

A6M33BIO - Biometrie

**Biometrické metody založené na
rozpoznávání hlasu II**

Doc. Ing. Petr Pollák, CSc.

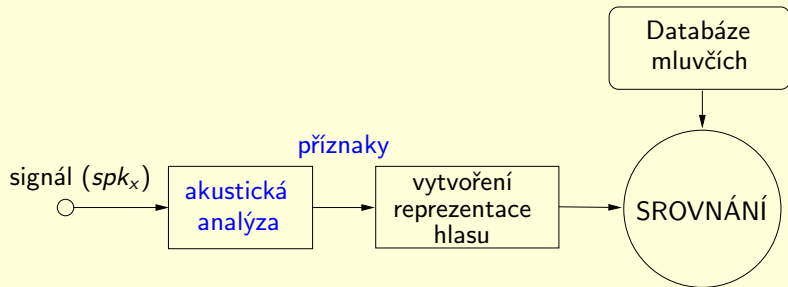
16. listopadu 2016 - 15:14

- **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

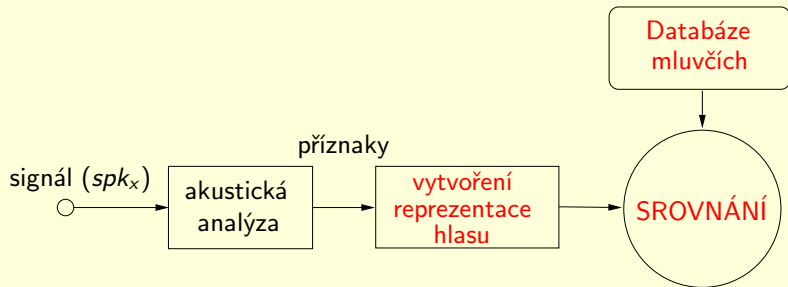
Metody automatického rozpoznávání mluvního



SHRNUTÍ - Používané příznaky:

- **MFCC** - obecně používané - textově nezávislé (možnost vyhlazení variability mezi mluvčími)
 - LPC keprální 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 mluvího



Používaná řešení :

- expertní rozhodování (fonetici, lingvisté)
- automatizované vyhodnocování

Základní úlohy automatického rozpoznávání mluvího :

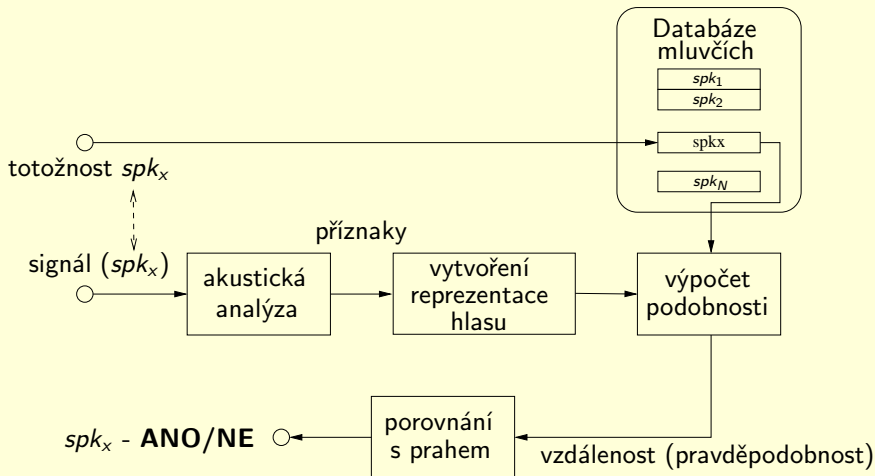
1 Verifikace mluvího

- ověření předpokládané totožnosti mluvího

2 Identifikace mluvího

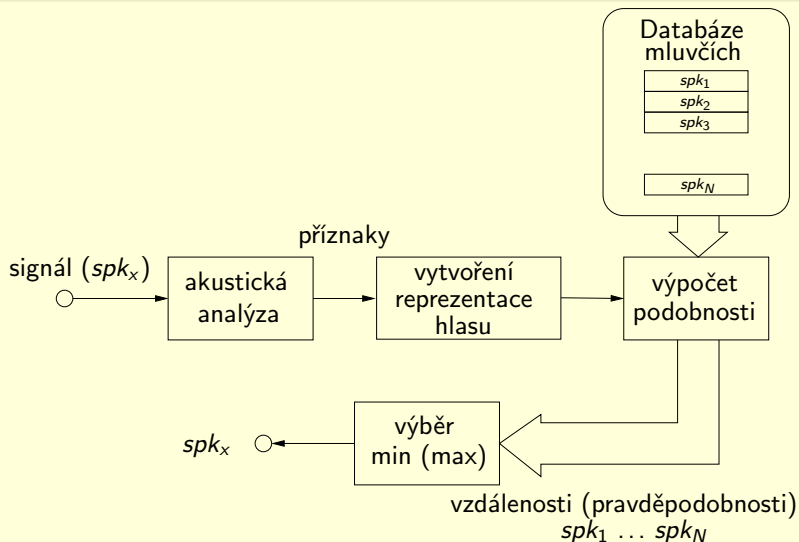
- **Identifikace v uzavřené množině**
rozpoznání neznámého mluvího z dané množiny mluvích
- **Identifikace v otevřené množině**
rozpoznání neznámého mluvího z neomezené množiny mluvích → identifikace & verifikace

