Zpracování řečových signálů – program kursu, úvod 2024 / 25

Honza Černocký

ÚPGM FIT VUT Brno, cernocky@fit.vutbr.cz

Plán

- Who's who
- Program kursu
- Hodnocení
- Literatura a Web

- Vědy vztahující se ke zpracování řeči
- Informační obsah řeči.
- Aplikační okruhy.
- Dolování dat z řeči.

Kdo Vás bude učit

- přednášky: Honza Černocký.
- jiné lidé (LVCSR, syntéza) — když tomu někdo rozumí lépe
- externisté: Vláďa Malenovský, VoiceAge, Jindra Matoušek, ZČU Plzeň a další.
- poč. cvičení: Katia Vendrame.
- `https://www.fit.vut.cz/person/ivendrame`
- zapisování není - jen 1 skupina.

Program kursu

- **Přednášky - A112**
- **Počítačové laboratoře - O204 ????? —** Python (thanks Honza Brukner 2021/22) účast **je kontrolována a hodnocena**,
- **Projekt —** 1, ručně vyhodnocován.
- **numerické cvičení –** jedno ke konci semestru, příklady na LPC, DTW, HMM.

Přednesy

1. Organisace kursu, aplikace, vědy.
2. Číslicové zpracování řečových signálů.
3. Předzpracování řeči, model tvorby řeči a cepstrum.
4. Lineárně-prediktivní model.
5. Určování základního tónu.
6. Kódování řeči I. [Vláďa Malenovský]
7. Kódování II. [Vláďa Malenovský]
8. Úvod do rozpoznávání, DTW, Guassian mixture models.
9. Skryté Markovovy modely (Hidden Markov models HMMs).
10. HMM II. Plynulá řeč s velkým slovníkem [Karel Beneš]
11. Neuronové sítě v ASR [Petr Schwarz]
12. Rozpoznávání mluvčího [Olda Plchot]
13. Syntéza řeči [Jindra Matoušek]
14. Lidské slyšení a tvroba řeči vs neuronové sítě [Hynek Heřmanský]

N1 Numerika 2 hodiny pro všechny: číslicový filtr, LPC - výpočet filtru, DTW, HMM.

Harmonogram přesněji ...

10.2. NIC
17.2. 4h prednaskovy blok - Honza uvod a zprac signalu
14.2. 2h prednaska - Honza zprac. signalu, prepro, ceps. + 2h cviko c. 1
3.3. 2h prednaska - Honza ceps + LPC zacatek
10.3. 2h prednaska - Honza LPC konec + zakladni ton + 2h cviko c. 2
17.3. 2h prednaska - Vlada Malenovsky - Kodovani 1
124.3. 2h prednaska - Vlada Malenovsky - Kodovani 2 + 2h cviko c. 3
31.3. 2h prednaska - Honza ASR zaklady a DTW + 2h cviko c. 4
7.4. (ICASSP) 2h prednaska - Karel Benes - LVCSR + 2h cviko c. 5
14.4. 2h prednaska Petr Schwarz - NN in ASR + 2h cviko c. 6
21.4. NIC (VELIKONOCE)
28.4. 2h prednaska - Jindrich Matousek - synteza + 2h numericke cviko (Honz
5.5. 4h prednaskovy blok - Oldrich Plchot SID + 2h Hynek Hermansky

Labs

L1 Zpracování řeči v Python: čtení/zápis zvukových souborů, základní operace, ukládání, nahrávání. Spektrogram. Segmentace, základní parametry - viz také totální remake ISS v ak. roce 2022/23!

L2 LPC a vektorová kvantizace, výpočet chyby.

L3 Detekce základního tónu a jednoduchý kodér.

L4 Gaussian Mixture model a jednoduchý klasifikátor něčeho.

L5 DTW a HMM.

L6 HMM.

Old labs

viz www ... pár, které by Vás mohly zajímat (real-time zpracování zvuku, syntéza).

HODNOCENÍ

co	za kolik
1 projekt	23
půlsemestrálka	14
semestrálka	51
laboratoře	12
celkem	100

- Obě zkoušky **bez materiálů** ☹, ale s **“cheat-sheets”** ☺
- bodování počítačových cvik ano.
- projekt - ruční čtení a pouštění kódu.
- možnost přijít s vlastním projektem, který bude nahrazovat standardní (diplomka, atd).

