Selecting an acoustic correlate for automated measurement of /ɪ/ production in children

Heather Campbell,¹ Daphna Harel,² Elaine Hitchcock,³ and Tara McAllister¹

¹ NYU Steinhardt School of Culture, Education, & Human Development
² NYU Center for Promotion of Research Involving Innovative Statistical Methodology
³ Montclair Department of Communicative Sciences and Disorders

Meeting of the Acoustical Society of America, June 26, 2017
Outline

I: Introduction
   1: Why /i/?
   2: Visual acoustic biofeedback
   3: Automated scoring for /i/
   4: Several acoustic measures to consider

II: Methods
   1: Data collection
   2: Measurement
   3: Statistical modeling

III: Results and Discussion

IV: Conclusions and next steps
Speech sound disorders (SSD) can impede academic, social, and psycho-emotional development (Hitchcock et al., 2015).

For some children, errors are resolve spontaneously, but others require long-term clinical intervention (Flipsen, 2015).

- May persist through adolescence and, for 1-2% of individuals, into adulthood (Culton, 1986).

More than 50% of school-based speech-language pathologists (SLPs) report having discharged children with treatment-resistant errors from their caseloads (Ruscello, 1995).
I: Introduction
   1: Why /ɪ/?
   2: Visual acoustic biofeedback
   3: Automated scoring for /ɪ/
   4: Several acoustic measures to consider

II: Methods
   1: Data collection
   2: Measurement
   3: Statistical modeling

III: Results and Discussion

IV: Conclusions and next steps
Why /ɹ/?

- Misarticulation of American English rhotics are the most common and challenging to treat. (Shuster et al., 1995; Ruscello, 1995).
  - Among the latest-acquired speech sounds (Smit et al., 1990).
  - Articulatorily complex: simultaneous anterior and posterior lingual constrictions (Espy-Wilson, 1992) can be achieved with a variety of lingual contours (Delattre and Freeman, 1968).

- Despite articulatory variability, accurate /ɹ/ has stable acoustic properties (Delattre and Freeman, 1968; Hagiwara, 1995)
  - Low third formant frequency (F3) relative to other vowels
  - Second formant frequency (F2) that is close to F3.
Outline

I: Introduction
   1: Why /ɪ/?
   2: Visual acoustic biofeedback
   3: Automated scoring for /ɪ/
   4: Several acoustic measures to consider

II: Methods
   1: Data collection
   2: Measurement
   3: Statistical modeling

III: Results and Discussion

IV: Conclusions and next steps
Visual-acoustic biofeedback intervention

- Takes advantage of acoustic consistency of /i/.
  - Display real-time linear predictive coding spectrum representing vocal tract’s resonant frequencies.
    - Display target showing correct production of sound.
    - Learner modifies output to align formants with target.
    - Focus is on lowering F3 to match accurate /i/ target.

- Demonstrated efficacy in single case experimental studies.

(McAllister Byun, 2017; McAllister Byun and Campbell, 2016; McAllister Byun and Hitchcock, 2012)
Visual-acoustic biofeedback intervention
App-based acoustic biofeedback

- Significant barriers to uptake of tech-based interventions:
  - Cost of the required technology ($2K-$5K).
  - Accessibility and user-friendliness of the technology.
  - Not always a quick solution; may require intensive schedule.
- App under development, in piloting stage (McAllister Byun et al., 2017).
Outline

I: Introduction
   1: Why /ɪ/?
   2: Visual acoustic biofeedback
   3: Automated scoring for /ɪ/
   4: Several acoustic measures to consider

II: Methods
   1: Data collection
   2: Measurement
   3: Statistical modeling

III: Results and Discussion

IV: Conclusions and next steps
Motivation for automated scoring

- Gains in treatment do not readily generalize to contexts without biofeedback; longer treatment durations needed (Edeal and Gildersleeve-Neumann, 2011).

