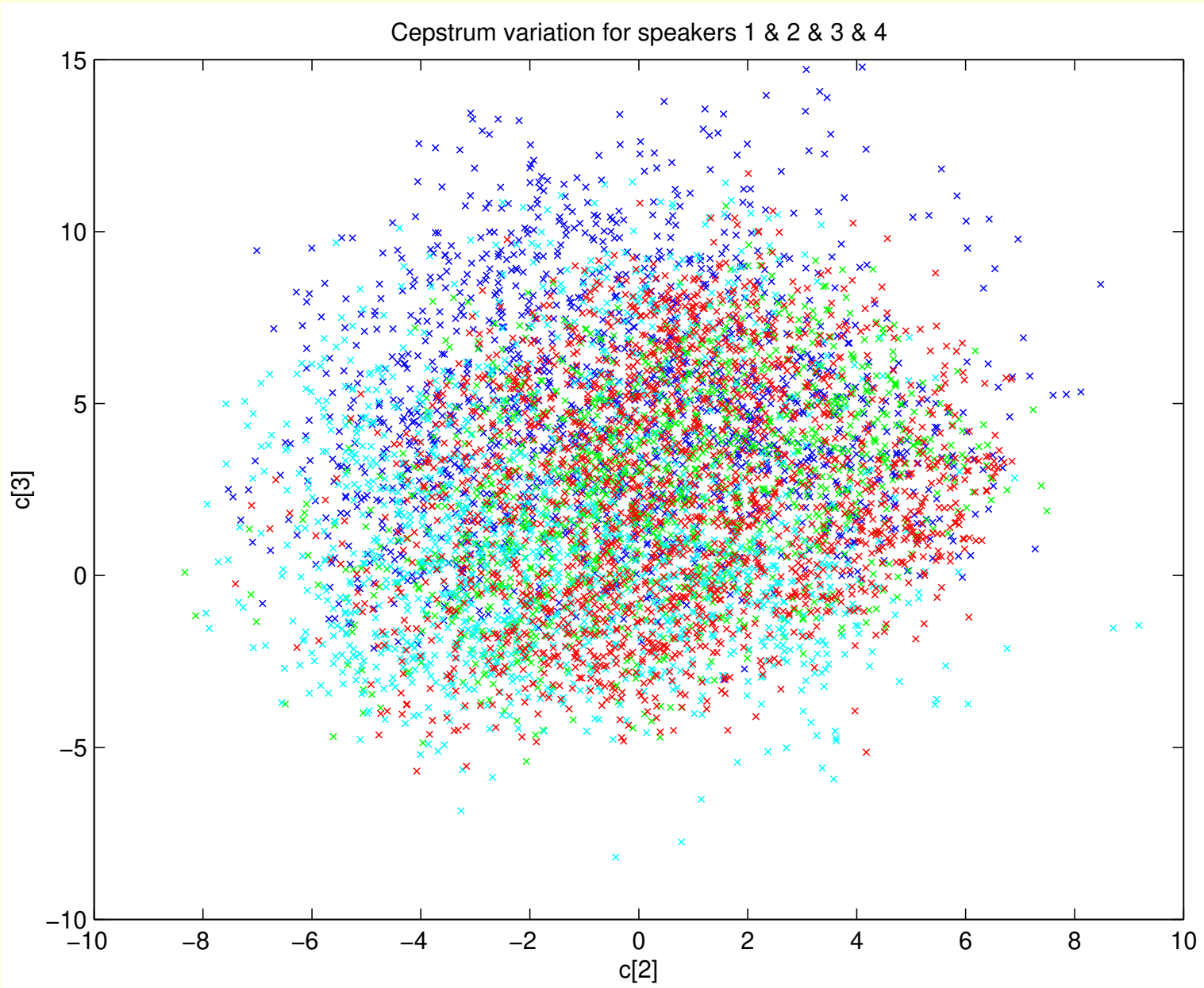
Možné reprezentace **hlasu/řečníka/promluvy:**

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

I. část

Textově nezávislá verifikace/identifikace na bázi VQ

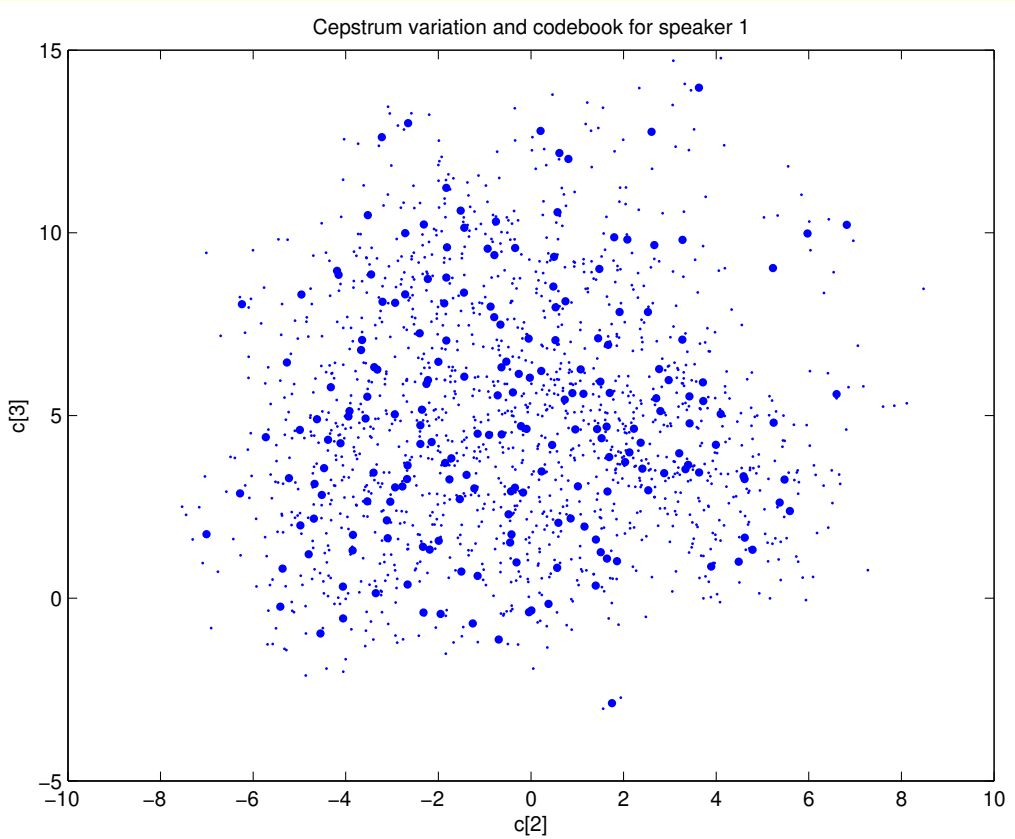
Rozložení kepstra ($c[2]$ vs. $c[3]$) - řečník 1 & 2 & 3 & 4



Variace kepstra a kódová kniha vybraného řečníka (VQ)

VQ - vektorová kvantizace

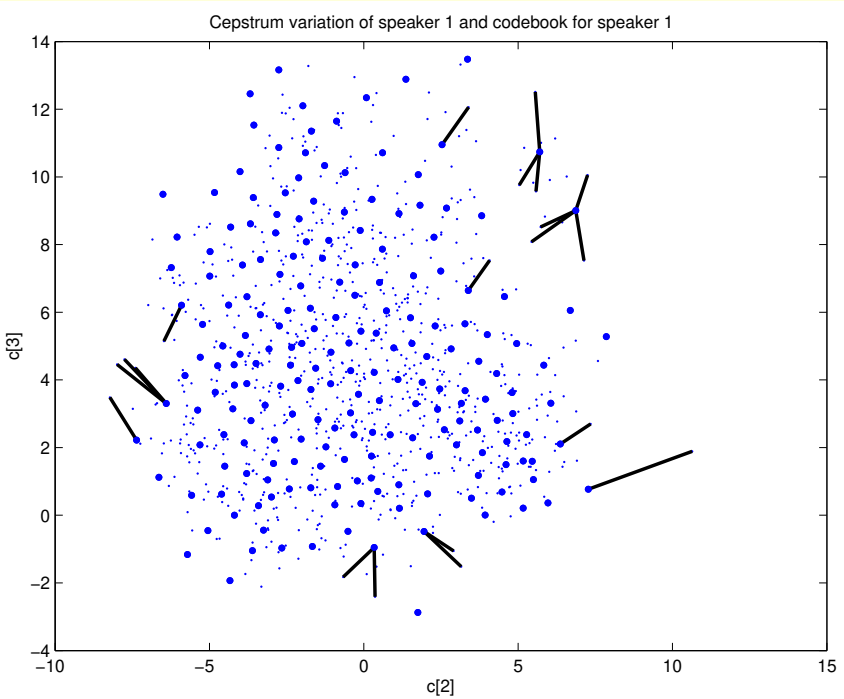
- redukce počtu (kvantování) uchovávaných příznakových vektorů
- rozložení příznaků je reprezentováno kódovou knihou - C^{spk} (konečný počet reprezentantů c_k^{spk} popisující variabilitu příznaků)
- výpočet na bázi K-means algoritmu
- odstranění neřečových segmentů, VAD (Voice Activity Detector)