Verifikace mluvího



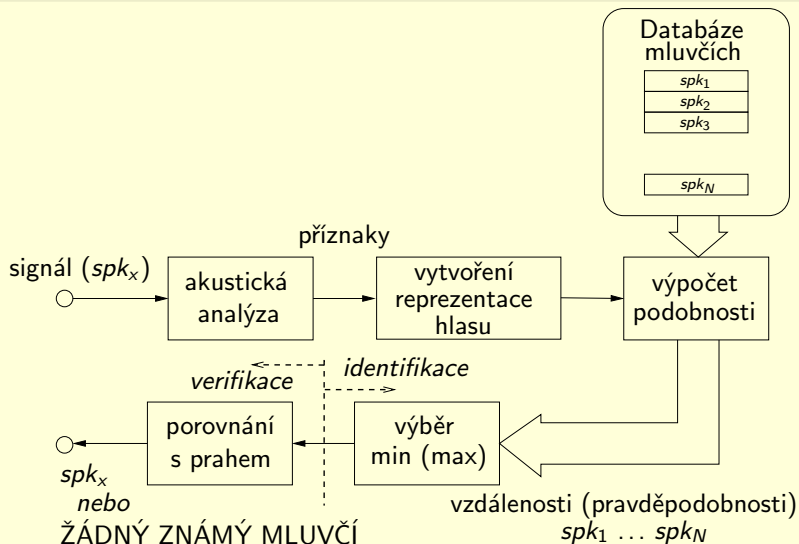
- ověření předpokládané totožnosti mluvího
- **VÝSLEDEK** = **přijetí** / **odmítnutí** předpokl. totožnosti

Identifikace mluvčího (v uzavřené množině)



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

Identifikace 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 mluvího 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 mluvího 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 mluvího
- výpočet pravděpodobnosti místo GMM

Rozpoznávání na základě časových funkcí

- vzorová promluva
- průběh energie - rytmizace promluvy
- průběh f_0 - 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

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

Vzdálenost mezi všemi segmenty (normovaná na délku):

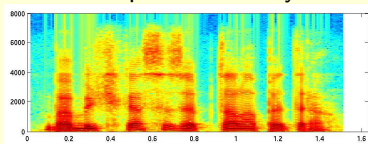
$$d_{i,j} = \frac{1}{NM} \sum_{k=1}^P (c_k[i] - \bar{c}_k[j])^2 \quad \text{pro } i = 1, \dots, N; j = 1, \dots, M$$

Kumulovaná vzdálenost algoritmem dynamického programování:

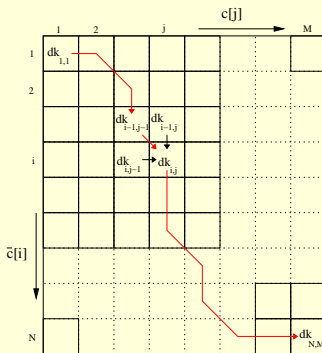
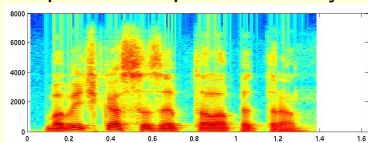
$$dk_{i,j} = \min (dk_{i-1,j} + d_{i,j}, dk_{i,j-1} + d_{i,j}, dk_{i-1,j-1} + d_{i,j}) \\ \text{pro } i = 1, \dots, N; j = 1, \dots, M$$

DTW - shrnující popis metody

Referenční promluva délky N



Rozpoznávaná promluva délky M



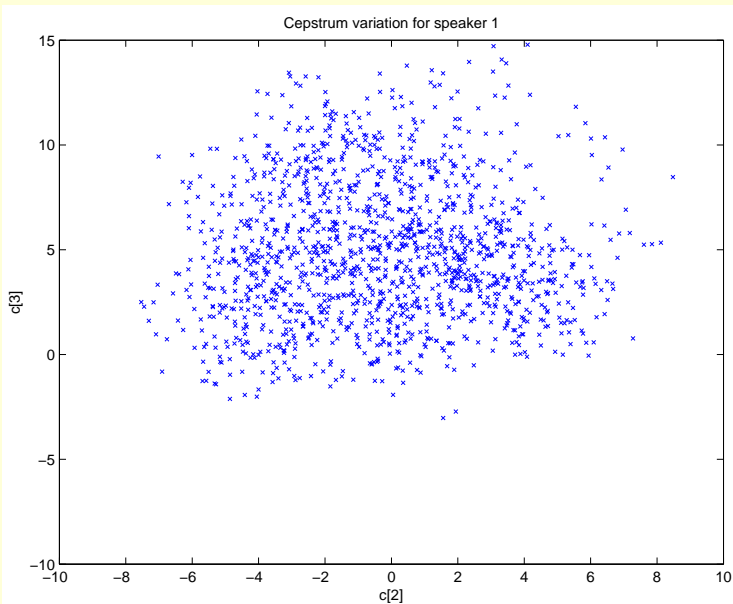
Identifikace řečníka \Rightarrow hledání minima $dk_{N,M}$ pro různé vzory

