Covariation of Stop Consonant Acoustics: Corpus Evidence and Implications for Talker Adaptation

Eleanor Chodroff and Colin Wilson

Johns Hopkins University
Department of Cognitive Science
Individual talkers vary significantly in the acoustic-phonetic realization of speech sounds

- Stop consonant voice onset time (VOT)
- Vowel formants
- Fricative spectral shape
- Glottalization
- etc.

E.g., Allen et al., 2003; Theodore et al., 2007, 2009; Yao, 2007; Peterson and Barney, 1952; Newman et al., 2001; Redi and Shattuck-Hufnagel, 2001

Many sources of variability in the speech signal:
- Phonetic category
- Contextual and global effects (e.g., speaking rate, word frequency, prosodic position)
- Talker (e.g., gender, dialect, sociolect, idiolect)
Structured variability

Listeners adapt to new talkers with relative ease in spite of variation

e.g., Clarke & Garrett, 2004; Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2006; Maye, Aslin, & Tanenhaus, 2008; Norris, McQueen, & Cutler, 2003; Bradlow and Bent, 2008

Structured variability:
Rapid and general adaptation to novel talkers will be facilitated by the knowledge of systematicity in how talkers vary.

- talker differences are not entirely random but obey strong regularities
- covariation of acoustic-phonetic cues across/within phonetic categories
  (cf. covariation of speech patterns across/within social classes; Labov, 1966)

Ex: a talker with a higher VOT for /p/ expected to have higher VOT for /t, k/
Evidence for structured variability

Covariance of talker means across vowels
Coordinate system (Joos, 1948) or frame of reference (Nearey, 1989)

Fig. 8. Frequency of second formant versus frequency of first formant for ten vowels by 76 speakers.

Evidence for structured variability

Covariance of talker means across stops

Scobie, 2008; Theodore et al., 2009; Yao, 2009
Structured variability in stop consonants

<table>
<thead>
<tr>
<th></th>
<th>[p^h]</th>
<th>[t^h]</th>
<th>[k^h]</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOT*</td>
<td>64</td>
<td>70</td>
<td>65</td>
</tr>
<tr>
<td>f0</td>
<td>213</td>
<td>210</td>
<td>222</td>
</tr>
<tr>
<td>rel. amplitude</td>
<td>16</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>mean frequency</td>
<td>2087</td>
<td>4053</td>
<td>2103</td>
</tr>
<tr>
<td>F1 onset*</td>
<td>485</td>
<td>510</td>
<td>500</td>
</tr>
<tr>
<td>vowel duration</td>
<td>113</td>
<td>89</td>
<td>96</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

\* = hypothetical values

*VOT* = Vocal Onset Time

Structured variability in stop consonants.
Outline

1. Introduction

2. Methods
   1. Mixer 6 Corpus
   2. Stop Consonant Measurements

3. Structured Variability
   1. Cross-category Correlations
   2. Within-category Correlations

4. Bayesian Model of Talker Adaptation

5. Discussion/Conclusion
Mixer 6 Corpus

Corpus

Read speech – utterances selected from Switchboard

Each speaker read the same sentences

Utterance length: 1-17 words (median: 7)

3 separate sessions, ~15 minutes each
~96 hours of speech

Available from the LDC

Speakers

129 native English speakers

69 female, 60 male

Age: 19 – 87 years old (median: 27)

Place of birth:
Pennsylvania: 68
Other mid-Atlantic and New England regions: 32
Other areas of the United States: 29

cf. corpus studies from: Byrd, 1993; Cole et al., 2004; Yao, 2007; Yuan & Liberman, 2008; Davidson, 2011; Gahl et al., 2012; Labov et al., 2013; Elvin & Escudero, 2015; Stuart-Smith et al., in press
Pre-processing

Reading and recording errors removed with a mixture of automatic and manual methods.

Automatic pre-processing with Penn Forced Aligner and AutoVOT


**AutoVOT:** locates onset of stop burst and following vowel

Measurement reliability:
- Manually measured VOT\(^+\) of \(~3000\) tokens
- RMSE = 12.9ms
- Population mean VOT\(^+\)s within range of that found in other studies
  (Lisker & Abramson, 1964; Zue, 1976; Byrd, 1993; Yao, 2007)

Additional \(~900\) tokens manually measured

Outlier exclusion threshold: \(\pm 2.5\) standard deviations from talker mean
Acoustic-Phonetic Cues of Interest

**Voice onset time (VOT\(^+\)):** duration from stop release to start of voicing
Focusing on *positive* voice onset time

N = 69,070 stops (outliers excluded)

* Primary cue to stop voicing
  (Lisker & Abramson, 1964)

