Towards the First Dictation System for Latvian Language

Askars Salimbajevs
Tilde, Latvia

Baltic HLT 2016
Introduction

• Acoustic properties of dictated speech
• Language properties of dictated speech
• Dictation commands
  • Punctuation
  • New line, new paragraph
  • Special symbols (&, #, emoticons)
  • Formatting and editing
• Real-time factor
Speech recognition: overview

- Speech signal
- Feature extraction
- Acoustic model
- Language model
- Combining both models
- List of best hypotheses
Acoustic model

• Cross-entropy trained HMM-DNN
• Based on Kaldi online/nnet2 recipe for “Switchboard”.
• 100h Latvian Speech Recognition Corpus[1]
• 8h from Latvian Dictated Speech Corpus[2] is added for domain adaptation
  • Contains punctuation and other commands
  • Contains parallel recordings from various devices
Language model

- Trained 44M sentences from web portals
  - Special preprocessing for dictation
- 800K vocabulary
- 2-gram model for 1-pass
- 3-gram model for rescoring
Adapting language model

• All punctuation and special symbols (#, %, &, ... ) are replaced with words.
• Number conversion from digits to words with correct inflection.
• Then formatting and other commands were artificially added as separate sentences.
• Finally, “New line” commands were appended after every second sentence in the text corpus.
Results

- Evaluation on 1 hour held out set of dictated speech

<table>
<thead>
<tr>
<th>ASR system</th>
<th>WER, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline with non-adapted LM</td>
<td>40.7%</td>
</tr>
<tr>
<td>Baseline with adapted LM</td>
<td>27.3%</td>
</tr>
<tr>
<td>Both AM and LM adapted</td>
<td>23.9%</td>
</tr>
</tbody>
</table>
Based on the full-duplex ASR system for Estonian [3]
Dictation client

- Based on dictate.js\[4\]
- Implementation of dictation commands
- Voice Activity Detection (VAD)
  - Reduces server load
  - Prevents iVector adaptation overfitting to silence
Shared rescoring language model

• Each recognition process has it’s own copy of rescoring LM
• These copies consume a lot of RAM, but a queried only for rescoring

• Idea – make “rescoring” LM shared between processes

• Advantages:
  • Smaller memory usage, more processes on the same machine

• Disadvantages:
  • Latency
Results

- Idle memory usage
Results

- Real-time performance evaluation
Conclusion

• Special text corpus preprocessing
  • 30% relative improvement
• Using Latvian Dictated Speech Corpus
  • Improved acoustic model (17%)
• WER 23.9 % on 1-hour set of dictated speech
  • 40% relative improvement against baseline system
• Integrated as a beta feature in the existing products
  • Voice activity detection
  • Dictation commands
  • Deployment on the same hardware