Literatura

- Studijní opora na `http://www.fit.vutbr.cz/study/courses/ZRE/public/`
- Psutka J.: Komunikace s s počítačem mluvenou řečí. Academia, Praha, 1995 [knihovna]
- Psutka J., Muller, L: Mluvíme s počítačem česky, Academia, 2006. [knihovna]
- Gold B., Morgan. N.: Speech and audio signal processing, John Wiley & Sons, 2000 [knihovna]
- S. Young, J. Jansen, J. Odell, D. Ollason, P. Woodland: The HTK book, Entropics Cambridge Research Lab., 1996, Cambridge, UK. Výborný, byť už trochu letitý, úvod do HMM, ke stažení na `http://htk.eng.cam.ac.uk/`
- **Superlectures** `http://www.superlectures.com/fit-zre/`

Streaming / záznamy

- A112 na fakultních video serverech a na YT:
  `https://www.youtube.com/playlist?list=PL_eb8wrKJwYueOOqt_aVId5bvjXAv1VZc`
- O204 nikde.

ZRE je pro Vás !

- Něco, co Vám v programu přednášek nebo labin chybí ?
- Něco, co už není "in" a být by tam nemělo ?
- Připomínky k hodnocení ?
- Podělte se o ně s přednášejícím na kterékoliv přednášce nebo pošlete mail (ale raději řekněte osobně . . .).

ÚVOD DO AUTOMATICKÉHO ZPRACOVÁNÍ ŘEČI

Definice

"Automatické zpracování řeči umožňuje hlasovou komunikaci mezi lidmi navzájem (kódování) nebo mezi člověkem a strojem."

Vědy ve zpracování řeči

Zpracování řeči je multi-disciplinární obor, využívající poznatků věd přírodních, technických i humanitních.

- **fysiologie:** porozumění funkci hlasového a sluchového ústrojí, pomoc při tvorbě různých modelů.
- **akustika:** studuje fyzikální mechanismy tvorby a slyšení řeči.
- **zpracování signálu:** mnoho zúčastněných oblastí: modelování, parametrisace, identifikace, spektrální analýza, kódování, teorie informace, klasifikace vzorů, atd.
- **humanitní vědy**
  - *fonetika* – o tvoření a slyšení zvukové stránky řeči.
  - *fonologie* – o fonémech a jejich systému v daném jazyce. *Foném* je hláska, která je v jazyce *významotvorná*, její změna může tedy měnit význam slova.
  - *prosodie* – o zvukové stránce jazyka (melodie, trvání hlásek, přízvuk ve slovech a ve větách).

- *lexikologie* – o slovní zásobě jazyka, jejím vývoji z vztazích mezi slovy (*synonyma:* 2 různá slova, stejný význam, *antonyma*: opačný význam, *homonyma:* stejné slovo, více významů). Napomáhá tvorbě *slovníků* a to i počítačových, důležité pro rozpoznávání řeči.
- *gramatika (mluvnice)* – soubor pravidel o obměnách slov a o jejich spojování ve věty. Důležité pro syntézu.
- *syntaxe* – část gramatiky zabývající se větou a souvětím.
- *sémantika* – o významu jazykových jednotek. Vychází obvykle od *slova*, důležitá pro porozumění řeči.

Různé úrovně zpracování řečového signálu

příklad na větě "Přišel jsem pozdě".

1. *akustická* - signál, spektrogram, parametry.
2. *fonetická* - p ř i š e l j s...
3. *lexikální* - sloveso: tvar "přijít", sloveso: tvar "být", příslovce.
4. *syntaktická* - (nevyjádřený podmět) přísudek příslovečné určení času.
5. *sémantická* - sdělení, že jsem přišel pozdě.
6. *pragmatická* - co jsem svým sdělením sledoval - omluva, prosté oznámení ?

Pragmatická úroveň musí vzít v úvahu význam slov vzhledem k *záměru člověka*. V aut. zpracování řeči je to zatím neřešená otázka.

INFORMAČNÍ OBSAH ŘEČI

snažíme se kvantifikovat *informační rychlost* (v bitech za sekundu, ve zkratce bit/s nebo bps), nutnou k popisu řeči v různých formách. Pro srovnání uvedeme formu fonetickou a akustickou s číslicovým vyjádřením signálu.