* Secondary cue to stop place of articulation
  (Klatt, 1975)

---

**Spectral center of gravity (COG):** energy-weighted average frequency of initial stop burst spectrum (smoothed)

N = 70,430 stops (outliers excluded)

* Primary cue to stop place of articulation
  (Winitz et al. 1972; Blumstein & Stevens, 1979)

* Secondary cue to stop voicing
  (Halle et al., 1957; Chodroff & Wilson, 2014)
Acoustic-Phonetic Cues of Interest

Spectral center of gravity (COG): energy-weighted average frequency of initial stop burst spectrum (smoothed)

- computed 64-point FFT for seven consecutive 3ms Hamming windows, shifted by 1ms
- first window centered on stop release
- power spectral densities averaged and COG computed on the smoothed spectrum

Hanson & Stevens, 2003; Flemming, 2007; Chodroff & Wilson, 2014
Acoustic-Phonetic Cues of Interest

f0

• first Praat-detected f0 at vowel onset (within 50 ms of stop offset)

\[ N = 52,887 \text{ stops (outliers excluded)} \]

* Secondary cue to stop voicing

(Haggard et al., 1970; Ohde, 1984; Whalen et al., 1990)

Following vowel duration (vdur)

• vowel onset defined by AutoVOT boundary; vowel offset by Penn Forced Aligner boundary

\[ N = 69,223 \text{ stops (outliers excluded)} \]

* Secondary cue to stop voicing

(Summerfield, 1981; Allen & Miller, 2004)
Stop Consonants for VOT$^+$ Analysis

69,070 word-initial prevocalic stop consonants
320 – 741 stop consonants per talker (median: 547)

Number of Tokens Per Talker

<table>
<thead>
<tr>
<th>Stop</th>
<th>Range</th>
<th>Median</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>47 – 98</td>
<td>77</td>
<td>9,686</td>
</tr>
<tr>
<td>T</td>
<td>17 – 77</td>
<td>46</td>
<td>5,906</td>
</tr>
<tr>
<td>K</td>
<td>55 – 114</td>
<td>93</td>
<td>11,765</td>
</tr>
<tr>
<td>B</td>
<td>70 – 138</td>
<td>99</td>
<td>12,681</td>
</tr>
<tr>
<td>D</td>
<td>70 – 192</td>
<td>140</td>
<td>17,441</td>
</tr>
<tr>
<td>G</td>
<td>59 – 122</td>
<td>91</td>
<td>11,591</td>
</tr>
</tbody>
</table>

Word types


*Function words except “to” retained in the analysis
Variation in Talker Means for VOT$^+$ (ms)
Cross-Place Correlations of Talker Means:
Voiceless (long-lag) Stops

Each point = talker mean
In brackets: 95% CIs based on 1000 bootstrap replicates
All ps < 0.0003 (alpha-corrected) unless otherwise indicated
Cross-Place Correlations of Talker Means:
Voiced (short-lag) Stops

Each point = talker mean
In brackets: 95% CIs based on 1000 bootstrap replicates
All ps < 0.0003 (alpha-corrected) unless otherwise indicated
Cross-Voice Correlations of Talker Means:

Each point = talker mean
In brackets: 95% CIs based on 1000 bootstrap replicates
All $p$s < 0.0003 (alpha-corrected) unless otherwise indicated
Spectral Center of Gravity (Hz)

- **P – B**
  - 95% CI: [0.52, 0.73]

- **T – D**
  - 95% CI: [0.61, 0.78]

- **K – G**
  - 95% CI: [0.59, 0.80]

- **P-T**
  - 0.44
  - [0.29, 0.56]

- **T-K**
  - 0.52
  - [0.38, 0.63]

- **K-P**
  - 0.57
  - [0.44, 0.66]

- **B-D**
  - 0.55
  - [0.33, 0.68]

- **D-G**
  - 0.68
  - [0.58, 0.77]

- **B-G**
  - 0.61
  - [0.48, 0.72]

Each point = talker mean
All *p* < 0.0003 (alpha-corrected) unless otherwise indicated
Each point = talker mean
All $p$s < 0.0003 (alpha-corrected) unless otherwise indicated
Vowel Duration (ms)

Each point = talker mean
All $p$s < 0.0003 (alpha-corrected) unless otherwise indicated
Outline

1. Introduction

2. Methods
   1. Mixer 6 Corpus
   2. Stop Consonant Measurements

3. Structured Variability
   1. Cross-category Correlations
   2. Within-category Correlations

4. Bayesian Model of Talker Adaptation

5. Discussion/Conclusion
Systematic relations among phonetic properties
Trading relations vs phonetic enhancement

Token-by-token correlations

(Schultz et al., 2012; Beddor et al., 2013; Dmitrieva et al., 2015; Kirby and Ladd, 2015; Clayards, submitted)

