Pokročilé metody rozpoznávání řeči

Přednáška 4

Zpřesňování akustického modelu

Výsledky a zkušenosti z minulé úlohy

Ukázka zaslaných výsledků:

Trénovací sady	SD	SI
1 sada (100 vět)	96 %	20 %
6 sad (600 vět)	92 %	73 %

HTK – slovník a slovní síť

Slovník definuje seznam slov a z jakých (dílčích) jednotek se skládají

Příklady: slovník pro rozpoznávání číslic vytvořený z celoslovních a hláskových jednotek

slovo	jednotka
NULA	nula
JEDNA	jedna
….	
DEVET	devet
SENT-END	[] sil
SENT-START	[] sil

symbol pro ticho

slovo	jednotky
NULA	n u l a
JEDNA	j e d n a
….	
DEVET	d e v j e t
SENT-END	[] -
SENT-START	[] -

Minuly semestr

Pozn. Před použitím v HTK je nutné slovník nutné „přeložit“ do angl. symbolů.

Gramatika – symbolický popis povolených sekvencí slov

$digit = JEDNA | DVA | TRI | CTYRI | PET | SEST | SEDM | OSM | DEVET | NULA;

(SENT-START ($digit) SENT-END) promluva musí obsahovat právě 1 číslici

---------- alternativně----------

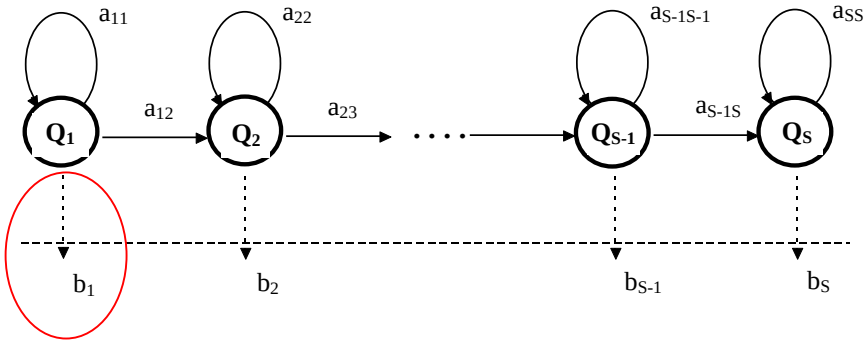
(SENT-START (<$digit>) SENT-END) promluva může obsahovat 1 nebo více číslic

Slovní síť – interní popis mezislovních přechodů

HParse grammar wordnet

Jak zpřesnit HMM?

Pomocí přesnější výstupní pravděpodobnostní funkce



Až dosud uvažováno klasické (multidimenzionální) gaussovské rozložení

$$b_s(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^P \det \Sigma_s}} . \exp[-\tfrac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_s)^T \boldsymbol{\Sigma}_s^{-1}(\mathbf{x}-\bar{\mathbf{x}}_s)]$$

Nyní ho nahradíme „vícemodálním“ gaussovským rozložením
(lineární kombinace – „směs“ - dílčích gaussovských rozložení)

$$b_s(\mathbf{x}) = \sum_{m=1}^{M} c_{sm} \frac{1}{\sqrt{(2\pi)^P \det \Sigma_{sm}}} . \exp[-\tfrac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_{sm})^T \Sigma_{sm}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_{sm})]$$

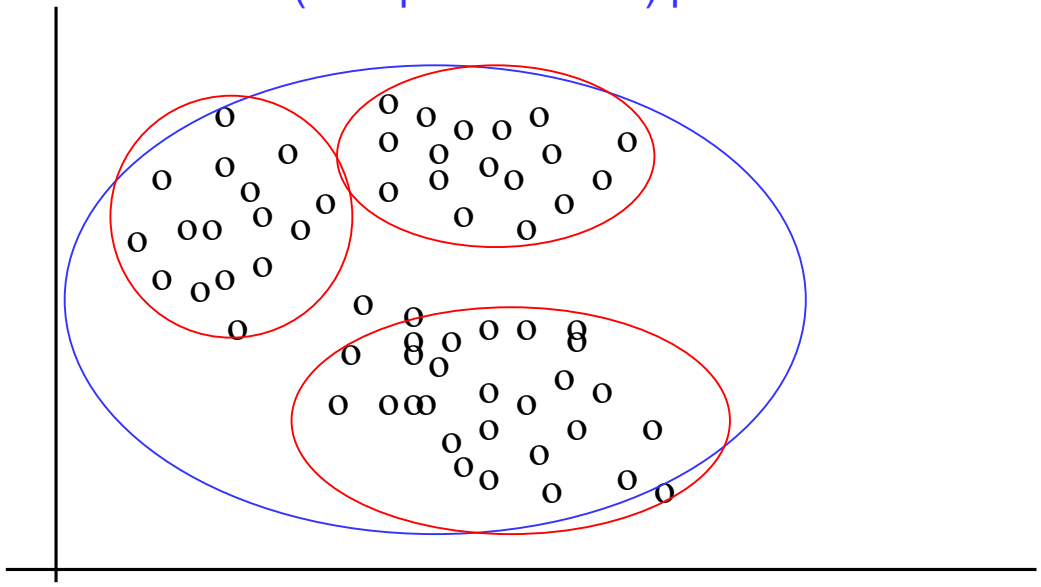
Princip vícemodálního rozložení

Data, z nichž se určují parametry gaussovského rozložení (**μ** a **σ**), jsou často uspořádána ve **shlucích (clustrech)**.