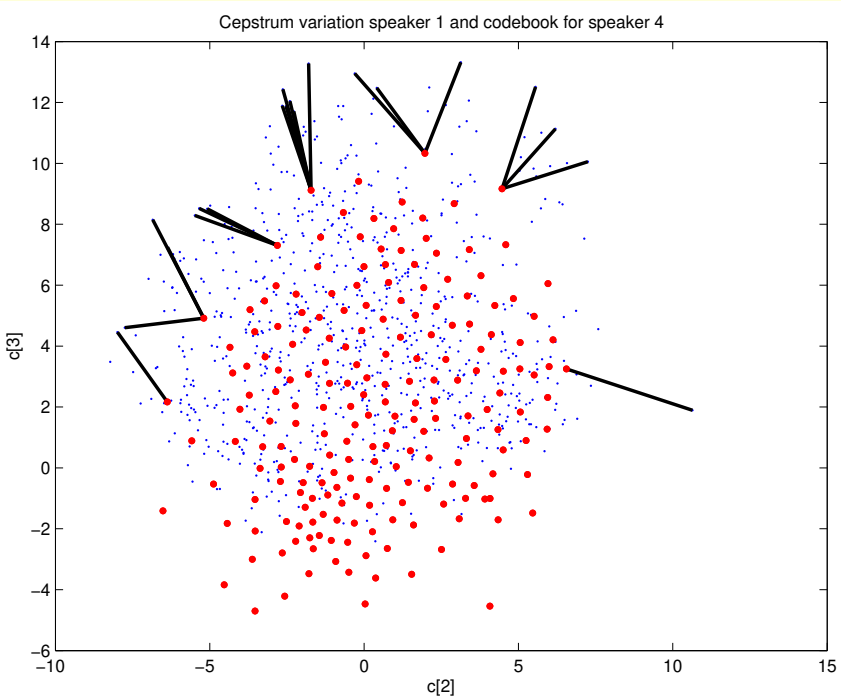


Realizace klasifikace na bázi VQ

Srovnání dvou mluvčích na bázi VQ

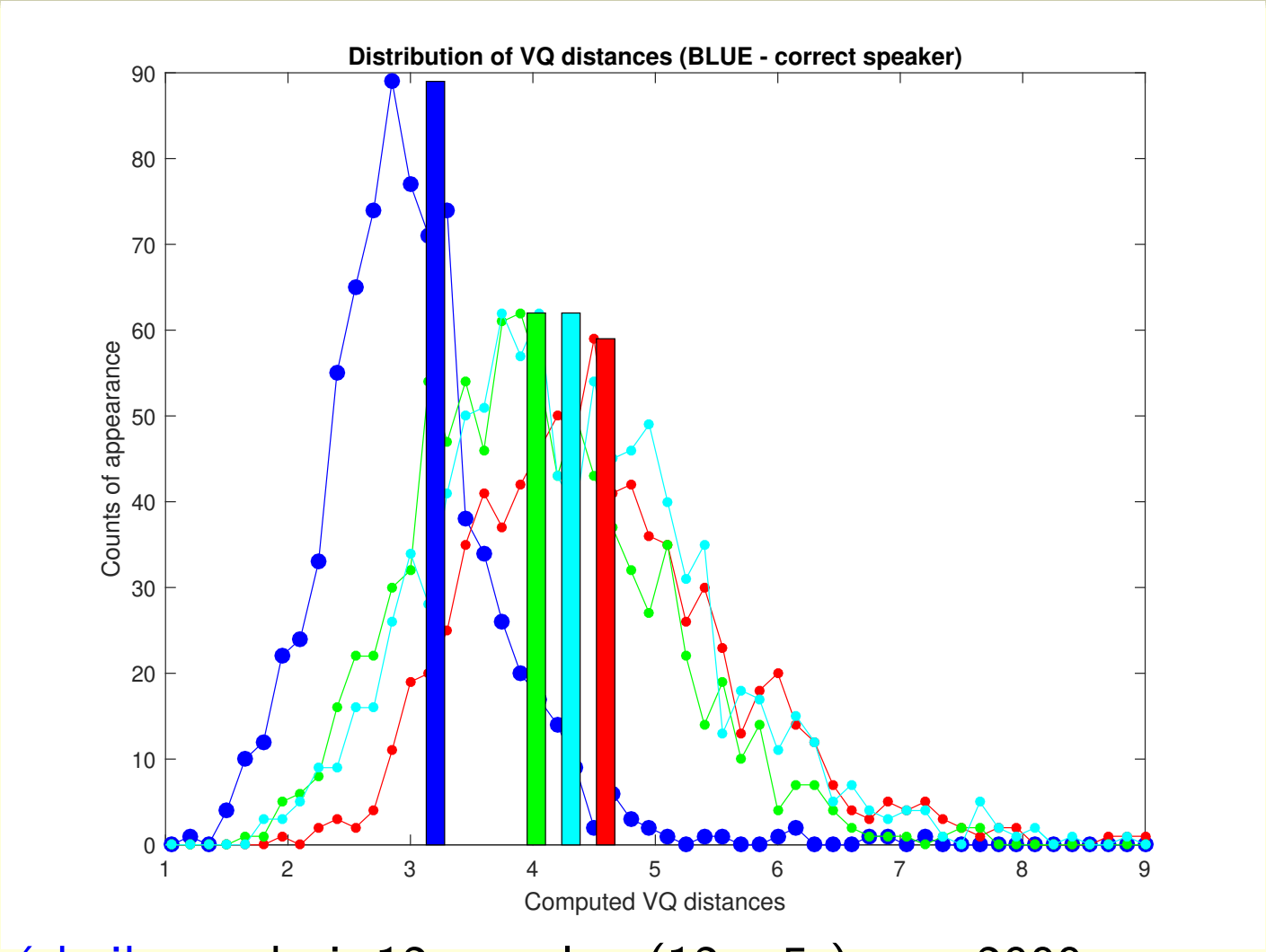
- měření průměrné vzdálenosti aktuálních příznakových vektorů od referenčních reprezentantů v kódové knize řečníka





$$score_{vq}^{spk} = \text{mean}\ (\ \text{CD}\ (\boldsymbol{c}_i, \underset{\boldsymbol{c}_k^{spk}}{\text{argmin}}\ \ \text{CD}\ (\boldsymbol{c}_i, \boldsymbol{c}_k^{spk})))$$

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 - $score_{vq}^{spk}$

II. část

Textově nezávislá verifikace/identifikace na bázi GMM

Rozpoznávání řečníka na bázi GMM

Statistika rozložení kepstra - **GMM** (Gaussian Mixture Model)

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

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

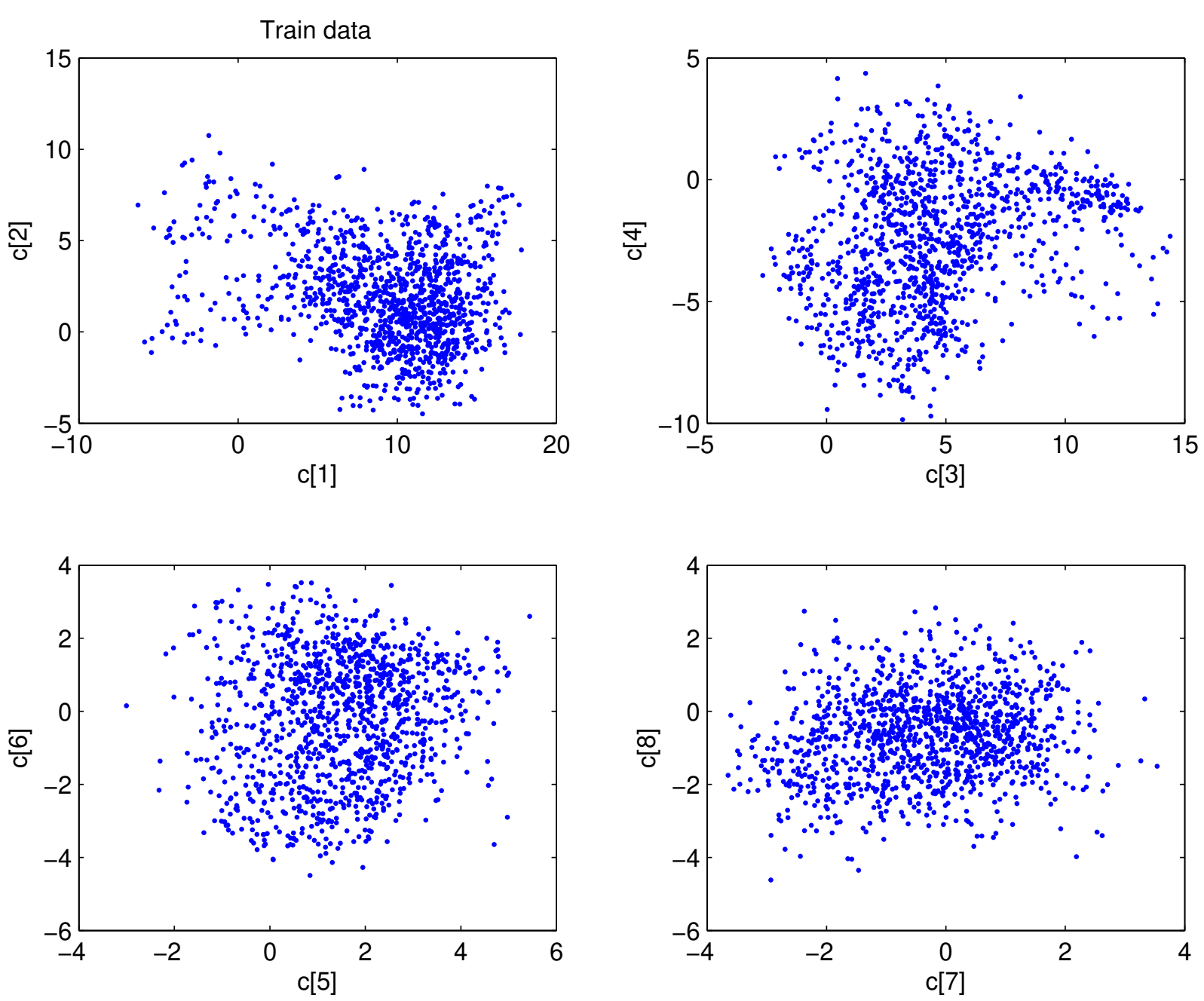
- λ^s ... GMM model pro řečníka *spk* s parametry:
  $\boldsymbol{\mu}_i^{spk}$ ... střední hodnota i-té složky
  $\mathbf{C}_i^{spk}$ ... kovarinační matice i-té složky
  c_i^{spk} ... váha i-té složky

