1. Introduction

Assessment of spoken English of non-native (L2) learners:
- Many people are learning English → want official qualifications
- To help meet this demand:
  - Automatically assess unscripted responses to prompts

Speech recognition is essential for assessment and error feedback
- Challenge: how to achieve good recognition accuracy?
  - wide variations from e.g. L1, proficiency level, recording conditions
  - spontaneous responses increase difficulty - disfluencies etc
  - "off-the-shelf" systems don’t work well
  - limited training data → "Limited Resource Language (LRL)"

- Lower proficiency level → larger L1 affect on pronunciations
- mismatch with pronunciation lexicons based on native speaker accents
- very costly (infeasible?) to tune lexicons to L1s
- Proposal: use graphemic lexicons
  - consistent improvements over phonetic systems for LRLs e.g. Babel
  - BUT English is highly irregular so graphemic systems generally worse
  - can this be used in pronunciation assessment?

2. Graphemic Lexicons

<table>
<thead>
<tr>
<th>Phonetic Lexicon</th>
<th>Approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABE'S</td>
<td>/eɪˈbəːm/</td>
</tr>
<tr>
<td>ABLE</td>
<td>/eɪˈbəːm/</td>
</tr>
<tr>
<td>ABOUT</td>
<td>/əˈbʌt/</td>
</tr>
<tr>
<td>ABOUT partial</td>
<td>/əˈbʌt/</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Graphemic Lexicon</th>
<th>Approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABE'S</td>
<td>/əˈbəːm/</td>
</tr>
<tr>
<td>ABLE</td>
<td>/Iˈbəːm/</td>
</tr>
<tr>
<td>ABOUT</td>
<td>/əˈbʌt/</td>
</tr>
<tr>
<td>ABOUT partial</td>
<td>/əˈbʌt/</td>
</tr>
</tbody>
</table>

- Root graphemes:
  - 26 letters of alphabet
  - Hesitations modelled by graphemes /GOO, G01/
- Attributes for context-dependent state tying:
  - word boundary information (I,M,F)
  - A apostrophe, P partial word
  - /a,e,i,o,u/ assigned to the vowel class
  - /vowel,y/ to the vowel class

3. Experimental Results: Automatic Speech Recognition

Data from Business Language Tests (BULATS)
- Up to 1 minute spontaneous responses to prompts
- ASR training data: 100 hours Gujarati L1 English speech
- Test sets: 225 speakers, A1-C grades
  - Gujarati L1 ‘Gujarati’ and 6 mixed L1s ‘Mixed’
- Decode
  - Stacked hybrid DNN-HMM acoustic model with trigram LM (HTK)

4. Experimental Results: Automatic Assessment

Data from Business Language Tests (BULATS)
- Grader training data: 1000 speakers Gujarati L1 English speech
- Test sets: as for ASR plus read speech and short prompt sections
- Pron features: 47 phones/1081 distances, or 26 graphemes/326 distances

5. Experimental Results: Automatic Assessment

<table>
<thead>
<tr>
<th>Decoder</th>
<th>Gujarati</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>(%)</td>
<td>(%)</td>
<td></td>
</tr>
<tr>
<td>PER</td>
<td>GER</td>
<td>PER</td>
</tr>
<tr>
<td>Ph</td>
<td>25.6</td>
<td>24.9</td>
</tr>
<tr>
<td>Gr</td>
<td>29.0</td>
<td>23.7</td>
</tr>
</tbody>
</table>

- GER is lower than PER
- PER increases with grapheme decode

6. Conclusions

- Non-native learner English: ‘limited resource language’:
  - large variation in spontaneous speech due to L1, proficiency etc
  - limited amount of training data
- Reduce lexical model mismatch with graphemic lexicon:
  - improves recognition performance
  - standard automatic grader performance equivalent to phonetic system
- Pronunciation distance features:
  - introduced to assess spontaneous speech pronunciations
  - graphemic-only system worse than phonetic system
  - phonetic features from graphemic output is successful