Lightly Supervised Acoustic Model Training on EPPS Recordings

Matthias Paulik and Alex Waibel
Interactive Systems Laboratories, Carnegie Mellon University / Universität Karlsruhe (TH)

(1) Introduction
- Train ASR for language $L_i$ by exploiting available EPPS data
- Problem: FTE can differ significantly from politician and interpreter speech
  ⇒ Apply lightly supervised acoustic model training
    - Bias initial ASR with available knowledge
    - Use hypotheses of biased system for training
- In this work: bias initial German ASR with
  - German FTE
  - German translations extracted from English and Spanish audio channels using spoken language translation (SLT)

(2) Experimental Setup
- Data: German, English, Spanish EPPS audio
  - Satellite live broadcast recorded @ interACT

<table>
<thead>
<tr>
<th>Sessions</th>
<th>Utterances</th>
<th>Audio [h]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>1</td>
<td>592</td>
</tr>
<tr>
<td>Test</td>
<td>1</td>
<td>885</td>
</tr>
<tr>
<td>Training</td>
<td>93</td>
<td>73,408</td>
</tr>
</tbody>
</table>

German audio data statistics
- German dev/test transcriptions created @ interACT
- No human transcriptions created for Eng, Spa
- FTEs extracted from Europarl_v3 corpus

- ASR Systems: German, English, Spanish
  - 2 pass systems; Janus Recognition Toolkit
    - Ger: AM - 70h BN data; LM - German FTEs
    - Eng, Spa: interACT '06 TC-STAR sub-systems
  - Typical WER: 12.13% Eng and 11.12% Spa

- MT Systems: English→German, Spanish→German
  - trained on Europarl_v3 corpus
    - GIZA++, interACT STTK decoder
    - Dev/test BLEU: 12.5/15.2 E→G, 11.9/13.4 S→G

(3) FTE & SLT based Supervision: Impact on WER
- Session specific supervision via biased LM
- 3 Types of supervision
  - GFTE sup.: FTE is part of LM training data
  - FTE sup.: FTE receives higher interpolation weight
  - SLT sup.: include 1000-best SLT hypotheses (per source language) in LM training, use interpolation
- No supervision for Eng/Spa ASR
- GFTE supervision for MT systems (TM and LM)

<table>
<thead>
<tr>
<th>PPL</th>
<th>-</th>
<th>GFTE</th>
<th>FTE</th>
<th>FTE &amp; SLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>219</td>
<td>206</td>
<td>161</td>
<td>138  142</td>
</tr>
<tr>
<td>Test</td>
<td>190</td>
<td>176</td>
<td>146</td>
<td>127  130</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WER</th>
<th>-</th>
<th>GFTE</th>
<th>FTE</th>
<th>FTE &amp; SLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>22.3</td>
<td>21.6</td>
<td>20.9</td>
<td>20.7  20.3</td>
</tr>
<tr>
<td>Test</td>
<td>21.0</td>
<td>20.1</td>
<td>19.4</td>
<td>19.1  19.2</td>
</tr>
</tbody>
</table>

German LM perplexity and WER for different types of supervision
⇒ Significant gain in transcription performance
⇒ Gain for SLT based supervision depends on number of languages used

(4) Acoustic Model Training: Results
- Decode training data with biased ASR
- Apply simple rule based filter to remove noisy and low confidence utterances
- 2 iterations of Viterbi training
- Add 3rd decoding pass to test new AM

<table>
<thead>
<tr>
<th>-</th>
<th>FTE &amp; SLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>Test</td>
</tr>
</tbody>
</table>

Original AM | 22.2 | 21.2 |
GFTE AM | 20.9 | 20.0 |
FTE AM | 20.7 | 19.9 |
FTE + SLT AM | 20.4 | 19.8 |

3rd pass WER with different acoustic models and applying either no supervision or FTE & SLT based supervision on dev/test

(5) Conclusion & Future Work
- 14.3% relative improvement in WER
- Future Work:
  - apply SLT based supervision on utterance level
  - include additional languages
  - more sophisticated filtering scheme based on ASR word confidences