- analogie ke kódové knize (střední hodnoty doplněny rozptyly)

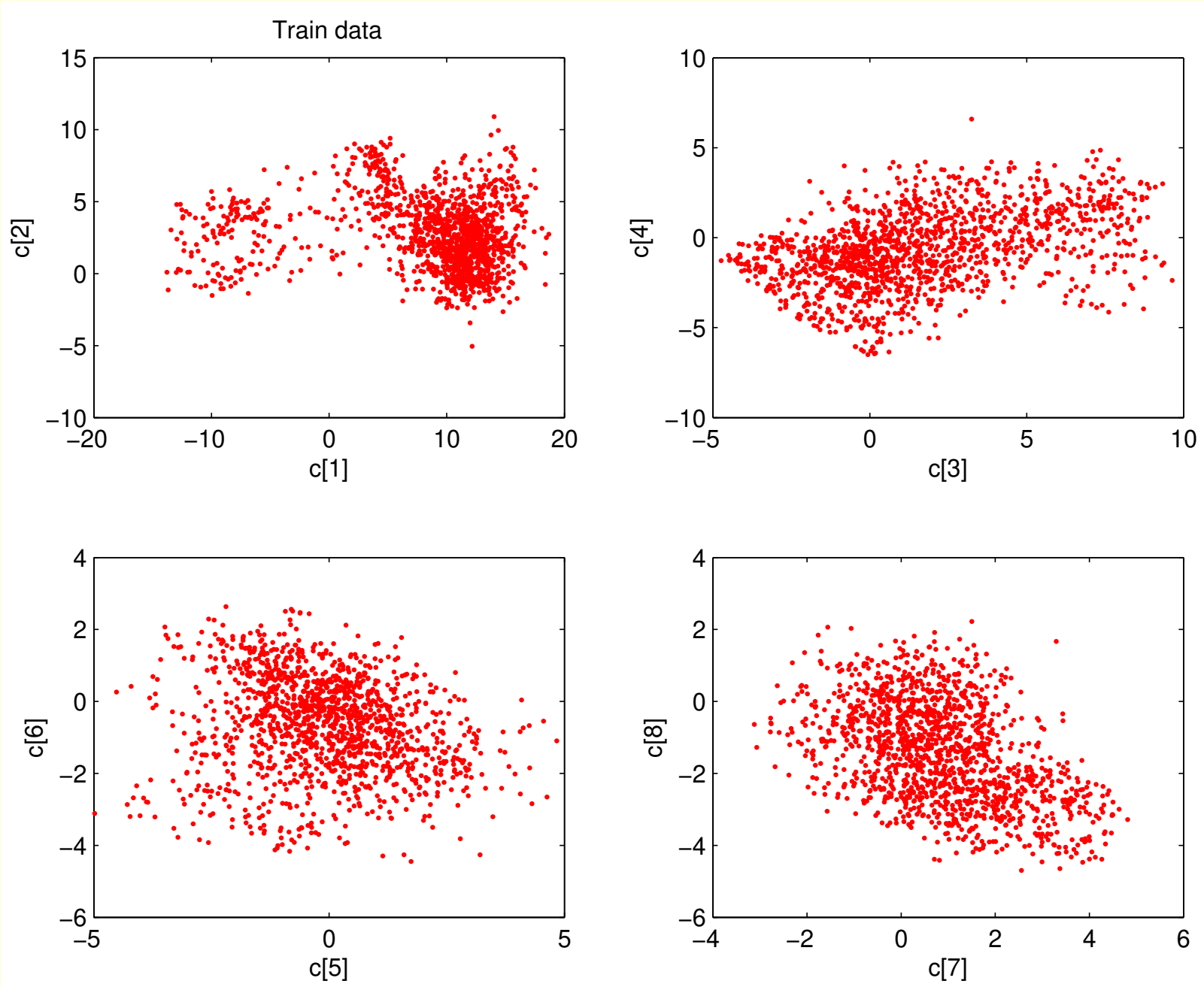
- více směsí modeluje variabilitu příznaků pro daného řečníka

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

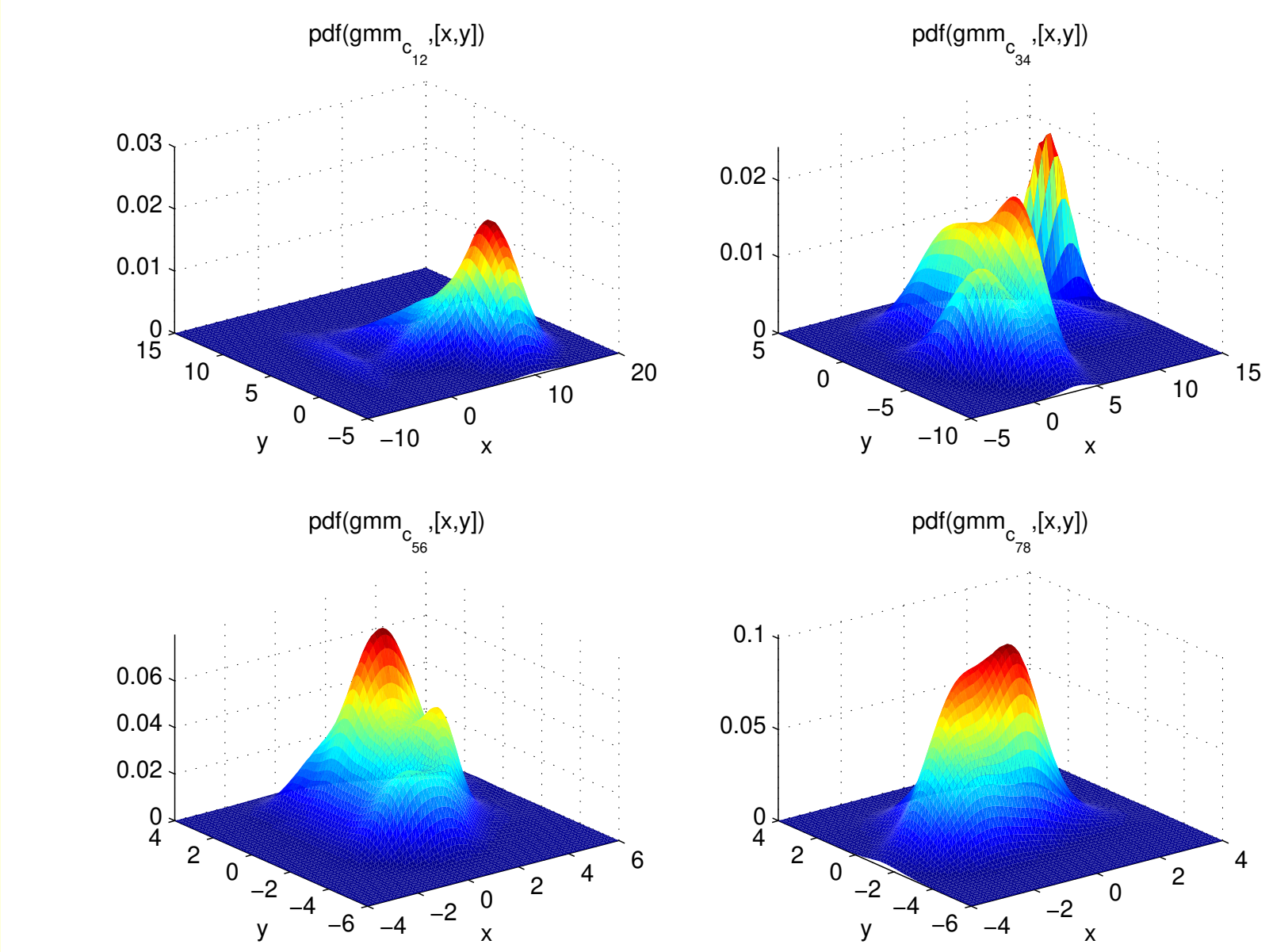
Rozložení prvků kódové knihy kepstra mluvčího A