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

- jednoduchá koncepce, málo dat
- konkrétní časová funkce = textově závislá metoda

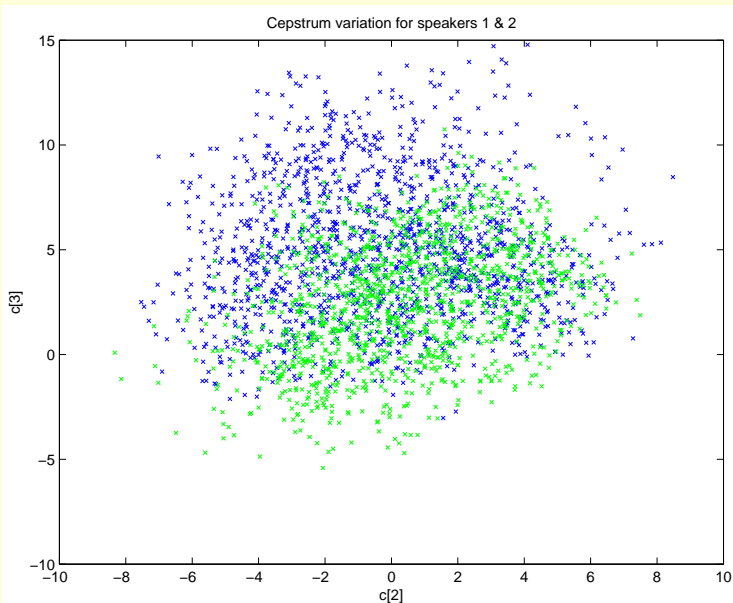
Klasifikace na bázi VQ

Rozložení kepstra řečníka - 1



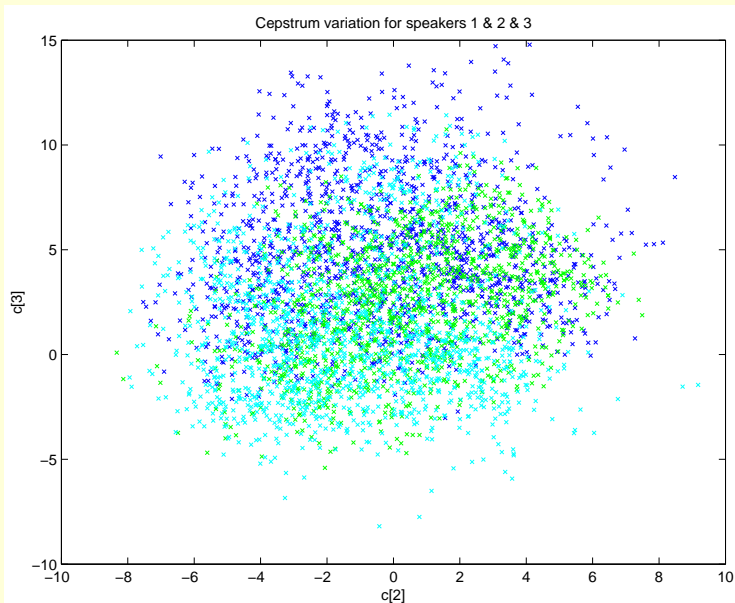
Klasifikace na bázi VQ

Rozložení kepstra řečníka - 1 & 2



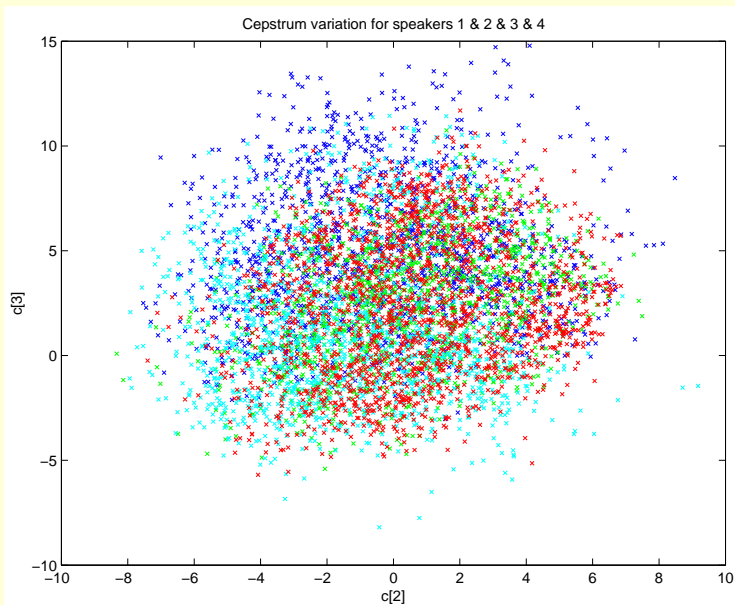
Klasifikace na bázi VQ

Rozložení kepstra řečníka - 1 & 2 & 3



Klasifikace na bázi VQ

Rozložení kepstra řečníka - 1 & 2 & 3 & 4