- **Home practice** may help increase the dosage of speech intervention while reducing the strain on SLP resources.
  - **Risk**: Without feedback from SLP, child will counterproductively reinforce incorrect speech patterns.

- **Current need**: Provide valid and reliable automated feedback and track progress during home practice with acoustic biofeedback.
The current study

- **Broad Goal**: Enable home practice with acoustic biofeedback through the incorporation of automated scoring.
- **Which acoustic measure corresponds best with clinician ratings of children’s /ɪ/ productions?**
- **Approach**: Compare models that include all possible acoustic values, with and without all possible interactions to find metric that best predicts accuracy.
Outline

I: Introduction
  1: Why /ɪ/?
  2: Visual acoustic biofeedback
  3: Automated scoring for /ɪ/
  4: Several acoustic measures to consider

II: Methods
  1: Data collection
  2: Measurement
  3: Statistical modeling

III: Results and Discussion

IV: Conclusions and next steps
Consider raw and derived measures for \(/\dot{\iota}/\)

- **F3:** Primary acoustic cue to rhoticity (Espy-Wilson et al., 2000).
  - Low height of F3 differentiates \(/\dot{\iota}/\) from acoustically similar sounds such as \(/l/\) and \(/w/\) (Polka and Strange, 1985).
- **F2:** Secondary acoustic cue to rhoticity (Polka and Strange, 1985).
  - F2 in close proximity to F3.
- Derived within-subject measures reflect the influence of both raw acoustic cues simultaneously (Flipsen et al., 2001; Lee et al., 1999).
  - F3-F2 Distance
  - F3/F2 Ratio

![Diagram of F3/F2](image)
Consider normalization relative to typical speaker data

- Raw and derived measures can be normalized relative to typical speaker data, e.g., Lee et al. (1999) for ages 5-19+.
  - Raw F2 and F3 means and SDs (Lee et al., 1999).
  - Derived F3-F2 and F3/F2 means and SDs (Flipsen et al., 2001).

<table>
<thead>
<tr>
<th>Age</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n^b )</td>
<td>F3 - F2</td>
</tr>
<tr>
<td>5 years</td>
<td>26</td>
<td>797 (343)</td>
</tr>
<tr>
<td>6 years</td>
<td>15</td>
<td>567 (152)</td>
</tr>
<tr>
<td>7 years</td>
<td>19</td>
<td>616 (138)</td>
</tr>
<tr>
<td>8 years</td>
<td>38</td>
<td>517 (175)</td>
</tr>
<tr>
<td>9 years</td>
<td>33</td>
<td>527 (145)</td>
</tr>
<tr>
<td>10 years</td>
<td>40</td>
<td>527 (169)</td>
</tr>
</tbody>
</table>
Consider interactions with acoustic measures

- Listeners may bring age- and sex-based expectations to a speech rating task that have the potential to interact with the properties of the raw acoustic signal.
  - Perceived age impacts accuracy ratings (Munson et al., 2010).
  - Perceived gender impacts accuracy ratings (Dart, 1991).

- Derivation and normalization may correct for some age- and sex-related differences (Flipsen et al., 2001), but it is unknown whether there are also interactions with these factors.
Outline

I: Introduction
   1: Why /ɪ/?
   2: Visual acoustic biofeedback
   3: Automated scoring for /ɪ/
   4: Several acoustic measures to consider

II: Methods
   1: Data collection
   2: Measurement
   3: Statistical modeling

III: Results and Discussion

IV: Conclusions and next steps
Participants

- Children receiving /ɪ/ treatment in 3 biofeedback studies.
  (McAllister Byun et al., 2014; McAllister Byun & Hitchcock, 2012; Hitchcock et al., in press)
  - Normal hearing and oral structure/function.
  - Word probes elicited throughout 8-10 weeks of intervention in a sound-shielded room with the CSL (KayPentax, Model 4150B)

