

Rozpoznávání řeči pomocí Kaldi

Diplomová práce o implementaci Kaldi rozpoznávače řeči pro Alex SDS

Ondřej Plátek

Matematicko-fyzikální fakulta Univerzity Karlovy

27. 5. 2014

Cíle práce

Zlepšit rozpoznávání řeči pro Alex Spoken Dialogue Systems
Obzvláště aplikaci poskytující informace o veřejné dopravě (800 899 998).

Zadané cíle práce:

- připravit akustické modely pomocí Kaldi toolkitu,
- vyvinout nový real-time rozpoznávač řeči, který podporuje inkrementální rozpoznávání řeči,
- integrovat rozpoznávač řeči do Alex SDS.

Kontinuální rozpoznávání řeči

Pattern matching

HMM — modelování časové řady řeči (monofóny/triphóny pro slova)

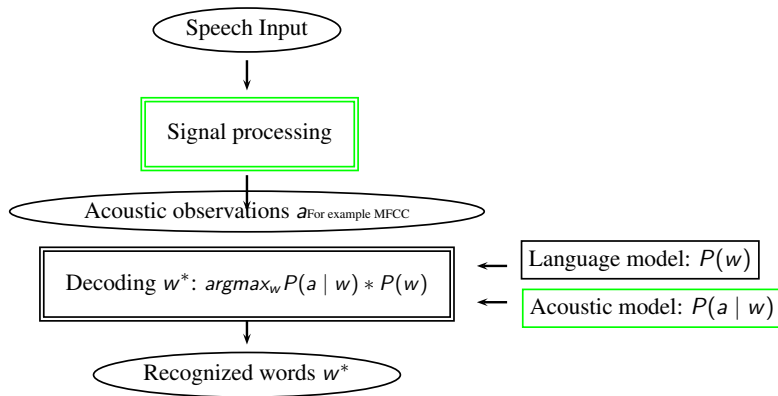
- Natrénovali a porovnali jsme několik HMM akustických modelů.

Prohledávání grafu - dekódování

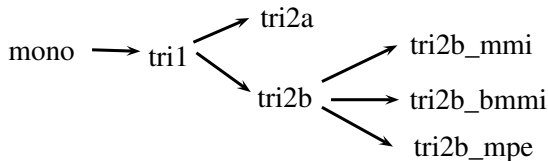
Viterbi algoritmus — dynamické programování

- Změna rozhraní.
- Normalizace výstupu rozpoznávače.
- Nalézt optimální parametry (beam, lattice-beam, max-active-states).

Akustické modelování



ASR trénování, výsledky



Czech

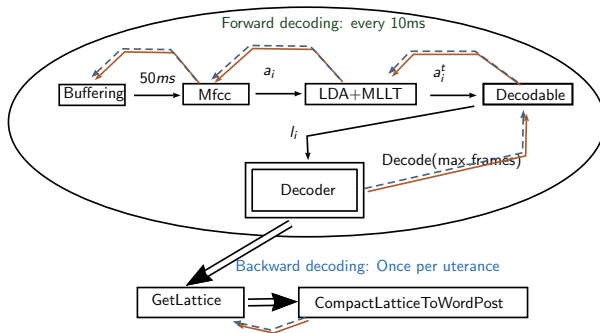
tri $\Delta + \Delta\Delta$	<i>HTK (60.4)</i> 56.6
tri LDA+MLLT	53.9
tri LDA+MLLT+MMI	49.5
tri LDA+MLLT+bMMI	49.3
tri LDA+MLLT+MPE	49.2

bigram

English

tri $\Delta + \Delta\Delta$	<i>HTK (17.5)</i> 16.2
tri LDA+MLLT	15.8
tri LDA+MLLT+MMI	10.4
tri LDA+MLLT+bMMI	10.2
tri LDA+MLLT+MPE	11.1

Komponenty on-line dekodování



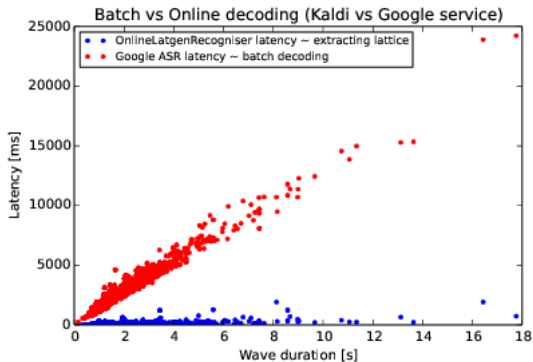
(Py)OnlineLatgenRecogniser rozhraní

- *Audioln* – zařazení audio do fronty k předzpracování
- *Decode* – dekodování určitý počet audio rámců (frame)
- *PruneFinal* – příprava datových struktur pro extrakci lattice (svazu)
- *GetLattice* – extrakce slovní posteriorní lattice
- *GetBestPath* – extrakce nejpravděpodobnější slovní hypotézy
- *Reset* – příprava rozpoznávače na novou promluvu