Přesnější popis rozložení dat pak dostaneme, když každý datový shluk reprezentujeme **vlastní gaussovkou** a tyto gaussovky pak **lineárně zkombinujeme**.

Příklady

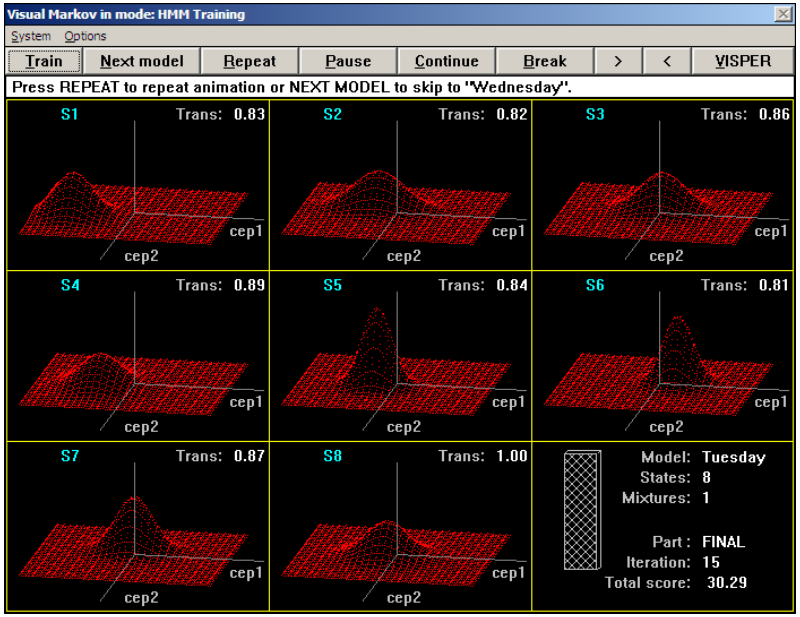
1) Rozložení tělesné výšky dospělé populace – 2 shluky dat (muži a ženy)
2) Data v dvourozměrném (dvoupříznakovém) prostoru



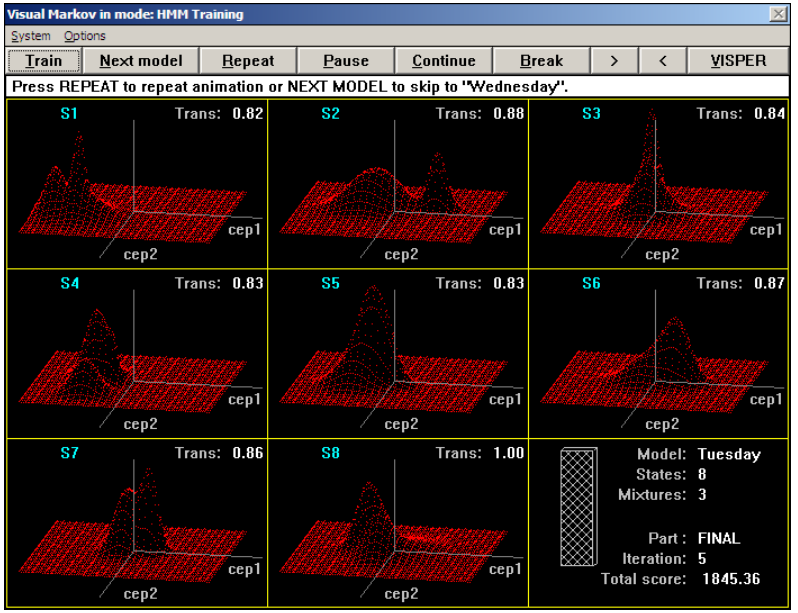
HMM s multimodálním gauss. rozložením

Směs gaussovek se v literatuře často označuje jako GMM – **Gaussian Mixture Model** a pro jednotlivé složky se používá hovorový termín **mixtura**. Při trénování je nově nutné určovat také c_{sm}

$$b_s(\mathbf{x}) = \sum_{m=1}^{M} c_{sm} \frac{1}{\sqrt{(2\pi)^P \det \Sigma_{sm}}} .\exp[-\tfrac{1}{2}(\mathbf{x}-\overline{\mathbf{x}}_{sm})^T \Sigma_{sm}^{-1}(\mathbf{x}-\overline{\mathbf{x}}_{sm})]$$



1-mixture model



3-mixture model

Trénování HMM s GMM v HTK (1)

Všechny trénovací rutiny v HTK umožňují trénovat vícemixturové HMM.