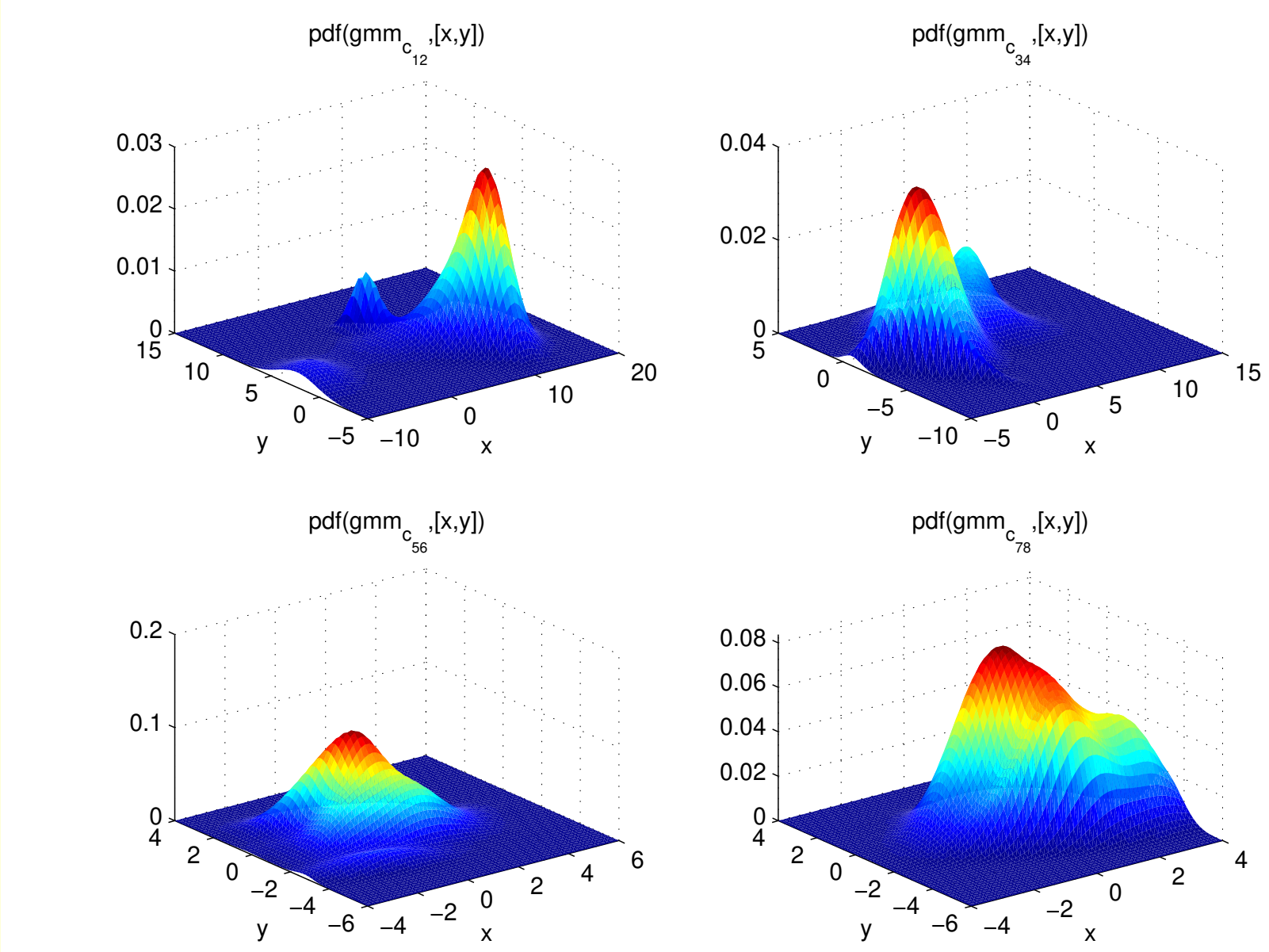
Rozložení prvků kódové knihy kepstra mluvčího B



GMM model rozložení kepstra mluvčího A



GMM model rozložení kepstra mluvčího B



Klasifikace na bázi GMM

Klasifikační míra : věrohodnost příznaku pro daný model

Věrohodnost se počítá z celé promluvy $\mathbf{O} = (\mathbf{o_1}, \mathbf{o_2}, ..., \mathbf{o_n})$
- pro všechny krátkodobého segmenty promluvy
- je třeba vypustit neřečové segmenty (aplikace VAD)

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

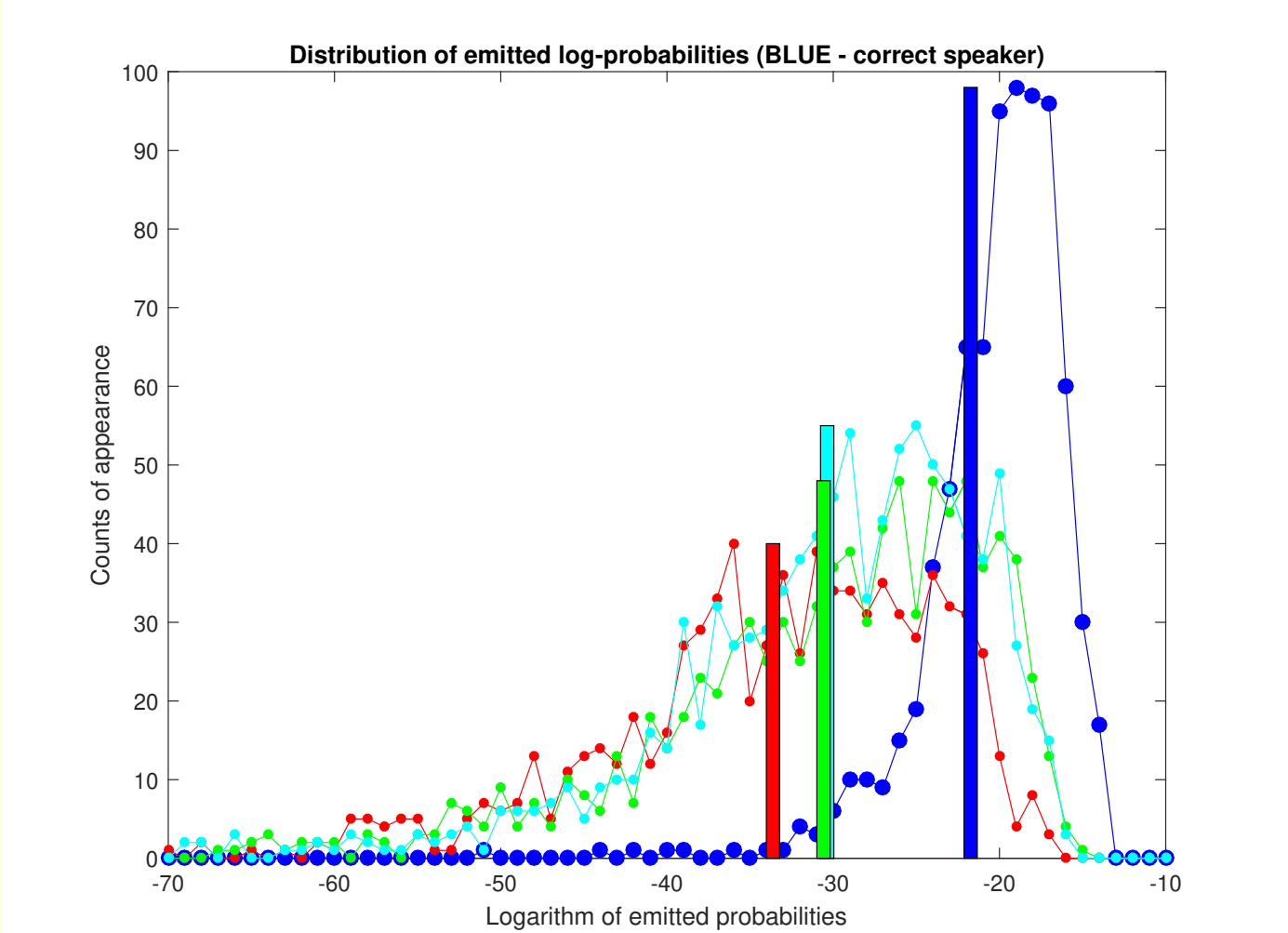
Logaritmická věrohodnost - součet logaritmů emitovaných pravděpodobností pro všechny krátkodobé realizace
(omezení možnosti podtečení)

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

Normovaná logaritmická věrohodnost na délku signálu
- průměrování emitovaných log. pravděpodobností

