The acoustics of [voice] in infant-directed speech and implications for phonological learning

Chandan Narayan
University of Toronto

Kyle Gorman
University of Pennsylvania

Daniel Swingley
University of Pennsylvania

BUCLD 33
Conflicting facts

- Infant-directed speech (IDS) provides the learner with enhanced cues to phonological contrast
  - Expanded $F1 \times F2$ space (Kuhl et al., 1997)
  - Larger $f_0$ range in tones (Liu et al. 2007)
  - Modally distributed cues to vowel length and quality (Werker et al., 2007)

- In early infancy, caregivers’ production of VOT shows more overlap between [+voice] and [-voice] than in late infancy (Sundberg & Lacerda, 1999; Baran et al., 1977)
How do infants get [voice]?
A challenge to enhancement

• Plenty of evidence that by 10-12 months, infants “know” the phonological categories (including [voice]) of their ambient language (Werker & Tees, 1984; numerous others)

• So, if IDS is providing infants with sloppy acoustic cues to phonological contrast, how might infants get to be so good at what they do?
Secondary cue to [voice]: $f_0$ perturbation

**Acoustics**
- Following [+voice] stops, the fundamental frequency ($f_0$) of the vowel in a CV syllable is lower than when following [-voice] stops

**Perception**
- When VOT is ambiguous, listeners report [+voice] when V has low $f_0$ and [-voice] when V has high $f_0$ (Lisker & Abramson, 1970; Abramson & Lisker, 1980; many others)
$f_0$ control

- Kingston & Diehl (1994; Francis et al., 2007) suggest that listeners control $f_0$, giving listeners extra low-frequency information in the vicinity of the stop closure for [+voice] perception.

- $f_0$ is an “enhance-able” acoustic feature that mothers might exploit when conveying [voice] information to their children.
Corpora

• **Brent Corpus (IDS)** (Brent and Siskind, 2001)
  – 4 mothers using infant-directed English speech
  – Natural mother-infant interactions at 9 months of age
  – 500 utterances/mother → over 1200 word-initial CVs

• **Buckeye Corpus (ADS)** (Pitt et al., 2007)
  – 4 women (3 with young infants, 1 with an older toddler) speaking a Midlands dialect with an adult
  – 20 minutes of speech/speaker → over 1000 CVs
Results: VOT

• Register (IDS vs. ADS) x Voicing ([+voice] vs. [-voice]) x Place (bilabial, alveolar, velar) ANOVA

• Voicing x Register interaction
  – Suggests that the difference in VOT between [+voice] and [-voice] is greater in ADS than in IDS
Results: VOT

• VOT in each register was used in a linear discriminant analysis to classify [voice]

<table>
<thead>
<tr>
<th></th>
<th>Predicted (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[+voice]</td>
</tr>
<tr>
<td>IDS</td>
<td></td>
</tr>
<tr>
<td>[+voice]</td>
<td>72.7 (n = 365)</td>
</tr>
<tr>
<td>[-voice]</td>
<td>11.2 (n = 78)</td>
</tr>
<tr>
<td>ADS</td>
<td></td>
</tr>
<tr>
<td>[+voice]</td>
<td>88.3 (n = 432)</td>
</tr>
<tr>
<td>[-voice]</td>
<td>10.9 (n = 62)</td>
</tr>
</tbody>
</table>
Results: $f_0$ perturbation

- Peak $f_0$ (z-Mel) following release of stop

- Voicing x Register ANOVA

- Significant effect of [voice] on $f_0$ but no interactions
  - Mean $f_0$ is higher following [-voice] stops
    $[+\text{voice}] = 0.16$ vs. $[-\text{voice}] = 0.38$
Modeling [voice]

• Predict presence of [voice] given VOT and $f_0$

• Hierarchical logistic regression
  – Allow speakers to vary in implementations of VOT and $f_0$ cues to [voice]
  – Random (per-subject) slopes for VOT and $f_0$
  – We’ll attempt to interpret subject effects to show the relative degree to which subjects implement the two cues
## Hierarchical model

|                  | Estimate | Std. Error | z value | Pr(>|z|) |
|------------------|----------|------------|---------|----------|
| (Intercept)      | 2.4272   | 0.1166     | 20.82   | <2e-16   |
| VOT              | -6.3426  | 0.4264     | -14.87  | <2e-16   |
| $f_0$            | -0.0872  | 0.1352     | -0.64   | 0.519    |
| VOT:$f_0$        | -1.1257  | 0.5419     | -2.08   | 0.038    |

