Accent Characterisation and Recognition Using Self-Normalisation

Mark Huckvale
University College London

Accents Research

**Top-down Research**
- Specify accent groups
- Accent group for a speaker is “known”
- Specify which acoustic features to measure
- Hypothesise which phonetic/phonological characteristics important

**Bottom-up Research**
- Accent groups emerge by clustering speakers
- Best accent group for a speaker emerges from distance to cluster centres
- Useful acoustic features discovered by comparing speakers
- Phonetic/phonological characteristics emerge by studying clusters

Bottom-up Accents Research

- Key problem in data-driven accent research is the availability of a good metric for comparing the similarity of two people’s accents.
- Because any acoustic measures will be influenced by
  - vocal tract length, voice pitch, voice quality, speaking style and speaking rate
- as well as by
  - phonetic and phonological changes due to accent

Acoustic Comparison Problem

<table>
<thead>
<tr>
<th>after</th>
<th>tea</th>
<th>father</th>
<th>fed</th>
<th>cat</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.5</td>
<td>0.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Experimental Data

- Accents of British Isles corpus
- (10 male+10 female) × 14 accent areas:
  - Birmingham | Cornwall | East Anglia | East Yorkshire | Glasgow | Inner London | Lancashire | Liverpool |
  - Newcastle | North Wales | Dublin | Scottish Highlands | South East | Ulster
- Looking at 20 short sentences from each speaker
- Phonetically annotated by HMM forced alignment
- Generates about 130 vowel measurements from each of 270 speakers or about 35,000 data points

Formant Metric

- Divide each vowel into two halves
- Find median value of first four formants in each half
- Compare vowels in matching words across speakers
- Use Euclidean distance, average over all pairs
- Assess performance by finding accent group of nearest neighbour

<table>
<thead>
<tr>
<th></th>
<th>Any sex</th>
<th>Other sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>44.2%</td>
<td>27.4%</td>
</tr>
</tbody>
</table>
Normalised Formant Metric
• Normalise formant values for each speaker independently
  – so each formant has zero mean and unit variance
• Compare vowels in matching words across speakers
• Use Euclidean distance, average over all pairs

<table>
<thead>
<tr>
<th></th>
<th>Any sex</th>
<th>Other sex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50.4%</td>
<td>34.7%</td>
</tr>
</tbody>
</table>

Spectral Metric
• 19-channel auditory filterbank
• Subtract mean energy, add as 20th value
• Take mean over first & second half of vowel
• Compare vowels in matching words across speakers
• Use Euclidean distance, average over all pairs

<table>
<thead>
<tr>
<th></th>
<th>Any sex</th>
<th>Other sex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>71.2%</td>
<td>43.4%</td>
</tr>
</tbody>
</table>

Minematsu Innovation

Analyze a speaker’s accent by studying the acoustic distances between segment realizations.

Segment Distance Tables
• Compute distances between segment realisations for each speaker independently
• Then compare speaker distance tables not spectral properties of segment realisations
• “Self normalisation” accommodates speaker characteristics not due to accent

<table>
<thead>
<tr>
<th>Birmingham</th>
<th>South East</th>
</tr>
</thead>
<tbody>
<tr>
<td>as/after:</td>
<td>0.00 3.48 2.14</td>
</tr>
<tr>
<td>as/father:</td>
<td>3.48 0.00 3.62</td>
</tr>
<tr>
<td>as/cat:</td>
<td>2.14 3.62 0.00</td>
</tr>
</tbody>
</table>

Distance Table Metric
• Compute 130×130 vowel distance tables for each speaker
• Use Correlation distance on pairs of tables, between pairs of vowels from matching words
• Choose accent group of speaker with highest correlation

<table>
<thead>
<tr>
<th></th>
<th>Any sex</th>
<th>Other sex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80.3%</td>
<td>75.9%</td>
</tr>
</tbody>
</table>

Accent Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>bnm</th>
<th>crn</th>
<th>ean</th>
<th>eyk</th>
<th>gla</th>
<th>ilo</th>
<th>lan</th>
<th>lvp</th>
<th>ncl</th>
<th>nwa</th>
<th>roi</th>
<th>shl</th>
<th>sse</th>
<th>uls</th>
</tr>
</thead>
<tbody>
<tr>
<td>bnm:</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
| crn:    | 0    | 13   | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 2    | 0    | 3    | 0    | 48.4%
| ean:    | 0    | 2    | 8    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 7    | 0    | 48.1%
| eyk:    | 0    | 1    | 0    | 11   | 0    | 0    | 0    | 0    | 6    | 0    | 2    | 0    | 2    | 55.0%
| gla:    | 0    | 0    | 0    | 18   | 0    | 0    | 0    | 0    | 0    | 2    | 0    | 0    | 90.0%
| ilo:    | 0    | 0    | 1    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 1    | 0    | 1    | 89.0%
| lan:    | 0    | 0    | 0    | 0    | 19   | 0    | 0    | 0    | 1    | 0    | 0    | 0    | 95.0%
| lvp:    | 0    | 0    | 0    | 0    | 0    | 17   | 1    | 2    | 0    | 0    | 0    | 0    | 0    | 95.0%
| ncl:    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0.0%
| nwa:    | 0    | 1    | 0    | 0    | 0    | 1    | 0    | 0    | 0    | 1    | 0    | 19   | 0    | 0    | 95.0%
| roi:    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 19   | 0    | 0    | 0    | 0    | 0    | 95.0%
| shl:    | 0    | 4    | 2    | 0    | 0    | 2    | 0    | 0    | 0    | 0    | 8    | 0    | 0    | 50.0%
| sse:    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 20   | 100.0%

Most common confusions:
- East Anglia → South East
- E. Yorkshire → N. Wales
In summary, this study

- Introduced a new metric for comparing the similarity of accent of two speakers
- Calculated from an analysis of a set of sentences spoken by both speakers
- Based on the correlation of inter-segment distance tables
- Shows better accent recognition performance than a metric based on acoustic comparisons across speakers
- Is relatively unaffected by speaker characteristics
- Opens up further possibilities for bottom-up accent research