Vhodný postup:

1. **Natrénovat jednomixturové modely**
2. **Zvýšit počet požadovaných mixtur na dvojnásobek.**
3. **Několika iteracemi natrénovat nový model s daným počtem mixtur.**
4. **Opakovat kroky 2. a 3. až do požadovaného počtu mixtur.**

Pozn. 1 Vhodný počet mixtur závisí na množství trénovacích dat. Při malém množství dat nemusí být požadovaný počet mixtur u některého stavu natrénován a zároveň modely nemusí být lepší.

Pozn. 2

Dle vlastních zkušeností: Máme-li k dispozici cca 5 hodin trénovacích dat, optimální počet mixtur by mohl být 16, u 10 hodin 32,…

Trénování HMM s GMM v HTK (2)

Konkrétní postup:

Využití programu HHEd (editor souborů, zejména label souborů)

Příklad: editor vezme stávající modely v adresáři mono a v adresáři multi/mono2.0 připraví nové modely tím, že rozdělí každou stávající mixturu na 2

HHEd -H mono/hmmdefs -M multi/mono2.0 com2mix monophones.lst

Soubor **com2mix** vypadá následovně

MU 2 {*.state[2-4].mix}

Následně se provedou 2 iterace trénování pomocí HERest

V dalším kroku

HHEd -H multi/mono2.2/hmmdefs -M multi/mono4.0 com4mix monophones.lst

com4mix

MU 4 {*.state[2-4].mix}

Trénování HMM s GMM v HTK (3)

Příklad jednoduché dávky na trénování až do 4 mixtur

```
HHEd -H mono/hmmdefs -M multi/mono2.0 com2mix monophones.lst

HERest -C config -I lab.mlf -S data.lst -H multi/mono2.0/hmmdefs -M multi/mono2.1
    monophones.lst
HERest -C config -I lab.mlf -S data.lst -H multi/mono2.1/hmmdefs -M multi/mono2.2
    monophones.lst


HHEd -H multi/mono2.2/hmmdefs -M multi/mono4.0 com4mix monophones.lst

HERest -C config -I lab.mlf -S data.lst -H multi/mono4.0/hmmdefs -M multi/mono4.1
    monophones.lst
HERest -C config -I lab.mlf -S data.lst -H multi/mono4.1/hmmdefs -M multi/mono4.2
    monophones.lst
```

U závěrečného modelu se vyplatí provést opět více iterací.

Zpřesňování modelu pomocí trifonů

Dosud uváděné modely byly založeny na modelech jednotlivých hlásek.

Hlásky však zní různě (a mají tudíž různé spektrum) v různém kontextu.

Kontextově nezávislé HMM (CI-HMM) – monofony.

Kontextově závislé HMM (CD-HMM) – nejčastěji trifony

Př. značení trifonu: a - h + o hláska ‘h’, kterou předchází ‘a’ a následuje ‘o’

Slovo ‘ahoj ‘ složené z trifonů (příp. difonů) a+h, a-h+o, h-o+j, o-j

Trifonový model má smysl trénovat, je-li k dispozici skutečně velké množství trénovacích dat (> 30 hodin). I tak se trénují tzv. **tied-state triphones** (trifony s posvazovanými neboli sdílenými stavy)

Výše uvedený model se nazývá **inter-word triphone model (IW),**
na rozdíl od **cross-word modelu (CW)**, který zahrnuje kontext i přes hranice slov.

Do příštího týdne

Nasdílejte si (opravená!) trénovací data.

Stáhněte si:

- další trénovací data z e-learningu (cca 90 x 100 vět)
  - nahrávky z předchozích ročníků PMR (pouze muži)
  - nahrávky od studentů KCJ (muži i ženy)
  - nahrávky ze studentského projektu (Staněk, muži a ženy)
- ukázky skriptů pro trénování mixtur
- nová testovací data – 24 českých jmen (7x24 nahrávek, muži i ženy)

Vytvořte si u sebe 4 složky trén. dat: PMR2023, PMR_old, KCJ, Stanek

Automaticky **zkontrolujte** všechny soubory PHN, zda neobsahují nepovolené symboly. Případné symboly dalších hluků (viz poslední slajd) nahraďte je symbolem „ticha“ (-). Následně vytvořte soubory LAB a MLF.

Do příštího týdne

Spočítejte základní statistiky trénovacích dat: počet osob (adresářů), celkový počet nahrávek (souborů WAV), celkový počet hodin (zjistěte celkovou velikost souborů WAV a vydělte příslušným číslem)

Na všech trénovacích datech natrénujte vícemixturové modely (1, 2, 4, 8, 16, 32) a otestujte na nahrávkách z minula (číslovky 0 – 9). Porovnejte výsledky.

Vytvořte si slovník 24 českých jmen ve formátu HTK a proveďte rozpoznávací testy na datasetu JMENA.

Pošlete mi výsledky (a statistiky) mailem, opět do konce týdne.