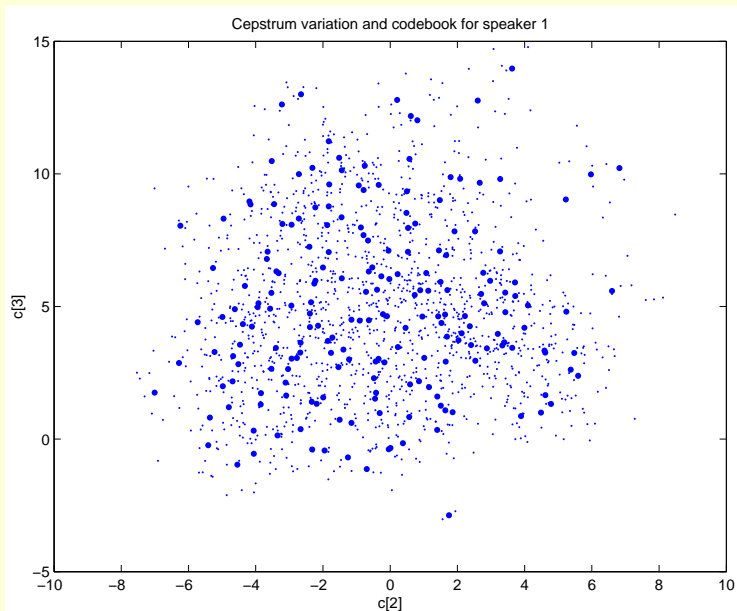
VQ - vektorová kvantizace

- kvantování obecných **příznakových vektorů**
 - rozložení příznaku je popsáno kódovou knihou
-

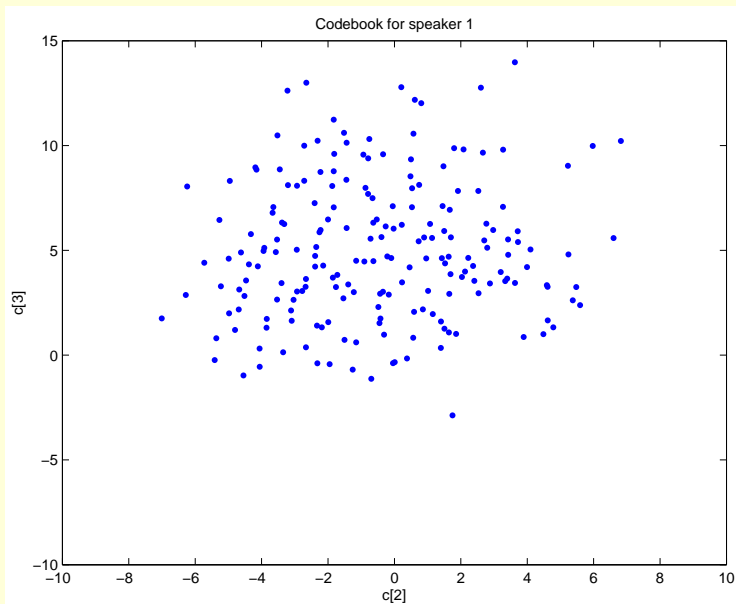
Kódová kniha - c_{VQ}^s ... **konečný počet reprezentantů**
variability příznaků
(výpočet na bázi K-means algoritmu)

VAD - vhodné použít pro **odstranění neřečových segmentů**
(energetický detektor může být postačující)

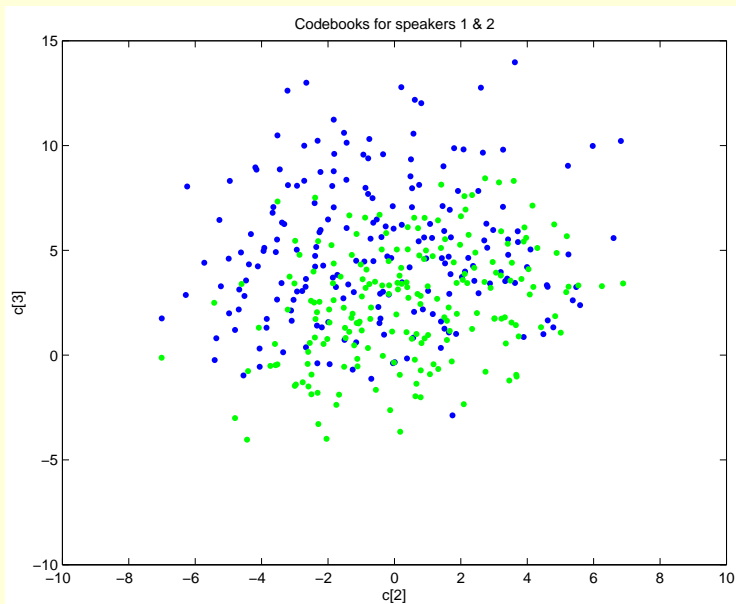
Variace kepstra a kódová kniha vybraného řečníka



