Speech recognition using Kaldi

Thesis about implementation Kaldi ASR for Alex SDS

Ondřej Plátek

Matematicko-fyzikální fakulta Univerzity Karlovy

14. 4. 2014

Goals of thesis

Improve speech recognition for Alex Spoken Dialogue Systems
Particularly public transport information application (800 899 998).

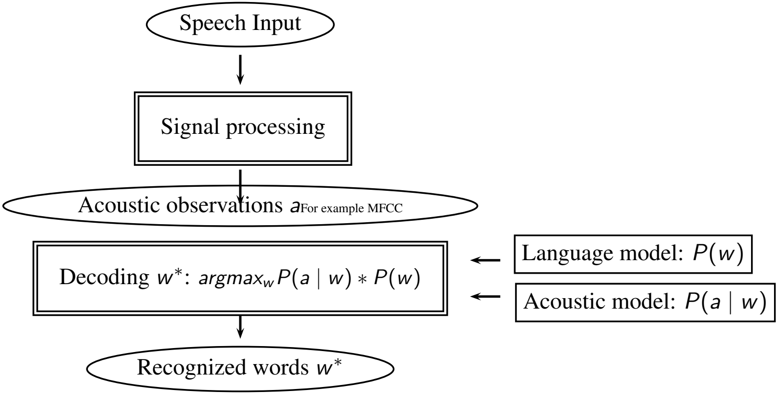
Goals of the thesis were:

- to build acoustic models using the Kaldi toolkit,
- to develop new real-time recogniser which supports incremental speech recognition,
- to integrate the recogniser into our Alex SDS.

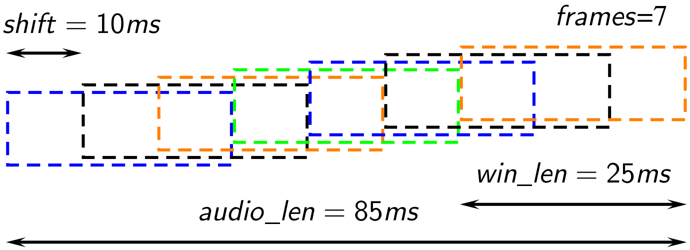
Content

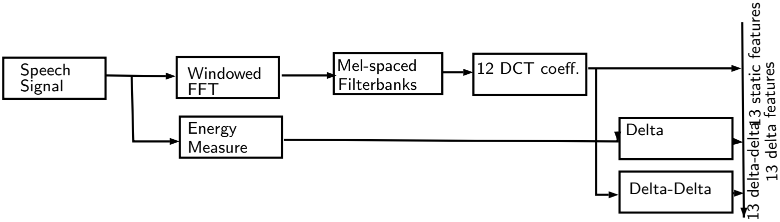
1. Task
2. ASR introduction
3. Evaluation in Public Transport Information domain
4. On-line recogniser
5. Acoustic modelling
6. Summary
7. Details

ASR components



Acoustic features, features preprocessing





Continuous Speech recognition

Pattern matching

HMM — speech time series modelling (phones/triphones for words)

- We trained several HMM acoustic models.

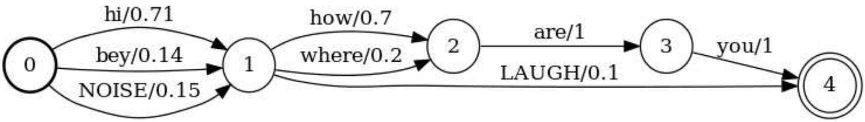
Graph search - decoding

Viterbi algorithm — dynamic programming

- We search for best parameters (beam, max-active-states).
- Normalise its output.
- Change interface.

Output formats

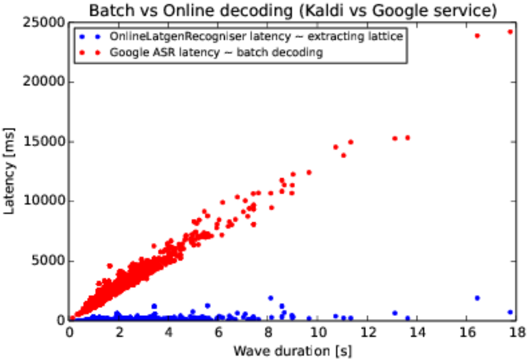
```
0.5 hi how are you
0.2 hi where are you
0.1 bey how are you
```



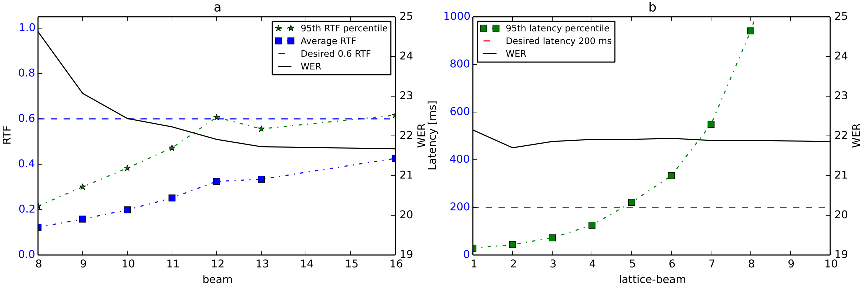
Evaluation measures

- Real Time Factor (RTF) of decoding – the ratio of the recognition time to the duration of the audio input,
- Latency – the delay between utterance end and the availability of the recognition results,
- Word Error Rate (WER).