<table>
<thead>
<tr>
<th>Study</th>
<th>Children</th>
<th>Ages (mean)</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acoustic (2012)</td>
<td>11</td>
<td>6-11 (9;0)</td>
<td>2109</td>
</tr>
<tr>
<td>Ultrasound (2014)</td>
<td>5</td>
<td>6-9 (7;8)</td>
<td>2926</td>
</tr>
<tr>
<td>EPG (2017)</td>
<td>6</td>
<td>6-10 (8;0)</td>
<td>1040</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>22</strong></td>
<td></td>
<td><strong>6075</strong></td>
</tr>
</tbody>
</table>

- Varied by phonetic context: Syllabic (808), Post-vocalic (1532), Singleton onset (774), Cluster onset (2961)
Outline

I: Introduction
  1: Why /ɪ/?
  2: Visual acoustic biofeedback
  3: Automated scoring for /ɪ/
  4: Several acoustic measures to consider

II: Methods
  1: Data collection
  2: Measurement
  3: Statistical modeling

III: Results and Discussion

IV: Conclusions and next steps
Ratings of perceptual accuracy

- Following the “industry standard” for perceptual rating in speech intervention studies (McAllister Byun et al., 2015), binary ratings were acquired in a blinded randomized fashion from 3 certified SLPs who exhibited at least 80% pairwise agreement.
  - Tokens were rated correct or incorrect.
  - Average across 3 raters was treated as an ordinal scale.
  - Ratings were unequally distributed across accuracy levels.

![Count of each accuracy rating](chart.png)
Acoustic measurement

- Trained graduate students measured formant frequencies from the minimum F3 in the rhotic interval of each word using Praat (Boersma and Weenink, 2014).
Outline

I: Introduction
   1: Why /ɪ/?
   2: Visual acoustic biofeedback
   3: Automated scoring for /ɪ/
   4: Several acoustic measures to consider

II: Methods
   1: Data collection
   2: Measurement
   3: Statistical modeling

III: Results and Discussion

IV: Conclusions and next steps
A series of ordinal mixed-effects regression models were fit on the aggregated data set while considering the following factors:

**Structural Variables**
- Fixed effects:
  - phonetic variant
  - age
  - sex
- Random effects:
  - child
  - word

**Acoustic Variables**
- Raw/Derived:
  - F3
  - F2
  - F3-F2
  - F3/F2
- Normed:
  - F3
  - F2
  - F3-F2
  - F3/F2

**Interaction Possibilities**
- none
- acoustics*age
- acoustics*sex
- acoustics*age & acoustics*sex

**32 models**
- 5 structural variables
- 1 acoustic variable
- 1 interaction possibility

**Select all**

**Select one**
Model selection

- Akaike & Bayesian Information Criteria (AIC/BIC) were used to select the best-fitting model.
  - Both take into account the number of predictors (Cohen et al., 2013).
  - BIC penalizes for each predictor, preferring fewer predictors.
  - Select the model with the lowest AIC and BIC.
- All analyses were conducted in R (RStudio, 2016).
  - Data compilation using 'tidyverse' packages (Wickham, 2016).
  - Regression models were fit using the “clmm” function in the 'ordinal' package (Christensen, 2015).
Results

- Controlling for age, sex, and phonetic context, the measure that accounted for the most variance in speech rating was F3-F2 distance normalized relative to a sample of age- and sex-matched speakers.
  - Higher normalized F3-F2 distance was associated with significantly lower accuracy ratings.
  - Best interaction possibility included acoustic variables interacting with both age and sex.

- Cluster onset tokens differed significantly from syllabic and vocalic targets.
Results

- Process for comparing all 32 models.
  - Best models among normalized metrics.

