Automated Assessment of Speaking

Judith Fairbairn, British Council
Dr Talia Isaacs, University of Bristol
Background

• Testing speaking in school
  – Expensive
  – Time consuming
  – Lack of local expertise

• Washback effect: no test – no teaching

• Hybrid human-machine scoring model
  – Lower human rater cognitive load
  – Increase rating reliability for some linguistic features
Speech recognition

• Audio recording of an acoustic signal
• Transcription of the sounds
• Automated assessment of writing
• Pattern recognition using an algorithm
• Probability of occurrence
False alarms and misses
Machine training process

- Collect large volume of speech responses
- Raters score response
- Create training corpus
- Statistical models calculate probabilities
- Likelihood of each sound/word occurring
# Speaking test examples

<table>
<thead>
<tr>
<th>Test scoring</th>
<th>Test mode of delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>IELTS</td>
</tr>
<tr>
<td></td>
<td>Cambridge English exams</td>
</tr>
<tr>
<td>Human</td>
<td>TOEFL iBT</td>
</tr>
<tr>
<td></td>
<td>Aptis</td>
</tr>
<tr>
<td>Machine</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Pearson PTE</td>
</tr>
<tr>
<td></td>
<td>Pearson Versant</td>
</tr>
<tr>
<td></td>
<td>ETS SpeechRater</td>
</tr>
</tbody>
</table>
Recent findings

- Individual vowel/consonant sounds and fluency easiest to measure

- Narrow construct measured (e.g. Pearson) but construct-irrelevance avoided

- Wider construct measured (e.g. ETS) but only used for low-stakes testing
Recent findings

- Correlations between computer/rater and rater/rater can be similar.

- Top end of rating scale (e.g. C1/C2) difficult to mark.
Building an automated test

- Task design and item writing
- Sound quality
- Piloting and trialling
- Data collection
- Annotating responses
Native vs non-native pronunciation

- Local vs global
- Inner circle English vs World English
- Comprehensibility / intelligibility construct
- Expert users of Vietnamese and English model
Conclusion

• Hybrid human-machine scoring model

• Tool to train human raters

• Double marking

• Vietnamese pronunciation

Thank you

Judith Fairbairn
Productive Skills Test Researcher
British Council

judith.fairbairn@britishcouncil.org