- No main effect of $f_0$ but it inversely covaries with VOT, as expected

- The effect size for $f_0$ is equivalent to about a semitone in mean pitch region
## Random effects

<table>
<thead>
<tr>
<th></th>
<th>VOT</th>
<th>$f_0$</th>
<th>VOT:$f_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.3072</td>
<td>0.193</td>
<td>-1.10</td>
</tr>
<tr>
<td>2</td>
<td>0.3237</td>
<td>0.260</td>
<td>-1.47</td>
</tr>
<tr>
<td>3</td>
<td>-0.4676</td>
<td>0.073</td>
<td>-0.38</td>
</tr>
<tr>
<td>4</td>
<td>-0.1164</td>
<td>-0.218</td>
<td>1.20</td>
</tr>
<tr>
<td>5</td>
<td>-0.0031</td>
<td>0.085</td>
<td>-0.49</td>
</tr>
<tr>
<td>6</td>
<td>-1.3383</td>
<td>-0.232</td>
<td>1.31</td>
</tr>
<tr>
<td>7</td>
<td>0.9212</td>
<td>-0.119</td>
<td>0.68</td>
</tr>
<tr>
<td>8</td>
<td>1.3105</td>
<td>-0.076</td>
<td>0.43</td>
</tr>
</tbody>
</table>
Register-specific (pooled) models of [voice]

<table>
<thead>
<tr>
<th></th>
<th>IDS</th>
<th>ADS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$</td>
<td>Std. Error</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>VOT ratio</td>
<td>-6.39</td>
<td>0.40</td>
</tr>
<tr>
<td>Peak $f_0$</td>
<td>-0.59</td>
<td>0.10</td>
</tr>
<tr>
<td>VOT ratio x Peak $f_0$</td>
<td>-2.03</td>
<td>0.44</td>
</tr>
</tbody>
</table>
Predicting [voice]: Discussion

• Given VOT and $f_0$:
  – VOT contributes to a [voice] prediction in both IDS and ADS
  – $f_0$ contributes more to a [voice] prediction in IDS than in ADS!

• There is enough consistent VOT information in ADS which essentially overrides the $f_0$ regularity

• When VOT is highly variable (with more overlap between categories) as in the case of IDS, $f_0$ information is useful in predicting [voice]
Misclassification analysis

• What is the practical import of the logistic models?

• The LR models were used as a [voice] classifier with only VOT as a predictor
Misclassifications using only VOT

- Twice as many misclassifications in IDS than in ADS
- Majority of misclassifications occur in the 0-0.5 range, which corresponds to the region of overlap in VOT

When misclassified tokens were \textit{re-classified} using $f_0$,

- 69\% of IDS tokens were correctly classified ($p < 0.001$)
- 52\% of ADS tokens were correctly classified ($p = 0.79$)
Take home message

• $f_0$ in the IDS sample preserves [voice] information when VOT information alone is ambiguous

• The emergence of $f_0$ as a stable contributor to [voice] prediction suggest a covert contrast that the learner might recover
Take this home too!

- So far, distributional models of phonetic category learning, that rely on *enhancement* as a hallmark of learning, are dimensionally *flat* when it comes to multiple cues to phonologically relevant features such as [voice]
A comprehensive theory of phonetic category learning (in infancy) must consider:

– The changing nature of the IDS across development: [voice] as presented to 9 month olds is different from [voice] to 14 month olds

– Multiple cues to relevant features: such as covarying VOT and $f_0$ for [voice], $F2$ onset and burst frequency for place, etc.
Thank you

• Chandan Narayan
  chandan.narayan@utoronto.ca

• Kyle Gorman
  kgorman@ling.upenn.edu

• Daniel Swingley
  swingely@psych.upenn.edu

Funded by IGERT(NSF) to CN & KG and the NIH to DS. Thanks to Penn Baby Lab.