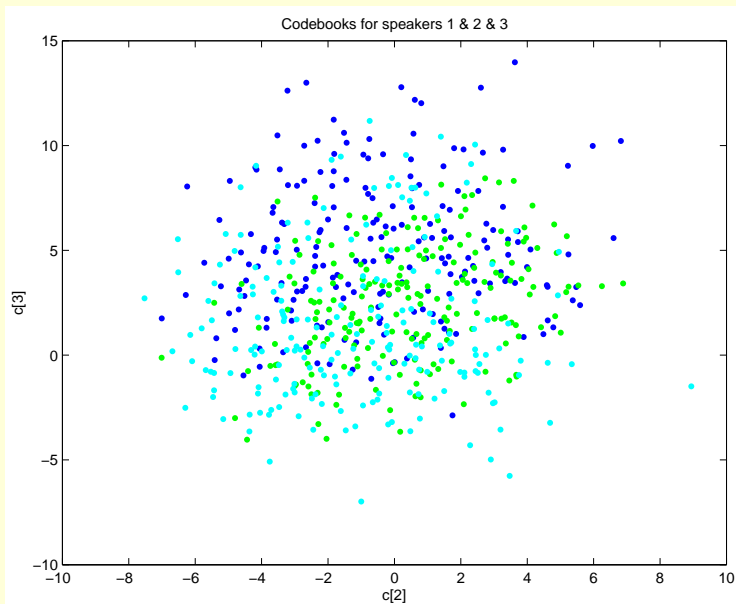
Kódová kniha řečníka - 1



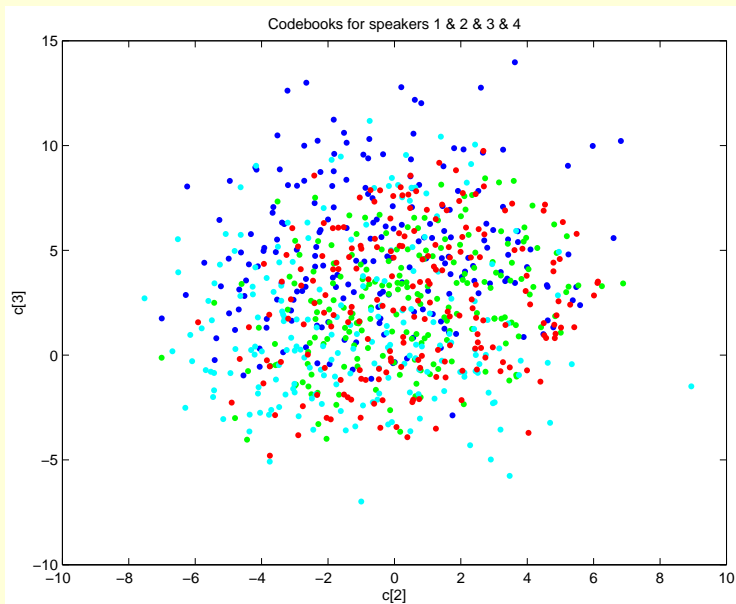
Kódová kniha řečníka - 1 & 2



Kódová kniha řečníka - 1 & 2 & 3

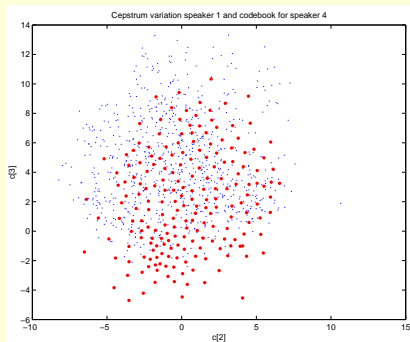
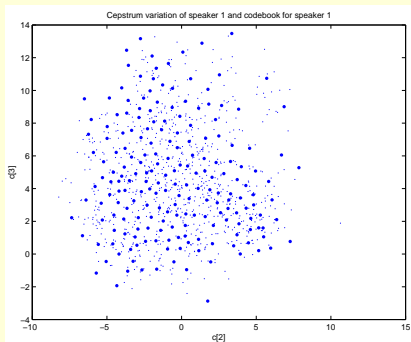


Kódová kniha řečníka - 1 & 2 & 3 & 4



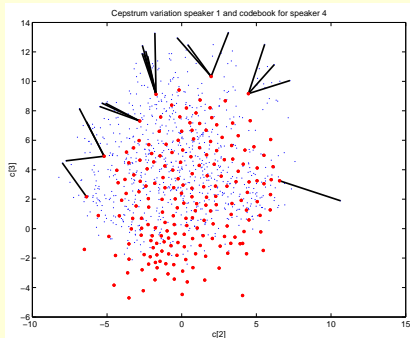
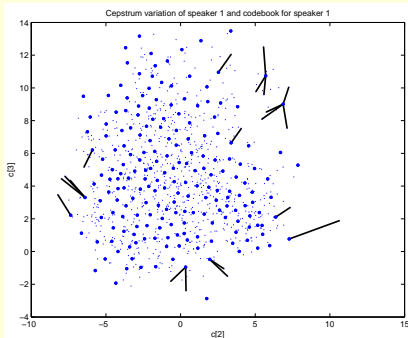
klasifikace na bázi VQ

- měření průměrné vzdálenosti aktuálních příznakových vektorů od uložených typických reprezentantů



