Talker variation and systematicity in voice onset time: A corpus study

Eleanor Chodroff and Colin Wilson
Department of Cognitive Science, Johns Hopkins University

4pSC29

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^h] and [k^h] covary across talkers (r = 0.82)
- Also find cross-voice VOT correlations (e.g., [t^h] 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).

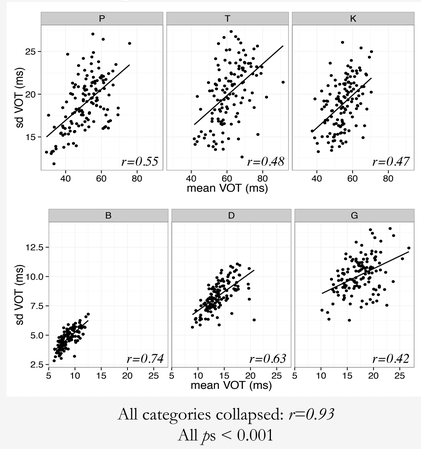
Results

Population Values

Population means and sds of talker-specific mean VOT (ms)

Stop	Mean	SD
P	51.1	9.4
T	61.3	9.1
K	54.8	7.2
B	8.7	1.5
D	13.9	2.4
G	17.4	3.0

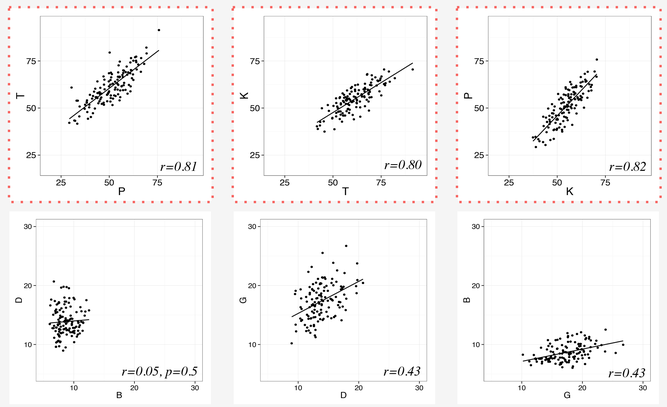
Speaker Means vs SDs



All categories collapsed: r=0.93
All ps < 0.001

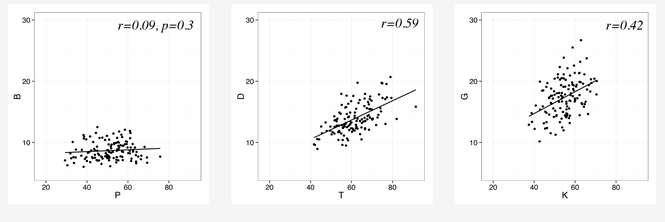
Cross-Place Correlations

All plots depict talker-specific mean VOTs in ms



Cross-Voice Correlations

All plots depict talker-specific mean VOTs in ms



All ps < 0.001, unless otherwise indicated

Modeling perceptual adaptation

Many approaches to speech perception assume that listeners employ talker-specific phonetic means (e.g., Nearey, 1978; Lobanov, 1971; McMurray & Jongman, 2011 for z-scoring or mean subtraction).

Investigate incremental Bayesian inference of talker-specific means with two models that differ in knowledge of structured variation:

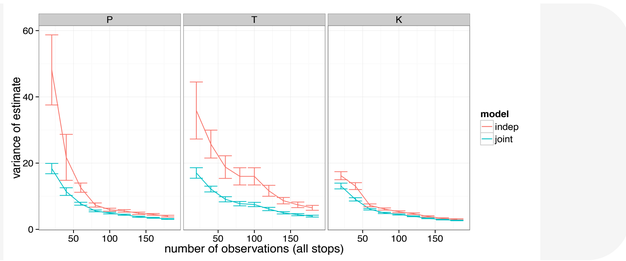
Model-Indep

(i = talker, j = stop)

- Estimate talker-specific mean μ_{ij} for each stop *independently*, with prior distribution on μ_{ij} specified by population values μ_{0j}, σ_{0j}

Model-Joint

- Estimate talker-specific vector of stop means μ_i *jointly,* using prior specified by population mean vector μ_0 and covariance matrix Σ_0



Talker-specific estimates are more precise — have lower variance — in the model that infers stop means jointly:

- Observations of different stops are mutually-informative.
- Increase in precision is significant for [p^h t^h k^h] (ps < 0.001).
- Models do not differ significantly in error of estimated means, or in predictions for voiced stops (which have small population sds).

Methods

Mixer-6 Corpus of read speech

- Randomly selected utterances from the Switchboard corpus
- Sentence length: 1-17 words (median: 7 words)
- Transcript read 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
  - Median: 27 years old

Number of Tokens

	P	T	K	B	D	G
Tokens (Per Speaker)	9285 (46-100)	5821 (18-78)	11492 (56-117)	12762 (72-133)	17459 (68-191)	11637 (59-118)

Total N > 68,400

Corpus Preparation

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

AutoVOT

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

- All stops aligned using the default AutoVOT acoustic models
- Minimum VOT duration
  - Voiceless stops: 15 ms
  - Voiced stops: 4 ms
- Window of analysis
  - Voiceless 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 = 12.9 ms)
- Values 2.5 sd away from the population grand mean were excluded from analysis

Discussion

- Knowledge of voiceless stop VOT correlations supported by evidence from perceptual adaptation and phonetic imitation
  - Listeners are able to identify that a long [k^h] is more characteristic of a talker with a long [p^h] even without hearing the talker produce the [k^h] category (Theodore *et al.*, 2010)
  - In imitation, listeners extrapolate a talker's characteristically long VOT of [p^h] to [k^h] 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 initially 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 (Gerstman, 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; Kleinschmidt & Jaeger, 2011, 2015; Pajak *et al.*, 2013)

Acknowledgments

We would like to thank Jack Godfrey, Sanjeev Khudanpur, Paul Smolensky, and Matt Goldrick for their useful input on acoustic analysis and overall project goals. We'd also like to thank Matthew Maciejewski, Wade Shen, Sharon Tam, Chloe Haviland, Elsheba Abraham, Alessandra Golden, Spandana Mandaloju, 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 DHS-USSS Forensic Services Division for supporting this research.