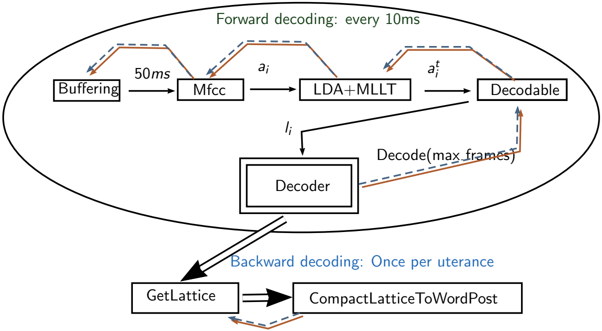
On-line vs batch decoding



Public Transport Information domain - results



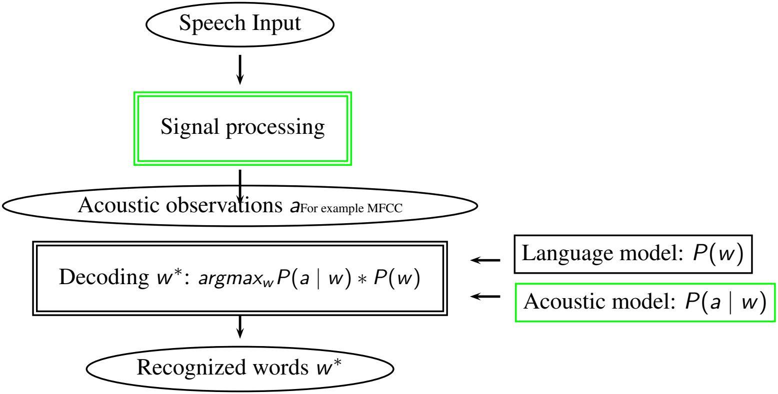
Components for on-line decoding



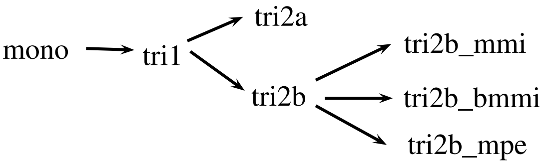
(Py)OnlineLatgenRecogniser interface

- *AudioIn* – queueing new audio for pre-processing
- *Decode* – decoding a fixed number of audio frames
- *PruneFinal* – preparing internal data structures for lattice extraction
- *GetLattice* – extracting a word posterior lattice
- *GetBestPath* – extracting a one best word sequence
- *Reset* – preparing the recogniser for a new utterance

Acoustic modeling



Acoustic models training



Training method name	Script shortcut
Monophone	mono
Triphone	tri1
$\Delta + \Delta\Delta$	tri2a
LDA+MLLT	tri2b
LDA+MLLT+MMI	tri2b_mmi
LDA+MLLT+bMMI	tri2b_bmmi
LDA+MLLT+MPE	tri2b_mpe

Vystadial dataset

Collected by UFAL Dialogue system group.

dataset	audio[hour]	# sentences	# words
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

ASR training results

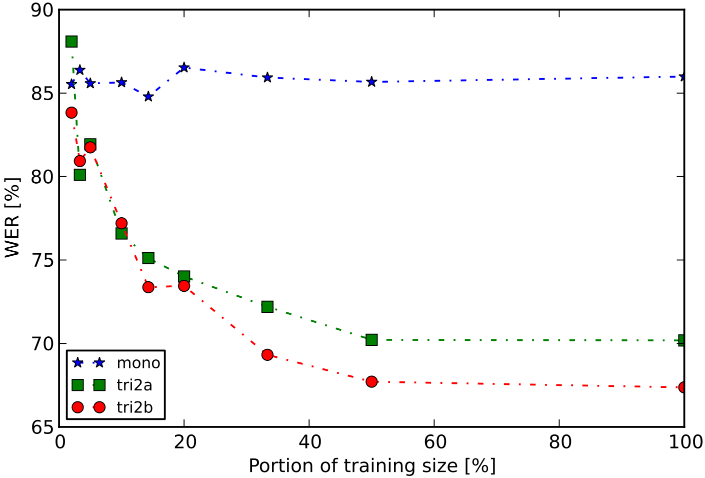
language/method	bigram
Czech	
tri $\Delta + \Delta\Delta$	56.6
tri LDA+MLLT	53.9
tri LDA+MLLT+MMI	49.5
tri LDA+MLLT+bMMI	49.3
tri LDA+MLLT+MPE	49.2
English	
tri $\Delta + \Delta\Delta$	16.2
tri LDA+MLLT	15.8
tri LDA+MLLT+MMI	10.4
tri LDA+MLLT+bMMI	10.2
tri LDA+MLLT+MPE	11.1

