Talker variation and systematicity in voice onset time: A corpus study
Eleanor Chodoroff and Colin Wilson
Department of Cognitive Science, Johns Hopkins University

Introduction
Talkers vary considerably in the acoustic realization of speech sounds, as demonstrated in studies of vowels (Peterson & Barney, 1952), fricatives (Newman et al., 2001), and stop consonants (Allen et al., 2003; Theodore et al., 2009).

How is acoustic-phonetic variation structured across and within speakers?

This study investigates variation in positive voice onset time (VOT) of voiced and voiceless stops in American English.

- Large corpus of read sentences from > 100 talkers
- Talkers differ in mean VOT of voiceless stops in particular, but strong correlations hold among talker-specific values.
- Ex. VOTs of [p] and [k] covary across talkers ($r = 0.82$)
- Also find cross-voice VOT correlations (e.g., [p] and [d]), and a positive relationship between VOT mean and sd for each stop.

Bayesian modeling of perceptual adaptation indicates that structured variation facilitates reliable estimation of talker means from minimal exposure, as well as generalization beyond the input (e.g. Clarke & Garrett, 2004; Kraljic & Samuels, 2005, 2006; Nielsen, 2007; Theodore et al., 2010).

Methods

Mixer-6 Corpus of read speech

- Randomly selected utterances from the Switchboard corpus
- Sentence length 1-17 words (median 7 words)
- Transcribed at three separate times by each of ~500 speakers (~15 min. per session)
The present analysis was performed over a coded subset of Mixer-6.

Participants:

- 129 native AE speakers
- Place of birth: 68 speakers from Pennsylvania; 32 speakers from other mid-Atlantic and New England regions
- 29 speakers from other US states
- Gender: 60 male, 69 female
- Age: 19-87 years old
- Medium 27 years old

Number of Tokens

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>T</th>
<th>K</th>
<th>B</th>
<th>D</th>
<th>G</th>
<th>Total N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speakers</td>
<td>9285</td>
<td>5821</td>
<td>11492</td>
<td>12762</td>
<td>17459</td>
<td>11637</td>
<td>68,400</td>
</tr>
</tbody>
</table>

Corpus Preparation

- Read speech audited for reading and recording errors through automatic and manual methods
- Cleaned transcript files aligned to audio using the Penn Phonetics Lab Forced Aligner
- Identified all-word-initial, prevoicing stops for VOT measurement (function words retained in the analysis, with the exception of ‘a’).

AutoVOT

AutoVOT automatically detects stop release and following vocalic onset, standard boundaries for positive VOT measurements (Kessler et al., 2014; Sundinberget & Kessler, 2012; Stuart-Smith et al., in press)

- All stops aligned using the default AutoVOT acoustic models
- Minimum VOT duration:
  - Voiced stops: 15 ms
  - Voiced stops: 4 ms

- Window of analysis:
  - Voiced stops: PFA boundaries ± 30 ms
  - Voiced stops: PFA boundaries ± 10 ms

- Reasonable agreement between automatic measurements and a validation set of 3,000 manually-extracted values (RMSE = 1.29 ms)

- Values 2.5 std away from the population grand mean were excluded from analysis

Population Values

<table>
<thead>
<tr>
<th></th>
<th>Stop</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>51.1</td>
<td>9.4</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>61.5</td>
<td>9.1</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>54.8</td>
<td>7.2</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>8.7</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>13.9</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>17.4</td>
<td>3.0</td>
<td></td>
</tr>
</tbody>
</table>

Cross-Place Correlations

Results

<table>
<thead>
<tr>
<th></th>
<th>Stop</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>51.1</td>
<td>9.4</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>61.5</td>
<td>9.1</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>54.8</td>
<td>7.2</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>8.7</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>13.9</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>17.4</td>
<td>3.0</td>
<td></td>
</tr>
</tbody>
</table>

Discussion

- Knowledge of voiceless stop VOT correlations supported by evidence from perceptual adaptation and phonetic imitation
- Listeners may rely on knowledge of a long [k] without hearing the talker produce the [k] category (Theodore et al., 2010)
- In imitation, listeners extrapolate a talker’s characteristically long VOT of [p] to [k] without prior exposure (Nielsen, 2007)
- Many previous studies have been limited in the extent to which they can explore cross-talker patterns in VOT (c.f. Yao, 2007; Theodore et al., 2009)
- Too few speakers (e.g., Abramson & Lisker, 1964; Zue, 1976; Cole et al., 2007)
- Not enough tokens per speaker (e.g., Byrd, 1993)
- Correlations between means and sds suggest a non-Gaussian distribution (e.g., gamma distribution, Goldrick et al., 2011)

- Listeners may rely on knowledge of structured variation to extrapolate from limited talker-specific evidence and refine talker-specific model with further exposure
- Adaptation via structured variation is similar to extrinsic normalization procedures, which employ information across many speech sounds (e.g., multiple vowels).

- Standard extrinsic normalization procedures, however, assume that cross-category data will come from the speaker at hand (Geertsema, 1968; Lobanov, 1971; Nearey, 1978, 1989)

Future Directions

- Determine correlations of VOT with other acoustic-phonetic cues to stop consonant place and voice
- Correlations with other acoustic-phonetic cues may facilitate talker adaptation and subsequent categorization
- Explore possible sources of VOT variation/covariation (e.g., speaking rate, physiology)
- Examine structured variability within and across other contexts beyond word-initial stops in stressed syllables
- How is structured variability manifested across other speech sounds (vowels, fricatives, etc.)
- Do the same patterns of structured variability emerge in spontaneous speech?
- Train AutoVOT models on hand-annotated data from Mixer-6 for possible improvement in the measurement
- While previous studies have demonstrated listener knowledge of cross-place correlations, do listeners have knowledge of cross-voice correlations?
- Integration with other models of perceptual learning and adaptation (e.g., Nielsen & Wilson, 2008; Kleinmich & Jaeger, 2011, 2015; Pajak et al., 2013)

Acknowledgments

We would like to thank Jack Gooffré, Sanjeev Khadupar, Paul Smolensky, and Matt Goldrick for their useful input on acoustic analysis and overall project goals. We’d also like to thank Na-Min Maciejczyk, Wade Shao, Sharon Tian, Chloe Holland, Eliehu Abraham, Aleksandra Golkin, Spandana Mandalapure, and Benjamin Wang for their help in data processing. Additional thanks goes to the NYU Phonetics and Experimental Phonology Lab for their insightful comments and suggestions. Finally, we acknowledge the NIH-NSF Fonetic Services Division for supporting this research.