Talker level correlations

(Nearney, 1989; Nearney and Assmann, 2007; Solé & Ohala, 2010; Beddor et al., 2013; Clayards, submitted)
# Correlations Within-Category

Correlations between talker-specific means within a stop category

<table>
<thead>
<tr>
<th></th>
<th>VOT x COG</th>
<th>VOT x f0</th>
<th>VOT x vdur</th>
<th>COG x f0</th>
<th>COG x vdur</th>
<th>f0 x vdur</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0.32*</td>
<td>-0.02</td>
<td>-0.19</td>
<td>-0.07</td>
<td>-0.01</td>
<td>0.14</td>
</tr>
<tr>
<td>t</td>
<td>0.34*</td>
<td>0.04</td>
<td>-0.09</td>
<td>0.08</td>
<td>0.07</td>
<td>0.17</td>
</tr>
<tr>
<td>k</td>
<td>0.25</td>
<td>0.18</td>
<td>-0.17</td>
<td>0.15</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>b</td>
<td>0.33*</td>
<td>-0.21</td>
<td>-0.14</td>
<td>0.10</td>
<td>-0.12</td>
<td>-0.08</td>
</tr>
<tr>
<td>d</td>
<td>0.70*</td>
<td>-0.11</td>
<td>-0.05</td>
<td>0.38*</td>
<td>-0.07</td>
<td>-0.16</td>
</tr>
<tr>
<td>g</td>
<td>0.50*</td>
<td>0.01</td>
<td>-0.25</td>
<td>0.33*</td>
<td>0.06</td>
<td>-0.23</td>
</tr>
</tbody>
</table>

* p < 0.001
Outline

1. Introduction

2. Methods
   1. Mixer 6 Corpus
   2. Stop Consonant Measurements

3. Structured Variability
   1. Cross-category Correlations
   2. Within-category Correlations

4. Bayesian Model of Talker Adaptation

5. Discussion/Conclusion
Bayesian Model of Talker Adaptation

Acoustic-phonetic evidence suggests that covariance within and across stop acoustics may facilitate rapid adaptation to novel talkers

Prior knowledge:
\[ p(m) = \mathcal{N}(m; \mu_{\text{pop}}, \Sigma_{\text{pop}}) \]

Adaptation to a novel talker:
Estimate posterior probability over talker means for each cue and stop
\[ q_0(m) = p(m|\mu, \Sigma) \]
\[ q_i(m) \propto p(x_i|m, l_i, \Sigma_{\text{within.talker}}) * q_{i-1}(m) \]

\( m \) = vector of talker-specific means (one entry per stop-cue combo)
\( \mu_{\text{pop}} \) = mean of \( m \) across the population,
\( \Sigma_{\text{pop}} \) = variance/covariance matrix on \( m \) in the population
\( (x_i, l_i) \) = one stop production from the talker (acoustic cues, label)
Complete Independence Model

Complete Covariance Model
### Covariance vs Independence Models

<table>
<thead>
<tr>
<th># of exposures</th>
<th>avg. density ratio</th>
<th>$\beta$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>64.07</td>
<td>-2.08</td>
<td>-8.71</td>
</tr>
<tr>
<td>20</td>
<td>18.17</td>
<td>-1.45</td>
<td>-9.58</td>
</tr>
<tr>
<td>30</td>
<td>8.17</td>
<td>-1.05</td>
<td>-9.85</td>
</tr>
<tr>
<td>40</td>
<td>4.76</td>
<td>-0.78</td>
<td>-7.36</td>
</tr>
<tr>
<td>50</td>
<td>3.06</td>
<td>-0.56</td>
<td>-6.15</td>
</tr>
</tbody>
</table>

- $\Sigma_{pop}$
- $\Sigma_{covar}$

**Graph**

- Log density of talker mean vs number of exposure stops.
Bayesian Model of Talker Adaptation: Application

Perceptual Generalization across Phonetic Categories

Listeners generalize a talker’s characteristic VOT across stop categories. (Eimas & Corbit, 1973; Theodore & Miller, 2010; Nielsen, 2011)
Lengthened VOT condition, $\Sigma_{\text{pop covar}}$

Lengthened VOT condition, $\Sigma_{\text{pop indep}}$

Shortened VOT condition, $\Sigma_{\text{pop covar}}$

Shortened VOT condition, $\Sigma_{\text{pop indep}}$
Implications

Covariance relations across speech sounds can be used as a prior to refine a talker-specific model.

