Acoustic data driven pronunciation lexicon for speech recognition

Liang Lu

University of Edinburgh

23 May 2013
Motivation

- A lexicon is one of the key components for ASR
- Building a lexicon is expensive
- Maintain a lexicon is hard
  - New words, terms, name entities, ....
- Can we learn the lexicon automatically?
Grapheme to phoneme (G2P) conversion

- G2P conversion model is one of the standard methods
- But it requires large initial pronunciation lexicon
- Do not make use of the acoustic data

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>s</th>
<th>p</th>
<th>ee</th>
<th>ch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phoneme</td>
<td>s</td>
<td>p</td>
<td>iy</td>
<td>ch</td>
</tr>
</tbody>
</table>
Purely acoustic data driven approach

Park and Glass, "Unsupervised pattern discovery in speech", IEEE TASLP, 2008
Combination of the two methods

- A weak G2P model trained with limited samples
- Pronunciation refinement using the acoustic data

McGraw et.al "Learning new word pronunciations from spoken examples", IEEE TASLP, 2013.
Pronunciation mixture model

• A standard ASR framework:

\[
\hat{W} = \arg \max_W p(O|W) P(W) \tag{1}
\]

• With explicit pronunciation mixture model:

\[
\hat{W} = \arg \max_W P(W) \sum_B p(O|B) P(B|W) \tag{2}
\]

• EM algorithm can be used to learn the pronunciation weights
Joint acoustic and lexicon model training

- Train the G2P model
- Generate the lexicon
- Train the acoustic model
- Update the lexicon
System setup & results

- **Switchboard corpus**
  - Training data: 110h
  - Initial lexicon: 5K
  - Expert lexicon: 30K
- **MFCC+LDA_MLLT**
- **ML-GMM system**
- **No SAT**
Results cont.

Table: Switchboard results with 286h data.

<table>
<thead>
<tr>
<th>System</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>110h-derived lexicon</td>
<td>35.7</td>
</tr>
<tr>
<td>+ lexicon update</td>
<td>35.1</td>
</tr>
<tr>
<td>Expert lexicon</td>
<td>34.2</td>
</tr>
</tbody>
</table>
Conclusions

- Low-resource pronunciation modelling
- G2P model + acoustic information
- Make the lexicon adaptable rather than fixed
- Unsupervised initialisation — get rid of the expert lexicon