Gépi Tanulás 4

Kornai András

KálmánCL 5/4

Vizsgakövetelmények

- A és C csoportok: prototipus-alapu modellezés jupyter notebookban
- B csoport: beszámoló Leslie Valiant cikkéből (PAC learning)
- Az utolsó órán meggondoljuk, hogy hogyan lehet, illetve lehet-e egyáltalán, PAC learninges feladatot csinálni a Peterson-Barney adatból
- Szorgalmi: Klautau cikk, Weka

A fonetikusok rehabilitálása

- Csomó hasznos dolgot találtak ki (beszédterápiában, hallásjavításban meg máshol). Nekünk ebből a jelfeldolgozás a legfontosabb
- A kezdeti időkben (Potter 1944) analóg áramköröket használtak
- A hetvenes években tértek át a digitális jelfeldolgozásra
- Nagyon hatékony módszereket dolgoztak ki, jó részük máig használatban van
- De akiknek a módszerekeket köszönhetjük azok nem nyelvészek hanem villamosmérnökök voltak
- A nyelvészeti fonetika (Catford, Fant, Gimson, Ladefoged, Maddieson, Ohala, Stevens – az a generáció akik az ötvenes-nyolcvanas években alkottak) nem sok nem-CL nyomot hagyott
- Azóta elsősorban LabPhon van, ami érdekes nyelvészet, de a CL feladatoktól (beszédfelismerés, szintézis) teljesen elszakadt

Hangminőség

- Megkérdezünk egy csomó embert, milyen a hang 1-től (érthetetlen) 5-ig (tökéletes)
- Ennek átlaga (standardizált mérési körülmények közt, ITU P.800) a MOS (mean opinion score)
- Ha MOS $>$ 4.5 akkor “broadcast quality”
- Ha MOS $\geq$ 4 akkor “toll quuality” azaz pénzt lehet kérni érte (és fordítva, ha a telefonhívásért pénzt kérnek, akkor legalább ilyen minőséget kell nyújtani)
- Ha MOS $\geq$ 3 akkor “communications quality” háborúban még ez is elfogadható
- Ez alatt emberi kommunikációra nem nagyon jó

Digitális jelfeldolgozás

- Analóg bemenet: gyorsan változó folytonos függvény
- Feltesszük, hogy nem túl hangos (az amplitúdó korlátos, a túl hangos részeket a maximummal helyettesítjük, legyen ez A)
- Gyakran veszünk mintát (44.1 kHz)
- A mért m_k értéket egy m_k/A-t közelítő számmal helyettesítjük, ehhez a $[0, A]$ intervallumot 2^n részre osztjuk
- Ezeket a számokat továbbítjuk (darabjához n bit kell) és a túloldalon ezekből a számokból szintetizálunk
- Ez az n-bites LPCM (linear pulse code modulation), toll quality-hoz legalább $n = 12$ bit kell, és 16kHz mintavételi gyakoriság
- CD-knél 44.1 kHz, 16 bit, 2 csatorna (sztereó) ez broadcast quality

Bitek

- Entrópia: $H = -\sum p_i \log_2 p_i$
- Ez bitekben van, ha természetes logaritmust használunk akkor nat-okról beszélünk
- Mindig egy valószinűségeloszlás tulájdonsága, nem egy egyedi elemé!
- Pont olyan statisztikai mérőszám mint az átlag és a szórás
- A maximálisan hatékony kódolás szerinti átlagos kódhosszt fejezi ki
- Az LPCM messze van a maximális hatékonyságtól, és a beszéddel még sokkal többet lehet trükközni mint a zenével!

Információtartalom

- CD-ROM Audio – 700 kbps (44.1khz, mintánként 16 bit)
- MPEG Audio – 112 kbps (44.1khz, mintánként 3 bit)
- telefon “toll quality” beszéd – 96 kbps (A-law, μ-law)
- ADPCM – 32 kbps (toll quality)
- LPC – 9.6 kbps (near toll quality)
- VQ homomorphic – 0.6 kbps
- symbolic – 0.2 kbps (0.05 kbps – négy nagyságrend!)

A gerjesztés-szűrés (Excitation-Filter) modell

Itt adták át a nyelvész-fonetikusok a stafétabotot a mérnököknek a hetvenes években, eredet Homer Dudley Vocoder (huszas évek)

- *Source (voiced):* glottal pulse train (half-rectified sine wave)
- *Source (unvoiced):* bandlimited white noise further up the vocel tract
- *Filter (oral):* hardwalled, lossless tube (only poles)
- *Filter (nasal):* two tubes (zeros as well)
- *Radiation losses:* (at lips, nostrils) mostly ignored

$e(t) \leftrightarrow E(\Omega)$
$v(t) \leftrightarrow V(\Omega)$
$s(t) = e(t) * v(t) \leftrightarrow S(\Omega) = E(\Omega)V(\Omega)$

Ekkoriban a fő beszédkódolási módszerek

- A-law, μ-law az amplitúdő-küszöböt manipulálják a jobb dinamika érdekében (máig standard)
- (Delta) Pulse Code Modulation
- Linear Prediction Coding (toll quality, 4.8-32 kbps)
- Filter Bank Analysis (Gabor-type filters) – Ez volt a nyelvészet halála
- Homomorphic (cepstral) Analysis – Ez pedig a mérnöklés kezdete, ez került a mobiltelefonokba is
- Van benne még egy pici fonetika (MEL-scale warping) de a lényeg az adatredukció!
- Kitettem egy friss cikket (min_2021.pdf) ez már wideband "zenei vonal" esetén is tudja a toll quality-t 1-2 kbps-nél

A klasszikus S-C-I megközelítés

input sound

Signal Processing

clean sound

Segmentation

finding phone endpoints

Feature Extraction

feature vectors

Classification

phone hypotheses

Identificaton

word and sentence hyphoteses

A HMM megközelítés

input sound

Signal Processing

clean sound

Feature Extraction

feature vectors

HMM

word and sentence hyphoteses

Difficulties with S-C-I

1. Segmentation decisions are largely context free
2. Segmentation errors propagate to identification stage
3. Global characteristics of speech largely lost
4. Language modeling only as postprocessing
5. Components optimized separately

A HMM-ek éppen ezek az okok miatt győztek az S-C-I-vel szemben

No recognition without segmentation

Legyen w a szegmentálatlan fájl, amiben az s és s' közti szegmens $w(s, s')$. Legyen egy szegmentáció $s_0, s_1, ..., s_n$ és az ehhez tartozó cimkézés $l_i = L(w(s_{i-1}, s_i))$ $(i = 1, ..., n)$
Ha a szegmentáló a $t_0, t_1, ..., t_m$ szegmentálást $P(t_0, t_1, ..., t_m|w)$ valószinűséggel adja vissza, és a cimkéző az r cimkét $Q(r|w(s, s'))$ valószínűséggel adja ki, akkor a helyes felismerés valószinűsége

$$\sum_{m=1}^{\infty} \int \cdots \int_{t_0,...,t_m} P(t_0, ..., t_m|w) \prod_{i=1}^{m} Q(l_i|w(t_{i-1}, t_i))dt_i$$

Ezért manapság a szegmentációt is HMM-mel csináljuk nem kézzel