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

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

Problém s trénování GMM modelu pro jednotlivého mluvčího

málo dat $\rightarrow$ malá schopnost generalizace

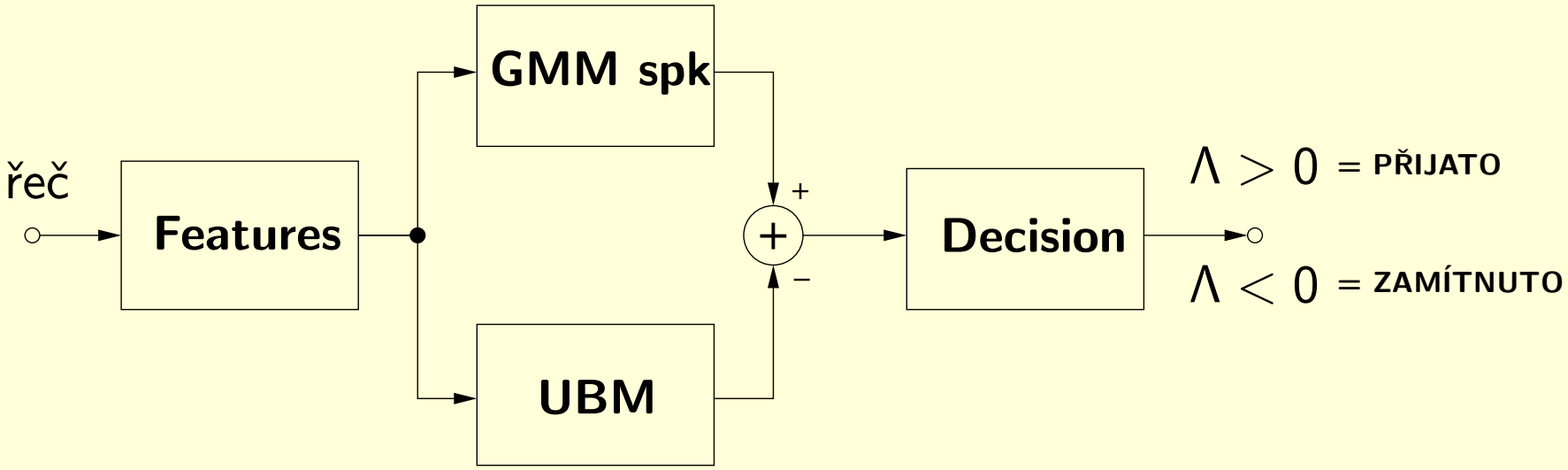
$\Downarrow$

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
    (typicky neexistujícího v množině pro trénování UBM)

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

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)

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

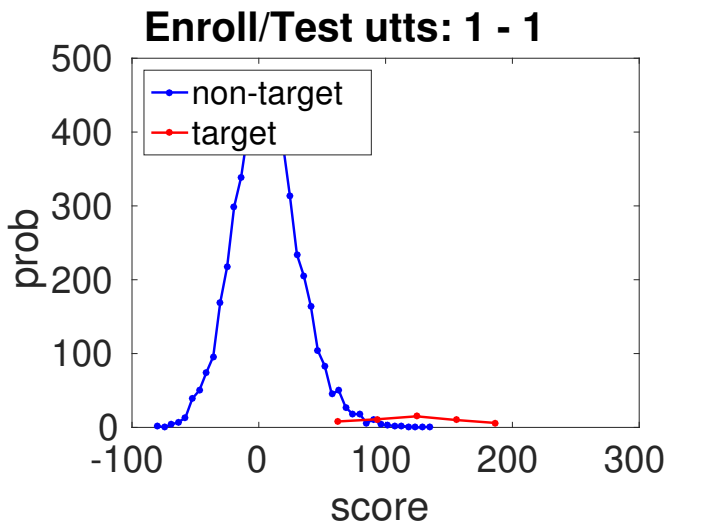
- variabilita akustických podmínek není explicitně modelována
- možnosti potlačení variability akustických podmínek :
  - LDA (nalezení podprostoru s optimální rozlišitelností tříd)
  - WCCN (normalizace kovariance uvnitř tříd)
  - PLDA - $\mathbf{x}_{r,s} = \boldsymbol{\mu} + \mathbf{V}\mathbf{y}_s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

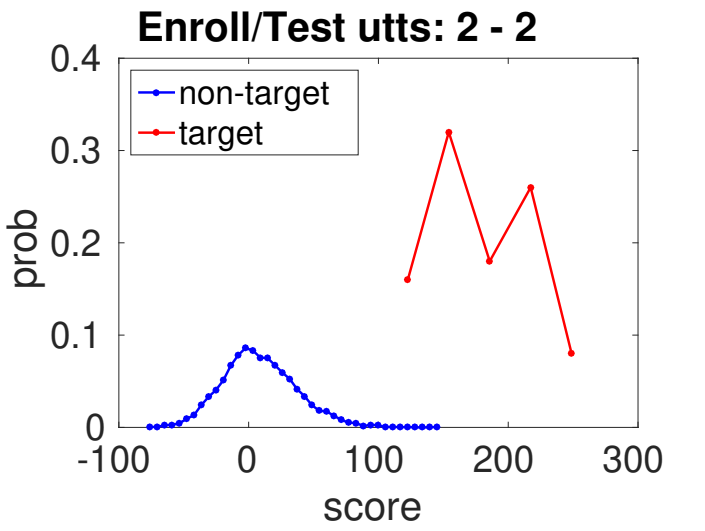
Příklad verifikace na bázi i-vektorů

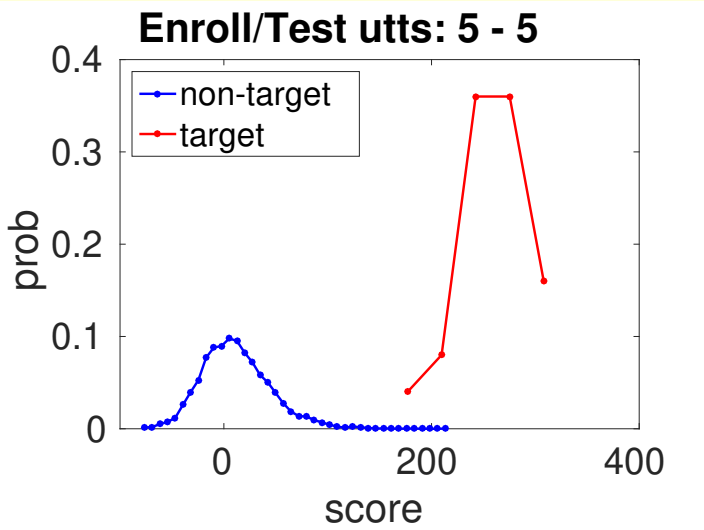
UBM : 50 spks, Enroll : 50 spks, Test : 102 spks

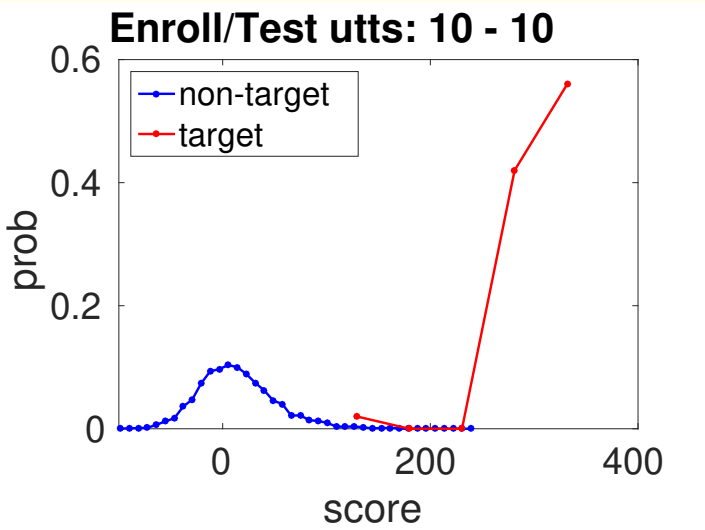
Skóre : kosinová vzdálenost

Varianty : různý počet enroll/test promluv (různá délka)









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:

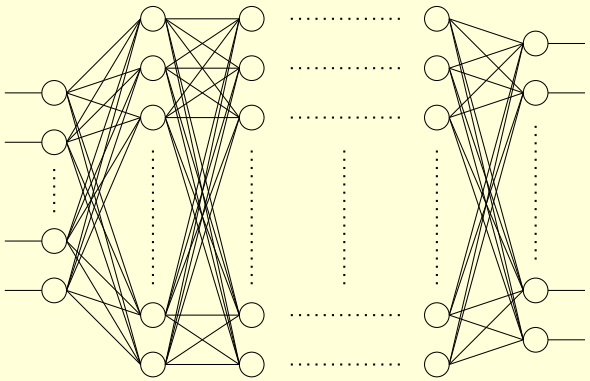
- možnost natrénování složitější funkce
- možné rozšíření příznakového vektoru
  (řetězení příznaků se širším kontextem)
- přesnější výsledky lze dosáhnout s DNN
  (vícevrstvé sítě s hlubokým učením - deep learning)
- speciální struktury sítí (RNN, TDNN, CNN, LSTM)