Evaluace, metriky

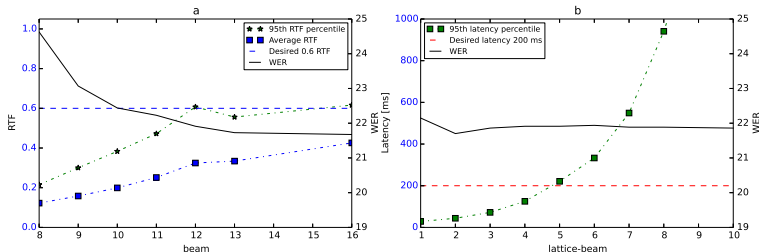
- Real Time Factor (RTF) dekodování – poměr času dekodování a délky promluvy,
- Latence – zpoždění mezi koncem promluvy a dostupností výsledků rozpoznávání,
- Word Error Rate (WER) – chyba nejlepší slovní transkripce.

On-line vs dávkové dekódování

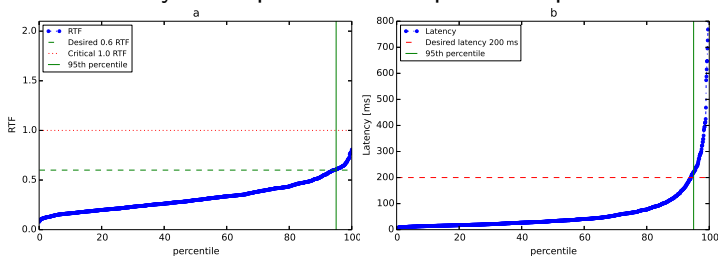


Pokles WER 45 % → 22% pro náš dialogový systém Alex

Public Transport Information doména - rychlost a přesnost



Dostatečně rychlé – pod 200 ms – pro 95 % promluv.



Shrnutí

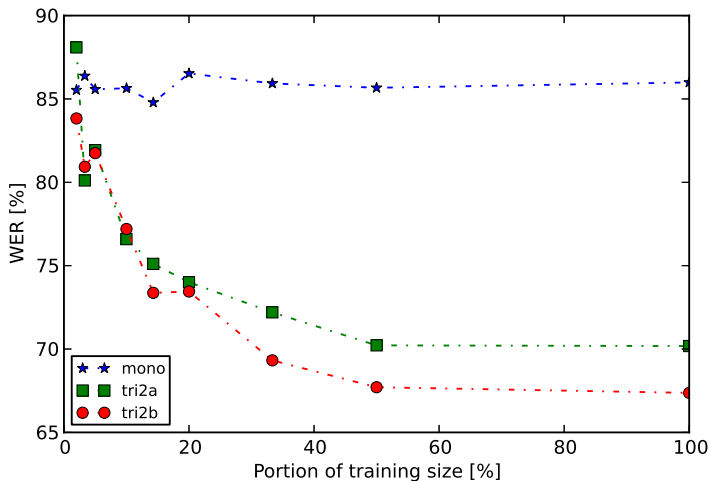
Výsledky

- V dialogovém systému WER 22, latence pod 200 ms.
Dříve 1900 ms a 48 WER.
- WER pro Vystadial skripty: angličtina 12, čeština mix domén 50

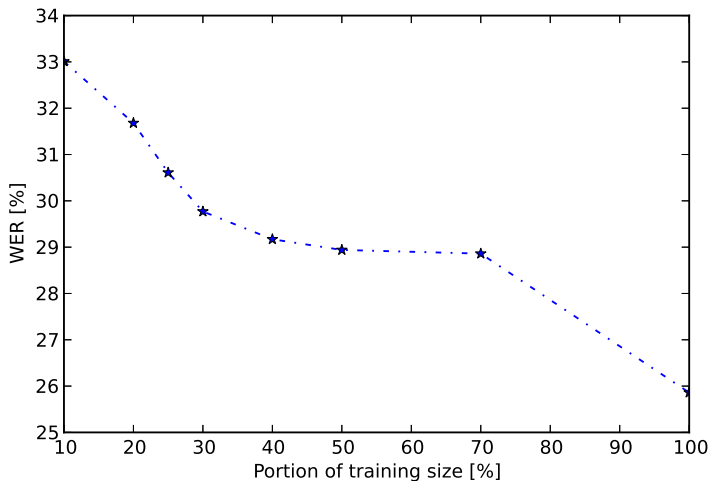
Závěry

- Testovaný real-time on-line rozpoznávač řeči
- Trénovací skripty pro češtinu a angličtinu (Vystadial)– přijmuto do Kaldi
- Integrace rozpoznávače řeči do dialogového systému, v reálném provozu na lince 800 899 998 (PTI doména)
- Spoluaautor akceptovaných článků na konferenci Sigdial, Lrec a TSD (Viz reference)