Implications for models of perceptual adaptation and generalization: Norris et al., 2003; Nielsen & Wilson, 2008; Kleinschmidt & Jaeger, 2011, 2015; McMurray & Jongman, 2011; Pajak et al., 2013

In line with results from perceptual generalization and phonetic imitation:

- Identify a long /k/ as more characteristic of a talker with a long /p/ even without hearing the talker produce the /k/ category (Theodore & Miller, 2010)

- Produce longer VOT for /k/ after exposure to lengthened VOT for /p/ (Nielsen, 2011)

(see also Eimas & Corbit, 1973)

Caveat: correlations are not perfect, so there is still room for talker-specific fine-tuning.
Conclusion

Cross-category means are highly correlated: VOT, COG, f0, following vowel duration

Examined in a large corpus of more natural (non-laboratory) speech in all 6 stop consonants

If listeners track them, they can adapt to talkers in a way that is more efficient and robust to noise, and that generalizes from one sound to another

Experimental results are consistent with rapid, generalized adaptation
Future Directions

What underlies the acoustic-phonetic correlations?

- physiological factors
- dialectal/sociophonetic
- phonology-phonetics interface
  - correlations guided by phonological features?
  - featural specification provides intermediate representation between individual speech sounds and all other sounds

Explore cross-talker patterns in other speech sounds and languages

Investigate cognitive status of correlations with other talker adaptation experiments
Thanks to:

Alessandra Golden
Jack Godfrey
Sanjeev Khudanpur

Audiences at:
JHU Center for Language and Speech Processing
NYU Phonetics and Experimental Phonology Lab
169th Acoustical Society of America
18th International Congress of Phonetic Sciences

Science of Learning Institute –
Johns Hopkins University

Department of Homeland Security –
USSS Forensic Services Division
Thank you!
### Correlations Within-Category: Token-by-token

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>D</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOT vs. COG</td>
<td>-0.18 – 0.70</td>
<td>-0.14 – 0.73</td>
<td>-0.11 – 0.81</td>
</tr>
<tr>
<td></td>
<td>mean: 0.30*</td>
<td>mean: 0.46*</td>
<td>mean: 0.49*</td>
</tr>
<tr>
<td>VOT vs. f0</td>
<td>-0.33 – 0.37</td>
<td>-0.38 – 0.29</td>
<td>-0.47 – 0.31</td>
</tr>
<tr>
<td></td>
<td>mean: -0.01</td>
<td>mean: -0.08*</td>
<td>mean = -0.04</td>
</tr>
<tr>
<td>VOT vs. vdur</td>
<td>-0.32 – 0.23</td>
<td>-0.27 – 0.35</td>
<td>-0.20 – 0.34</td>
</tr>
<tr>
<td></td>
<td>mean: -0.06*</td>
<td>mean: 0.01</td>
<td>mean: 0.10*</td>
</tr>
<tr>
<td>COG vs. f0</td>
<td>-0.40 – 0.45</td>
<td>-0.52 – 0.45</td>
<td>-0.42 – 0.41</td>
</tr>
<tr>
<td></td>
<td>mean: -0.03</td>
<td>mean: -0.07*</td>
<td>mean: -0.01</td>
</tr>
<tr>
<td>COG vs. vdur</td>
<td>-0.26 – 0.41</td>
<td>-0.34 – 0.31</td>
<td>-0.40 – 0.32</td>
</tr>
<tr>
<td></td>
<td>mean: 0.03</td>
<td>mean: 0.00</td>
<td>mean: 0.04</td>
</tr>
<tr>
<td>f0 vs. vdur</td>
<td>-0.44 – 0.24</td>
<td>-0.53 – 0.25</td>
<td>-0.58 – 0.20</td>
</tr>
<tr>
<td></td>
<td>mean: -0.10*</td>
<td>mean: -0.19*</td>
<td>mean: -0.20*</td>
</tr>
</tbody>
</table>
Correlations of VOT after removing effect of speaking rate:

P-T: .82, p < .001
T-K: .78, p < .001
K-P: .80, p < .001
B-D: .02, p = .8
D-G: .25, p < .01
G-B: .36, p < .001
P-B: -.10, p = .2
T-D: .43, p < .001
K-G: .26, p < .01
Correlations for vowel duration after removing effect of speaking rate:
P-T: .79, p < .001
T-K: .71, p < .001
K-P: .66, p < .001
B-D: .70, p < .001
D-G: .78, p < .001
G-B: .79, p < .001
P-B: .35, p < .001
T-D: .66, p < .001
K-G: .73, p < .001