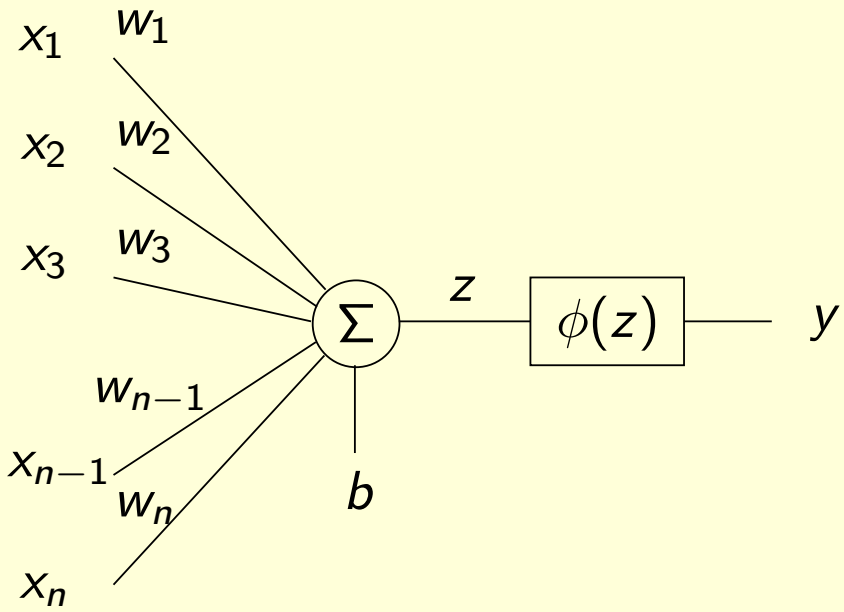
NEVÝHODY:

- 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)

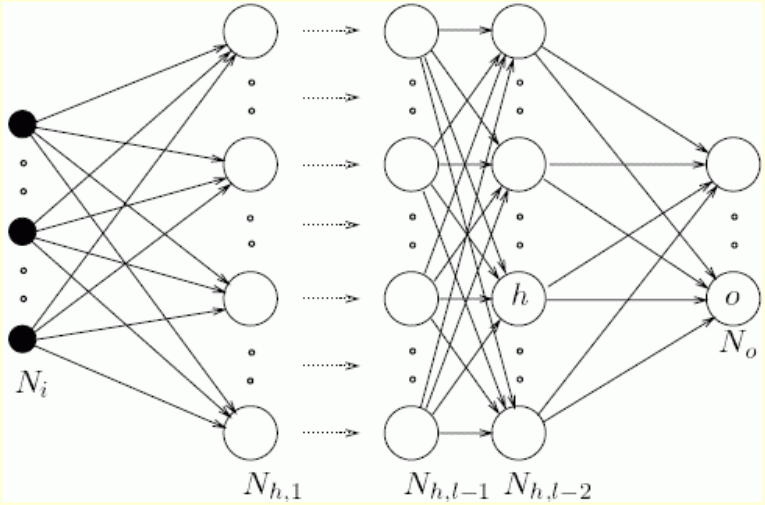
Inicializace sítě před trénování:

- 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

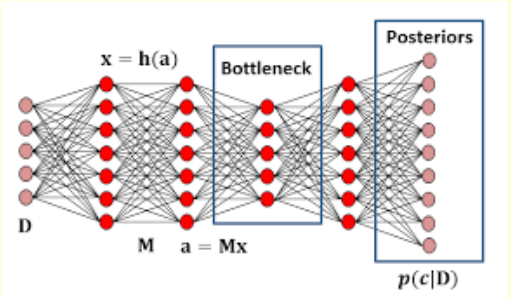


VSTUP: BF, kepstrum (MFCC), možný kontext několik oken

SKRYTÉ VRSTVY: 4-10

VÝSTUP: Softmax (aposteriors)

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

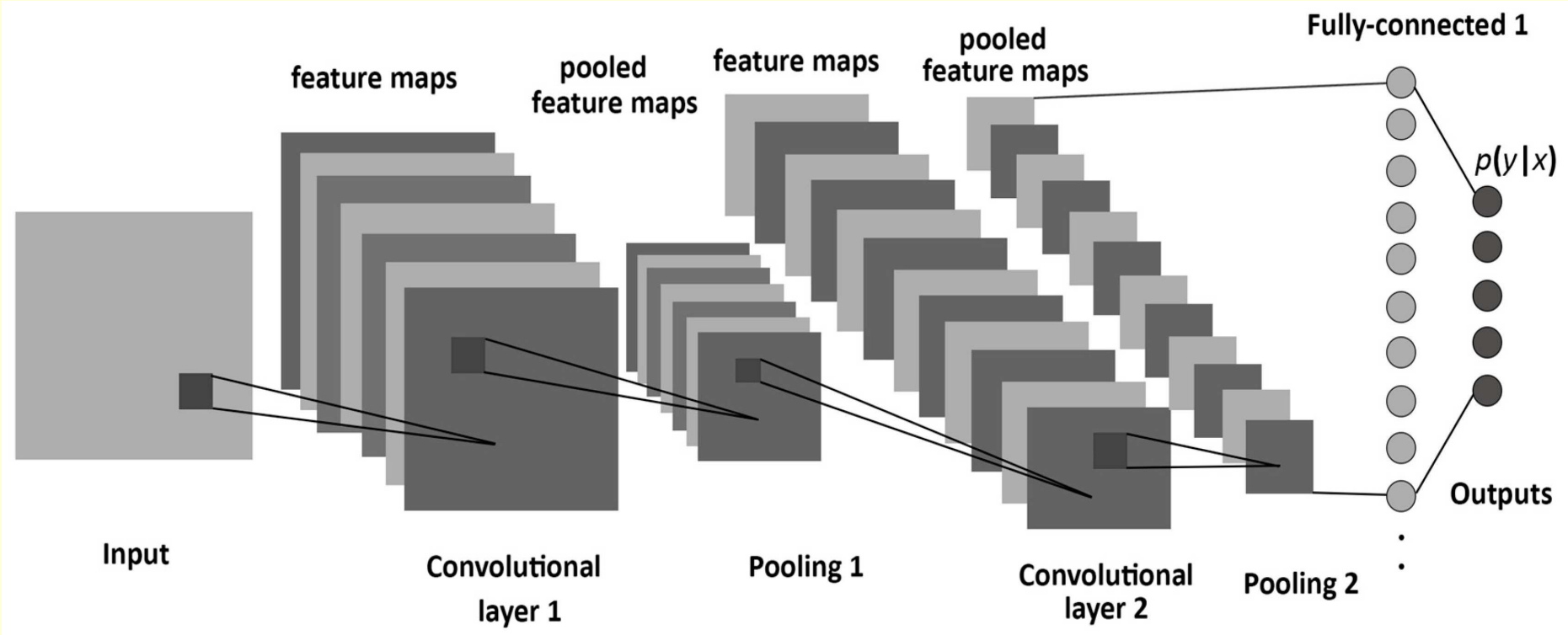


Nepřímé použití DNN: výstup bottleneck vrstvy + MFCC

→ příznaky pro i-vektorový systém

Přímá klasifikace na bázi CNN (Convolution Networks)

Principiální schéma CNN:

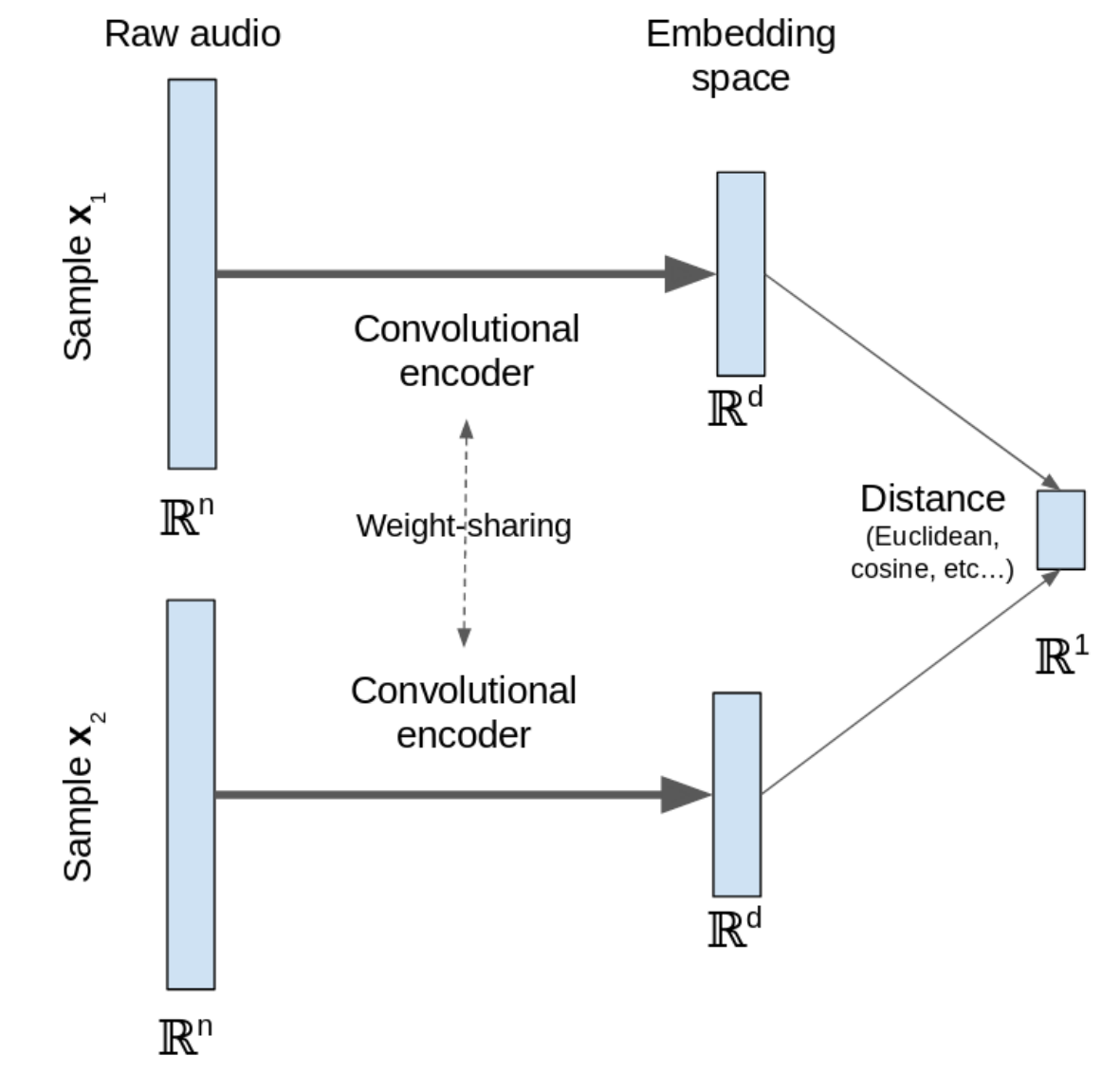


Nejčastější aplikace CNN ve zpracování obrázků

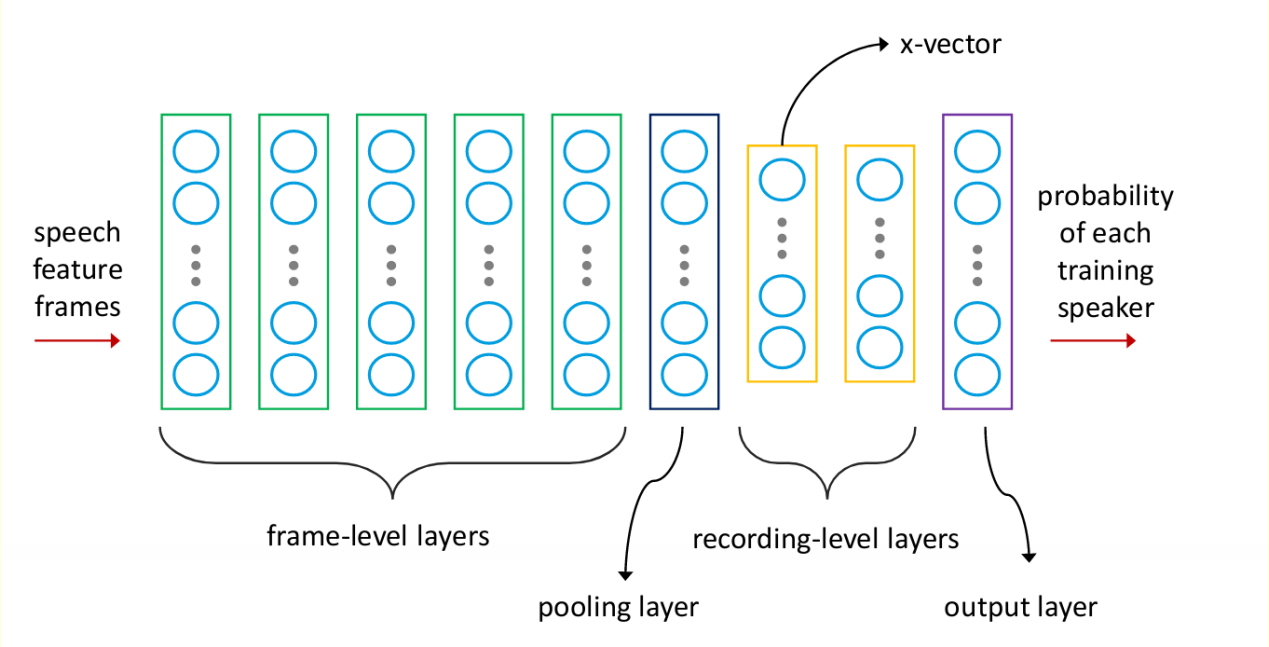
Aplikace pro SRE: vstupem je spektrogram (obrázek) či signál

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

CNN Siamese Speaker Verification

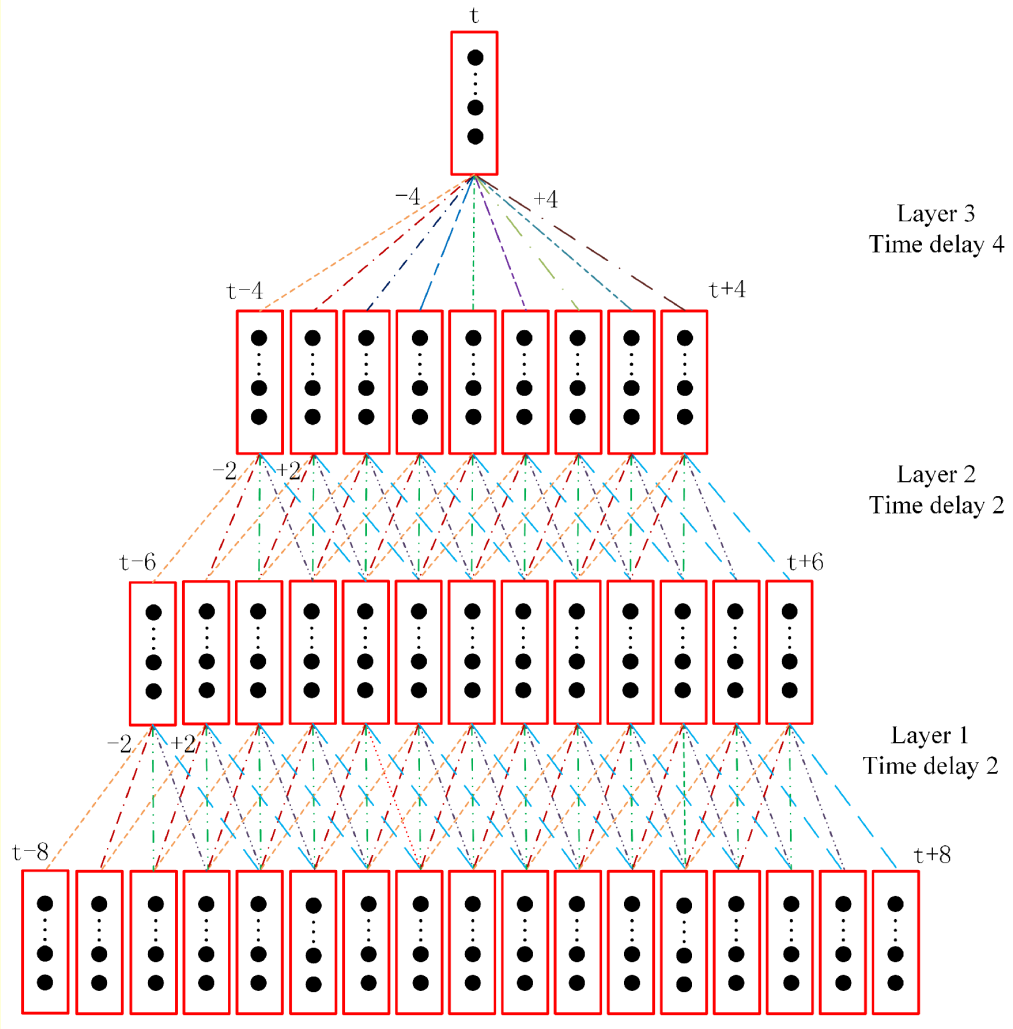


Aktuální DNN standard: systém na bázi x-vektorů

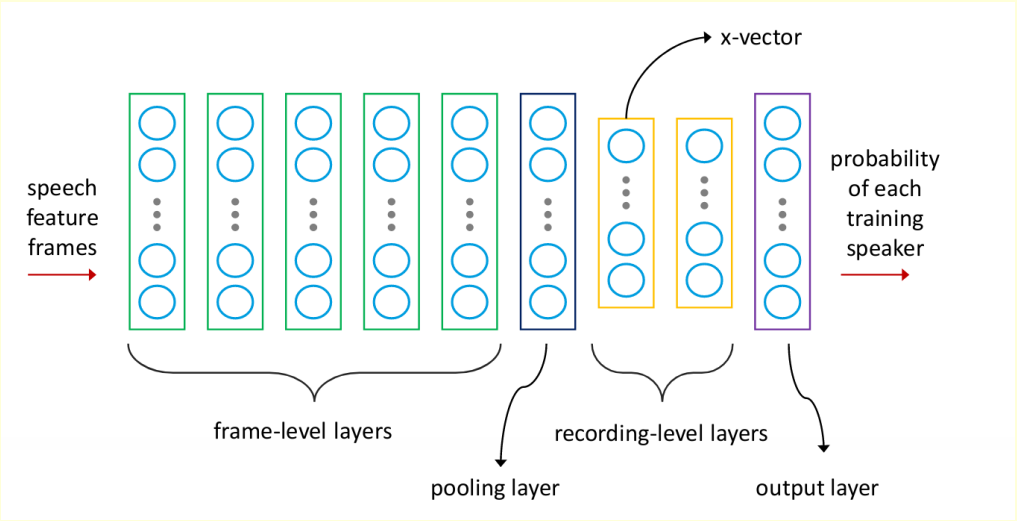


- vstupní příznaky jsou zpracovány v 5 TDNN vrstvách, zvyšující se zpoždění zahrnuje potřebnou kontextovou informaci (Δ příznaky nejsou používány)
- 6. poolingová vrstva počítá střední hodnoty a standardní odchylky výstupu 5. vrstvy přes všechny segmenty nahrávky
- dopředné (bottleneck) vrstvy 7 a 8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