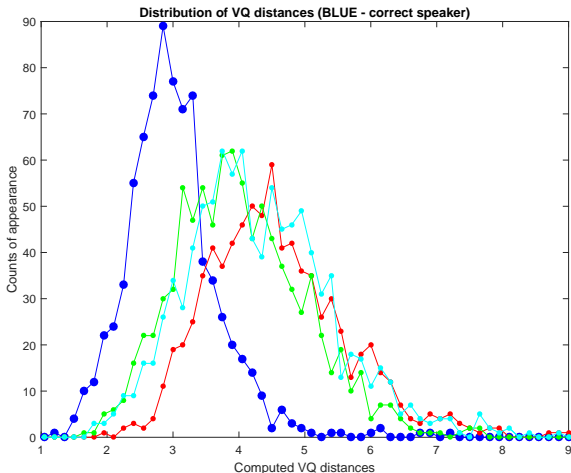
klasifikace na bázi VQ

- měření průměrné vzdálenosti aktuálních příznakových vektorů od uložených typických reprezentantů



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

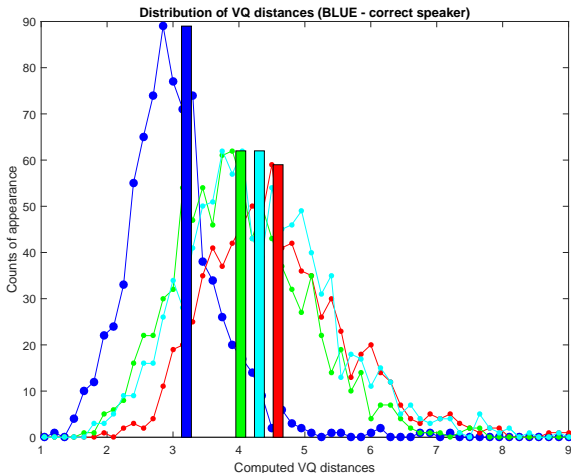
Statistiky výsledků pro 4 řečníky a 1 kódovou knihu



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

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

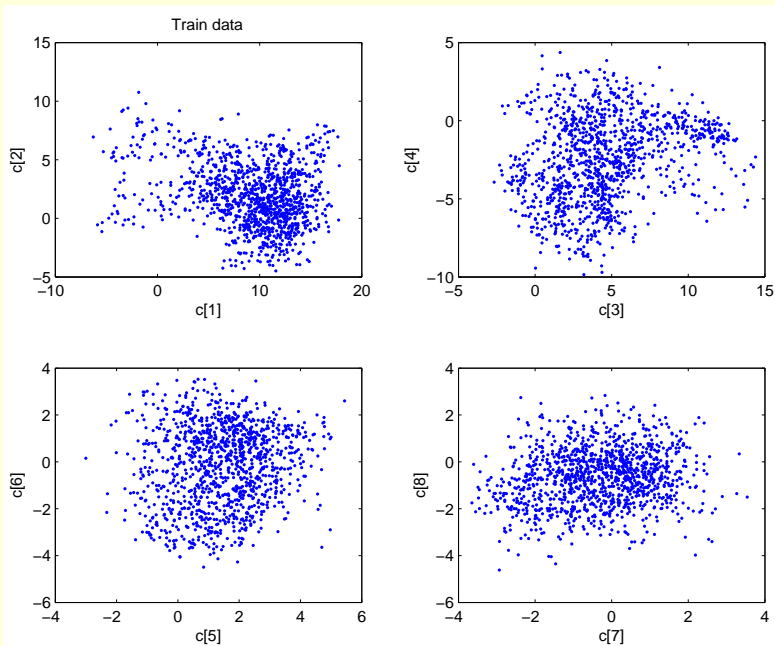
Průměrné hodnoty vzdálenosti

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

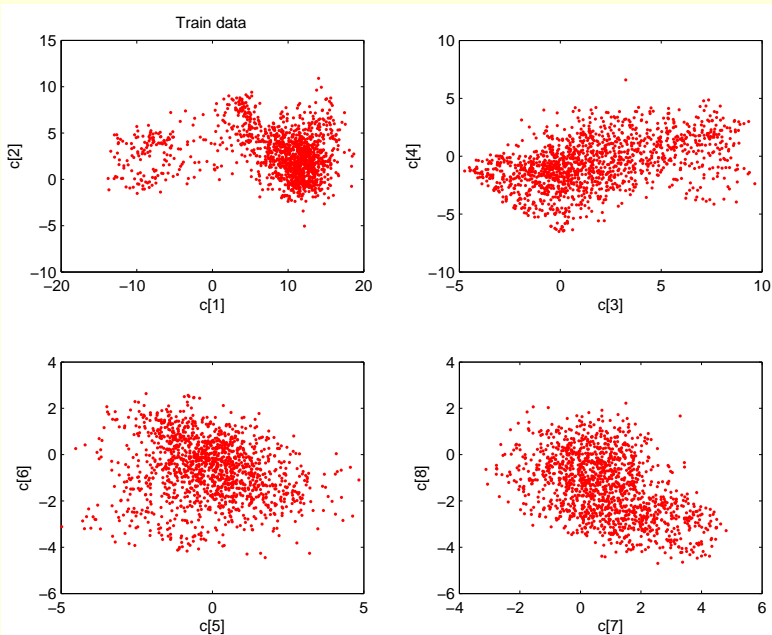
$$p(\mathbf{o}|\lambda^s) = \sum_{i=1}^{M_s} c_i^s \cdot \mathcal{N}(\mathbf{o}, \boldsymbol{\mu}_i^s, \mathbf{C}_i^s)$$