<table>
<thead>
<tr>
<th>Acoustic Measure Included in Model</th>
<th>Main Effects</th>
<th>Main Effects + Acoustics*Age</th>
<th>Main Effects + Acoustics*Sex</th>
<th>Main Effects + Acoustics*Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AIC</td>
<td>BIC</td>
<td>AIC</td>
<td>BIC</td>
</tr>
<tr>
<td>Normalised F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalised F3</td>
<td></td>
<td>7900.2</td>
<td>7980.7</td>
<td></td>
</tr>
<tr>
<td>Normalised F3-F2</td>
<td></td>
<td>7704.9</td>
<td>7778.7</td>
<td>7680.4</td>
</tr>
<tr>
<td>Normalised F3/F2</td>
<td></td>
<td>7617.3*</td>
<td>7704.5*</td>
<td>7672.0</td>
</tr>
</tbody>
</table>

Note: The highlighted values indicate the best models among normalized metrics.
Results

- Surprising that normalized version of F3-F2 performed better than non-normalized version of F3-F2.
- Limitations of normative data from Lee et al. (1999):
  - Based on 9-25 individuals in each age/sex group.
  - Speakers from a limited geographic region.
  - Only stressed vocalic /ɜ/ in the word “bird.”

<table>
<thead>
<tr>
<th>TABLE I. Distribution of subjects by age (in years) and gender.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td><strong>Male</strong></td>
</tr>
<tr>
<td><strong>Female</strong></td>
</tr>
<tr>
<td><strong>Total</strong></td>
</tr>
</tbody>
</table>
## Results

- Process for comparing all 32 models.
  - Best models among normalized metrics.
  - Best models among non-normalized metrics.

<table>
<thead>
<tr>
<th>Acoustic Measure Included in Model</th>
<th>Main Effects</th>
<th>Main Effects + Acoustics*Age</th>
<th>Main Effects + Acoustics*Sex</th>
<th>Main Effects + Acoustics<em>Age + Acoustics</em>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AIC</td>
<td>BIC</td>
<td>AIC</td>
<td>BIC</td>
</tr>
<tr>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F3</td>
<td></td>
<td>7871.1</td>
<td>7951.7</td>
<td></td>
</tr>
<tr>
<td>F3-F2</td>
<td>7752.0</td>
<td>7825.8</td>
<td>7739.7</td>
<td>7820.2</td>
</tr>
<tr>
<td>F3/F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalised F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalised F3</td>
<td></td>
<td>7900.2</td>
<td>7980.7</td>
<td></td>
</tr>
<tr>
<td>Normalised F3-F2</td>
<td>7704.9</td>
<td>7778.7</td>
<td>7680.4</td>
<td>7760.9</td>
</tr>
<tr>
<td>Normalised F3/F2</td>
<td>770.2</td>
<td>7775.2</td>
<td>7672.0</td>
<td>7752.5</td>
</tr>
</tbody>
</table>

* indicates a significant difference.
Conclusions

- For future automated scoring of children’s /r/ productions:
  - If normative data are appropriate, use the externally normalized F3-F2, in interaction with the child’s age and sex.
  - Otherwise, we recommend the non-normalized version of F3-F2, in interaction with age only.
- App-based treatment with automated scoring may facilitate increases in treatment dosage by allowing home practice.
Next steps

- Collect more representative normative values, including:
  - A larger sample of children.
  - A more geographically diverse sample.
  - Phonetic contexts other than the syllabic rhotics.

- Improve current aggregated data set:
  - Obtain gradient ratings rather than binary ratings (McAllister Byun et al., 2016; Schellinger et al., 2016; Munson et al., 2012, 2017).
  - Obtain crowd-sourced ratings from naïve listeners (McAllister Byun et al., 2015), which may differ from SLP ratings (Klein et al., 2012).
  - Include potential control for different phonetic context: duration (Klein et al., 2012).
References


RStudio (2016). Rstudio: integrated development for R.


Thank you!

Questions?

heather.campbell@nyu.edu

Acknowledgment

Thanks to members of the Biofeedback Intervention for Speech Lab and the Montclair speech lab for assistance with data collection and processing.