Fonetická forma

počet fonémů v češtině: 36. Pro vyčíslení *množství informace* budeme uvažovat zdroj generující na sobě navzájem nezávislé prvky x_i z množiny $X = \{x_1, \ldots, x_S\}$, kde S je konečný počet prvků. Každý z prvků má určitou pravděpodobnost $p(x_i)$ a prvky tvoří úplnou soustavu, takže:

$$\sum_{i=1}^{S} p(x_i) = 1. \tag{1}$$

Informační obsah i-tého prvku je dán počtem bitů, které potřebujeme k jeho vyjádření:

$$I(x_i) = -\log_2 p(x_i) \quad \text{[bit]}. \tag{2}$$

Entropie zdroje (střední hodnota informace) je pak dána:

$$H(X) = -\sum_{i=1}^{S} p(x_i) \log_2 p(x_i). \tag{3}$$

Pro české fonémy, předpokládáme-li zjednodušeně stejnou pravděpodobnost jejich výskytu, je tato entropie $H(X)$=5.2 bit. Vezmeme-li v úvahu jejich pravděpodobnost, je to $H(X)$=4.6 bit. Vezmeme-li navíc ještě v úvahu vzájemné *závislosti* mezi fonémy (bigramy: $p(x_j|x_i)$, trigramy: $p(x_k|x_i x_j)$, atd.), dosahuje entropie hodnoty $H(X)$=3–3.5 bit.

V běžném českém hovoru produkuje člověk cca 80–130 slov za minutu, tedy asi 10 fonémů za sekundu. Informační rychlost je tedy asi C_{phn} = **30–40 bit/s**. Psychoakustické testy ukázaly, že člověk je schopen zpracovat příchozí informace do rychlosti asi 50 bit/s.

Akustická forma

Je-li řeč representována číslicovým signálem, musí být splněn Nyquistův–Shannonův–Kotelnikovův teorém:

$$F_s > 2F_m, \tag{4}$$

kde F_s je vzorkovací kmitočet a F_m je nejvyšší kmitočet obsažený ve spektru signálu. Každý vzorek je kvantován jednou m možných kvantovacích hladin, k jejich vyjádření potřebujeme $N = \log_2 m$ bitů. Odstup signálu od šumu v dB je úměrný hodnotě $6N$. Vyjádříme-li pomocí těchto veličin informační rychlost, dostaneme:

$$C_{ak} = \frac{I}{t} = \frac{\log_2 m}{T_s} = NF_s. \tag{5}$$

Příklady:

1. Signál s Hi-Fi kvalitou, F_s=44.1 kHz, N=16 bit. Výsledná informační rychlost je C_{ak} = **705 kbit/s**.
2. Signál s telefonní kvalitou (pásmo omezeno od 300 do 3400 Hz): F_s=8 kHz, N=8 bit. Výsledná informační rychlost je C_{ak} = **64 kbit/s**.

Ponaučení ?

Z příkladů je patrné, že akustická forma má oproti fonetické formě obrovskou *redundanci* (nadbytečnost). Mimo informace obsažené ve fonetické formě nese totiž informaci o osobnosti mluvčího, o barvě jeho hlasu, náladě, prostředí, atd. Mozek posluchače pak provádí dekódování a separaci různých typů informací — Bohu žel, nevíme přesně jak. Přesto je vhodné využít alespoň základních informací o *tvorbě řeči* k **omezení této informační rychlosti.**

APLIKAČNÍ OBLASTI ZPRACOVÁNÍ ŘEČI

- kódování: pro přenos a ukládání. **Cíl:** vyjádření řeči co nejmenším počtem bitů.
  Požadavky:
  - náročnost ↓,
  - zpoždění ↓,
  - srozumitelnost ↑,
  - přirozenost ↑,
  - odolnost proti chybám v kanálu ↑.
- syntéza: od konkatenace (MHD Brno) až po TTS (text-to-speech).
- Další aplikace
  - medicína: zkoumání anomálií a chorob hlasového traktu.
  - psychologie a kriminalistika: detektor lži, identifikace alkoholu v krvi, detekce stresu či únavy,...
  - pomoc handicapovaným (učení hluchých dětí správné výslovnosti, atd.)
- **Co se dělá ve Speech@FIT** ... viz samostatná *barevná* presentace.