- $\mathcal{N}(\mathbf{o}, \boldsymbol{\mu}, \mathbf{C})$... N -rozměrná gaussovská funkce daná vektorem středních hodnot $\boldsymbol{\mu}$ a kovarianční maticí \mathbf{C}
- více směsí modeluje lépe variabilitu příznaků pro daného řečníka
- typické počty směsí: 8-256 (model řečníka), 512-2048 (univerzální model) - *počty směsí závisí na množství trénovacích dat*

Rozložení prvků kódové knihy kepstra mluvčího A

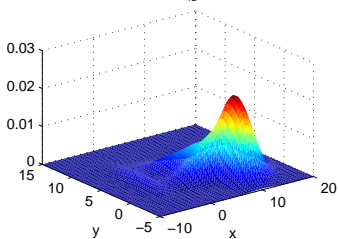


Rozložení prvků kódové knihy kepstra mluvčího B

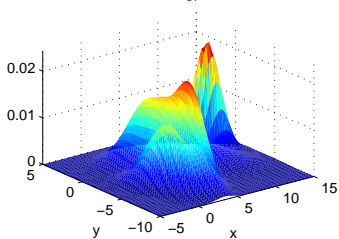


GMM model rozložení kepstra mluvčího A

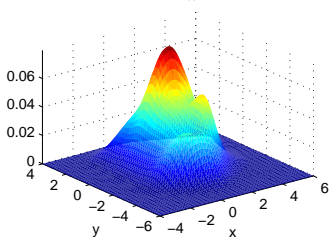
pdf($\text{gmm}_{c_{12}}$, [x,y])



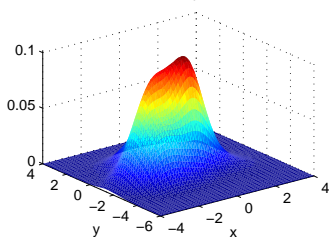
pdf($\text{gmm}_{c_{34}}$, [x,y])



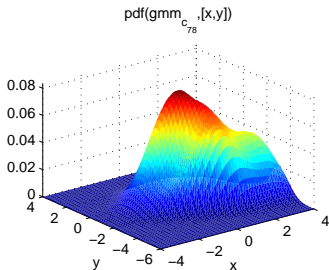
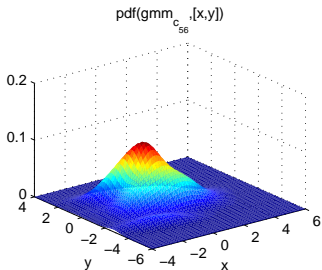
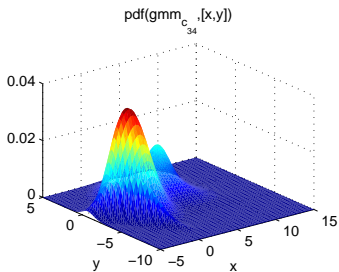
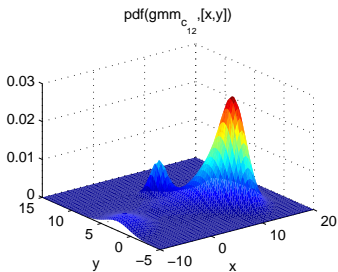
pdf($\text{gmm}_{c_{56}}$, [x,y])



pdf($\text{gmm}_{c_{78}}$, [x,y])



GMM model rozložení kepstra mluvčího B



Klasifikační míra:

→ **věrohodnost příznaku pro daný model** - $p(\mathbf{o}_j|\lambda^s)$

- hodnota věrohodnosti se počítá z celé promluvy

$$\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_n)$$

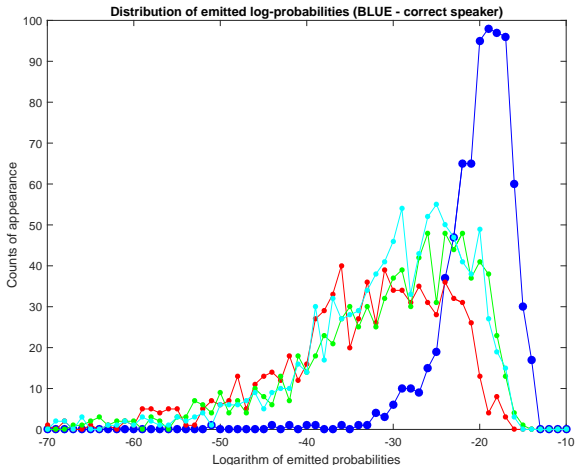
$$P(\mathbf{O}|\lambda^s) = \prod_{j=1}^N p(\mathbf{o}_j|\lambda^s)$$