TDNN vrstvy - kontextová informace



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

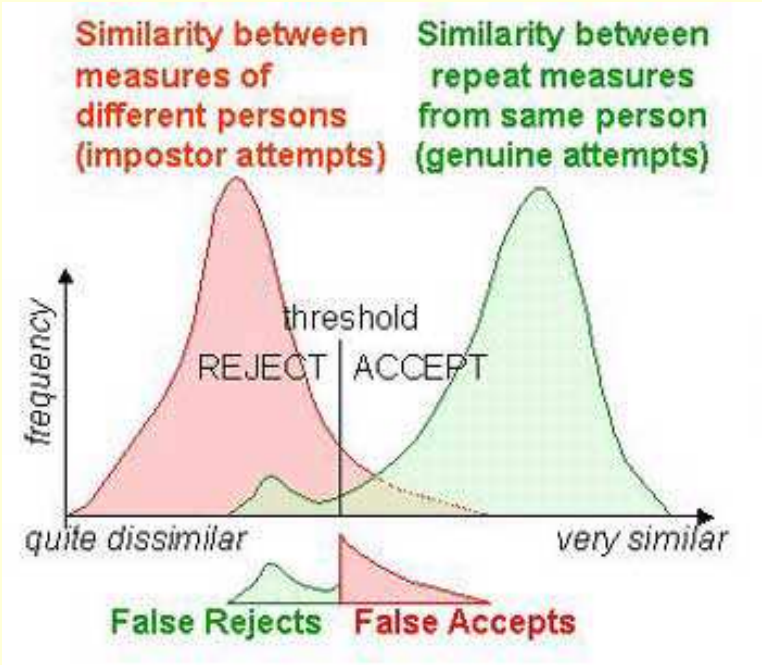
vstup 24 pásem melovské BF, pooling přes počet segmentů T, N řečníků

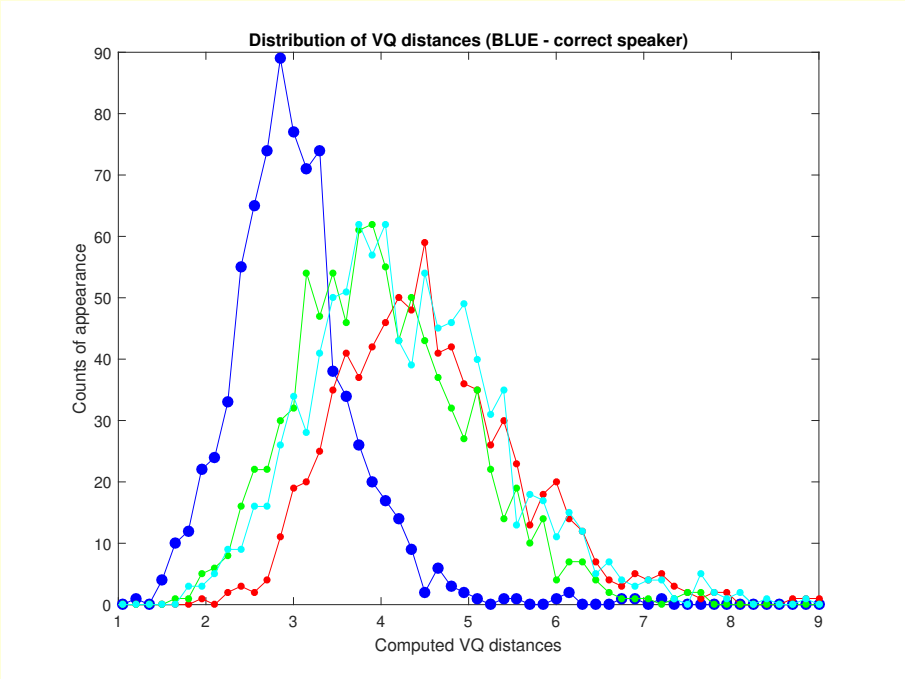
Více detailů viz: Snyder et al: X-Vectors: Robust DNN Embeddings for Speaker Recognition. In ICASSP 2018

V. část

Příklady systémů rozpoznávání řečníka

Hodnotící kritéria při verifikaci mluvčího - Míra stejné chyby



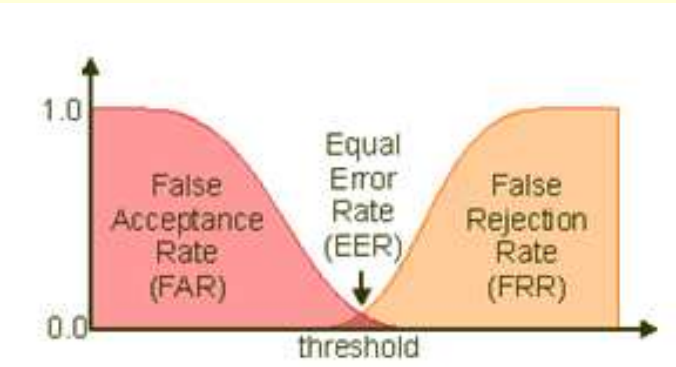


TA - True acceptance

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 :

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

Historické systémy 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)

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

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:

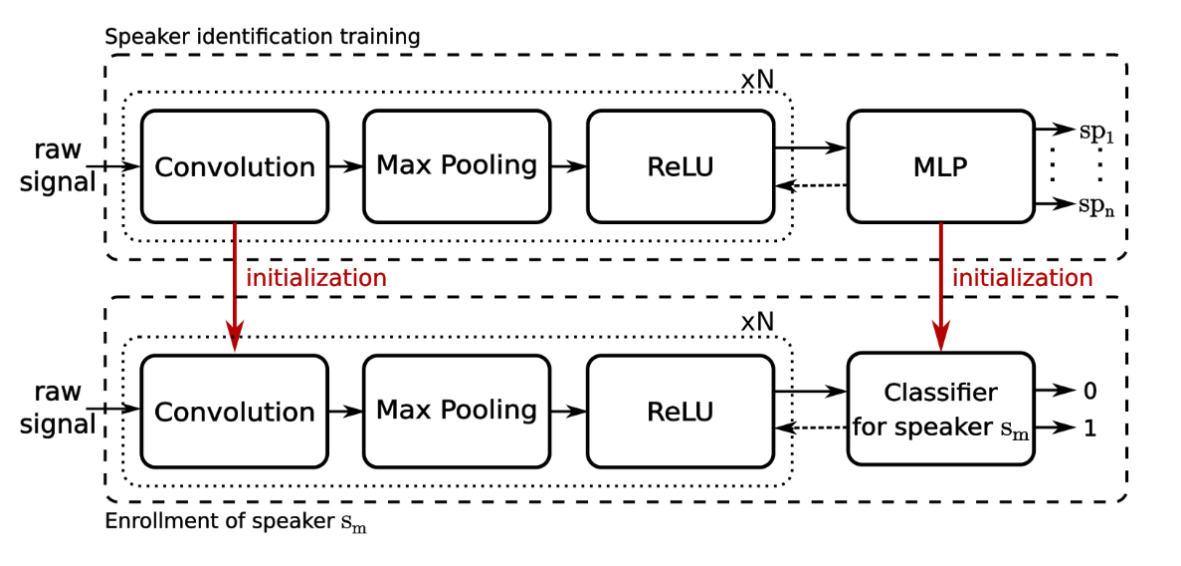
The IBM Speaker Recognition System
EER 2.11% - GMM-UBM (i-vector) - MFCC - LDA
EER 1.49% - GMM-UBM (i-vector) - MFCC - NDA
(NDA - Nearest-neighbour discriminant analysis)
EER 0.59% - DNN-fMLLR-NDA (English)

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 romluvy k verifikaci
- významná závislost na fonetickém kontextu
  (proto textově závislé i textově nezávislé úlohy)
- vetšina 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 !