HTK and Kaldi acoustic models

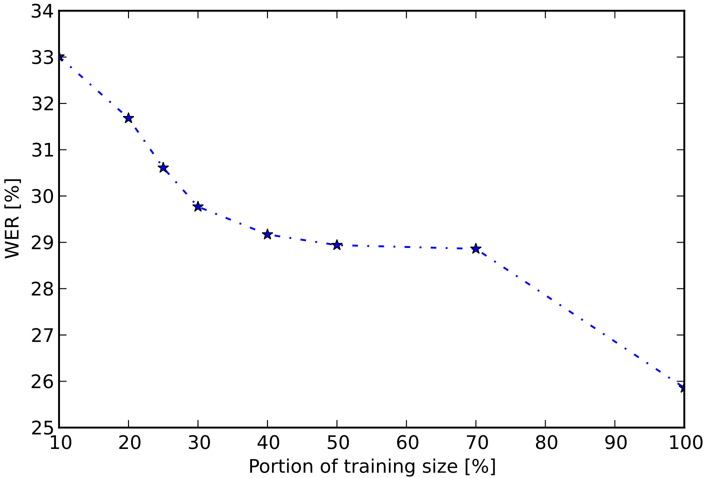
HTK method	**bigram**
Czech	
tri $\Delta + \Delta\Delta$	60.4
English	
tri $\Delta + \Delta\Delta$	17.5

Kaldi method	**bigram**
Czech	
tri $\Delta + \Delta\Delta$	56.6
English	
tri $\Delta + \Delta\Delta$	16.2

Acoustic model accuracy based training data size



Speech recognition accuracy based on LM training data size



Achievements

- Working real-time on-line speech recogniser
- Developed acoustic modeling scripts for Czech and English - accepted to Kaldi svn trunk
- Integration of ASR into Alex Dialogue Systems Framework
- Improved speech recognition for toll-free line 800 899 998

Results

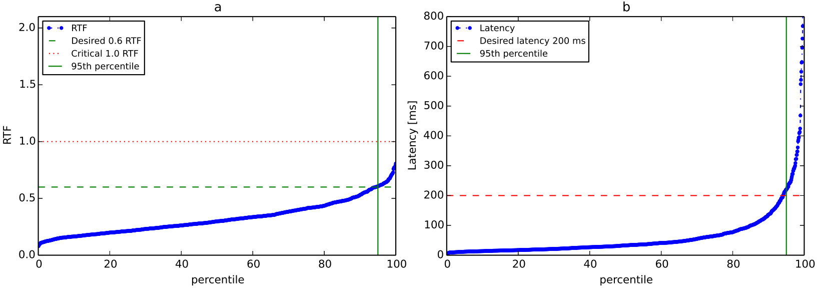
- WER 22, latency under 200 ms on Public Transport Information domain (Czech)
- WER 50 for Czech on Vystadial dataset (Czech - complex domain)
- WER 12 for English on Vystadial dataset

Functional (Py)OnlineLatgenRecogniser demo

```python
d = PyOnlineLatgenRecogniser()
d.setup(argv)
while audio_to_process():
    d.audio_in(get_raw_pcm_audio())
    dec_t = d.decode(max_frames=10)
    while dec_t > 0:
        decoded_frames += dec_t
        dec_t = d.decode(max_frames=10)
d.prune_final()
lik, lat = d.get_lattice()
```

Speed - RTF and Latency

Fast enough for 95 % of utterances.



Problem

Spoken dialogue systems needs speech recognition
OpenJulius — crashes, PocketSphinx — no posteriors, RWTH decoder — license
Cloud based services Google and Nuance — no customisation + license issues

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

Links and references

Thank you for your attention!

Related links

- Thesis and this slides
  https://github.com/oplatek/kaldi-thesis
- OnlineLatgenRecogniser implementation and AM training scripts
  https://github.com/UFAL-DSG/pykaldi
- Alex implementation
  https://github.com/UFAL-DSG/alex
- Contact & CV
  http://www.linkedin.com/in/ondrejplatek

References

- Vystadial dataset – Matěj Korvas, Ondřej Plátek, Ondřej Dušek, Lukáš Žilka, and Filip Jurčíček, 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.