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

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

- vhodné aplikovat detektor řeči pro vyřazení neřečových úseků

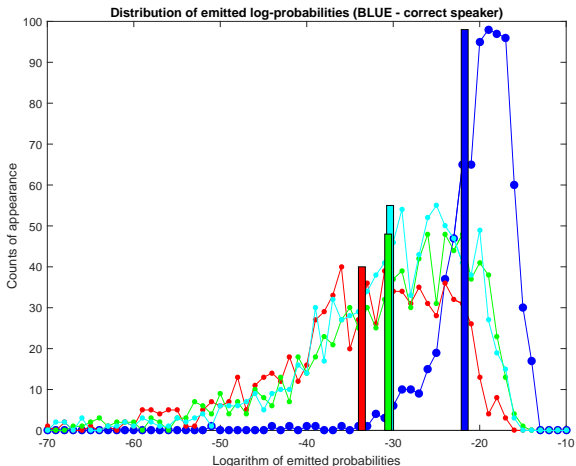
Statistiky výsledků pro 4 řečníky a 1 GMM model



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

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

Trénování GMM modelu pro jednotlivého mluvčího je problematické:

- málo dat \rightarrow malá schopnost generalizace

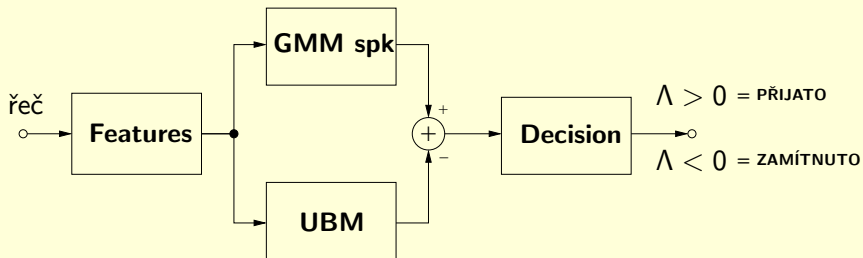


UBM-GMM modelování

- UBM - Universal Background Model
(generalizující model popisující společný prostor parametrů pro všechny mluvčí)
- GMM model mluvčího se získává adaptací UBM
 - nejčastěji MAP (Maximum A posteriori Probability)
 - optimální pro větší množství adaptačních dat
 - adaptují se jenom "viděné" parametry
- 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

\rightarrow **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)

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

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

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

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

- 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 - $\mathbf{x}_{r,s} = \boldsymbol{\mu} + \mathbf{V}\mathbf{y}_s + \mathbf{U}\mathbf{w}_{r,s} + \boldsymbol{\epsilon}_{r,s}$
(modelování variability akustických podmínek)

- **identifikace**

$$spk = \underset{s}{\operatorname{argmin}} \operatorname{dist}_s \quad \text{pro } s = 1, 2, \dots, L$$

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

- **verifikace**

$\operatorname{dist}_s < \operatorname{dist}_{thr}$ předpokládaná identita PŘIJATA

$\operatorname{dist}_s > \operatorname{dist}_{thr}$ předpokládaná identita ZAMÍTNUTA

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

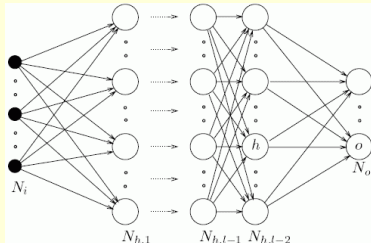
$P(\mathbf{O}|\lambda^s) < P_{thr}$ předpokládaná identita ZAMÍTNUTA

Použití ANN (DNN)

- ANN (DNN) se používají pro výpočet pravděpodobnosti místo GMM či přímo pro klasifikaci
- možnost natrénování složitější funkce
- přesnější pro DNN s více skrytými vrstvami (deep learning)
- náročnější na trénování, zejména z hlediska množství dat (nutnost nastavení mnoha vnitřních parametrů sítě)

Přímá klasifikace pomocí DNN

- DNN ve funkci odhadu aposteriori pravděpodobnosti řečníka

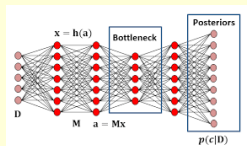


VSTUP: BF, kepstrum (MFCC),
možný kontext několik oken

SKRYTÉ VRSTVY: 4-10

VÝSTUP: Softmax (aposteriors)

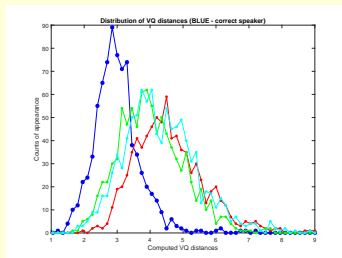
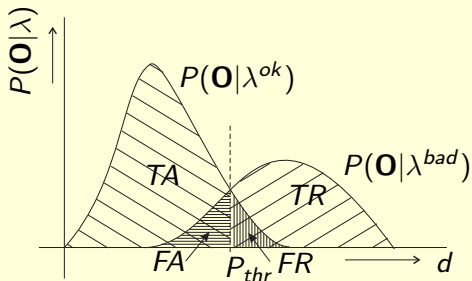
VARIANTA - DNN síť s bottleneck vrstvou (komprese):



III. část

Příklady systémů rozpoznávání řečníka

Hodnotící kritéria při verifikaci mluvčího - Míra stejné chyby



Vyhodnocování klasifikace:

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**) :

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

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)

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

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

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

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, et al, 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

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%

Queensland University of Technology, Australia : EER = 8.69%

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 !