Přesnost akustických modelů dle velikosti dat



Přesnost akustických modelů dle velikosti trénovací dat LM

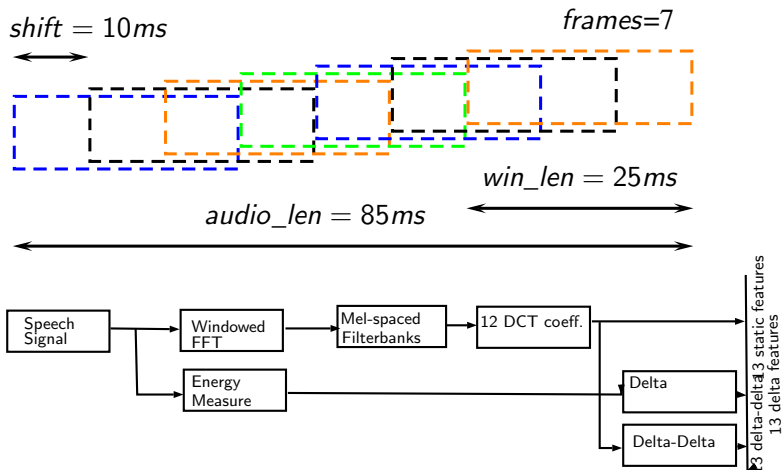


Vystadiad dataset

Posbíráno skupinou UFAL Dialogovým systémů.

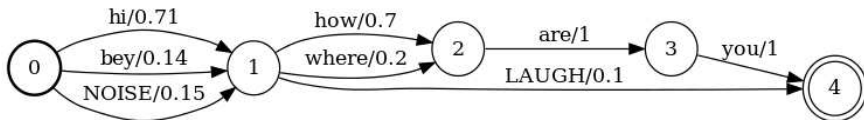
dataset	audio[h]	# vět	# slov
English			
training	41:30	47,463	178,110
development	01:45	2,000	7,376
test	01:46	2,000	7,772
Czech			
training	15:25	22,567	126,333
development	01:23	2,000	11,478
test	01:22	2,000	11,204

Acoustic features, features preprocessing



Výstupní formáty

0.5 hi how are you
0.2 hi where are you
0.1 bey how are you



Funkční (Py)OnlineLatgenRecogniser demo

```
1 d = PyOnlineLatgenRecogniser()
2 d.setup(argv)
3 while audio_to_process():
4     d.audio_in(get_raw_pcm_audio())
5     dec_t = d.decode(max_frames=10)
6     while dec_t > 0:
7         decoded_frames += dec_t
8         dec_t = d.decode(max_frames=10)
9 d.prune_final()
10 lik, lat = d.get_lattice()
```

Problém

Dialogové systémy potřebují rozpoznávání řeči

OpenJulius — padá, RWTH decoder — licence

Cloudové služby Google a Nuance — žádná adaptace + problémy s licencemi

Semiring

Name	\mathcal{K}	\oplus	\otimes	$\bar{0}$	$\bar{1}$
Real	$[0, \infty)$	$+$	$*$	0	1
Log	$(-\infty, \infty)$	$-\log(e^{-x} + e^{-y})$	$+$	∞	0
Tropical	$(-\infty, \infty)$	min	$+$	∞	0

Odkazy a reference

Děkuji za pozornost!

Související odkazy

- Diplomová práce
<https://github.com/oplatek/kaldi-thesis>
- OnlineLatgenRecogniser implementace a AM trénovací skripty
<https://github.com/UFAL-DSG/pykaldi>
- Alex implementace
<https://github.com/UFAL-DSG/alex>

Reference

- Vystadial dataset – Matěj Korvas, Ondřej Plátek, Ondřej Dušek, Lukáš Žilka, and Filip Jurčiček, Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2014), 2014, p. To Appear.
- Free on-line speech recogniser based on Kaldi ASR toolkit producing word posterior lattices – Ondřej Plátek and Filip Jurčiček, Proceedings of the Sigdial 2014 conference
- Integration of an online Kaldi speech recogniser to Alex Dialogue Systems Framework – Ondřej Plátek and Filip Jurčiček, TSD conference 2014