Structured Variability in Stop Consonant Realization:

A Corpus Study of Voice Onset Time in American English

Eleanor Chodroff[1], John Godfrey[2], Sanjeev Khudanpur[2], Colin Wilson[1]

Johns Hopkins University
[1]Department of Cognitive Science
[2]Center for Language and Speech Processing

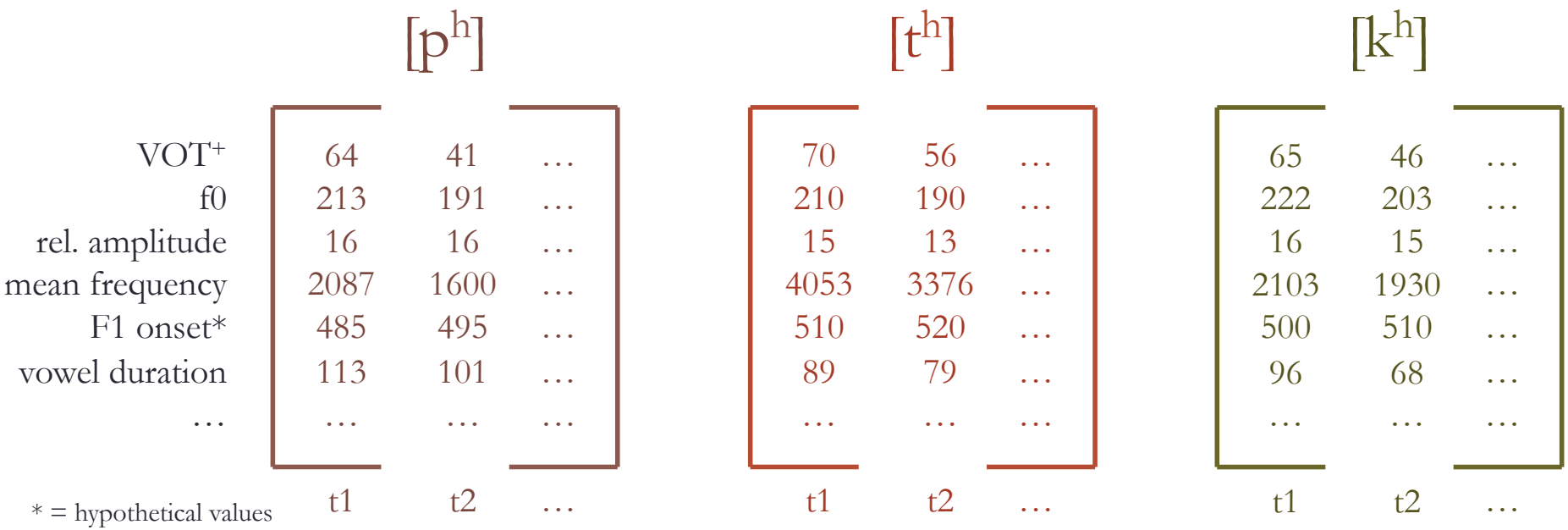
Individual talkers vary significantly in the phonetic realization of speech sounds

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

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



Many adaptation models posit that listeners estimate talker means (e.g., McMurray & Jongman, 2011), but independent estimation of many means would require considerable exposure.

Listeners generalize a talker's characteristic VOT across stop categories.
(Theodore et al., 2010; Nielsen, 2011)

Today's talk:

Evidence of structured variability in stop consonant VOT+ in the acoustic signal.

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

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

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

Acoustic measurement

Automatic pre-processing with Penn Forced Aligner and AutoVOT

PFA: Yuan & Liberman, 2008; AutoVOT: Keshet et al., 2014; Sonderegger & Keshet, 2010, 2012

Positive VOT (VOT$^+$): AutoVOT

Outlier exclusion

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)

Speaking rate: mean word duration in an utterance from PFA word boundaries

e.g. Summerfield, 1981; Miller et al., 1986; Miller & Volaitis, 1989; Pind, 1995; Kessinger & Blumstein, 1997, 1998; Allen et al., 2003

Stop Consonants for Analysis

68,297 word-initial prevocalic stop consonants
320 – 741 stop consonants per talker (median: 540)

Number of Tokens Per Talker

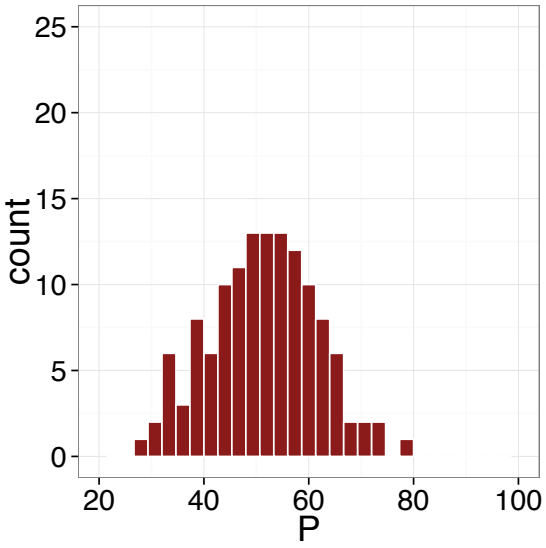
Stop	Range	Median	Total
P	46 – 98	72	9,287
T	17 – 77	45	5,834
K	55 – 114	91	11,491
B	70 – 138	98	12,671
D	70 – 192	140	17,432
G	59 – 122	91	11,582

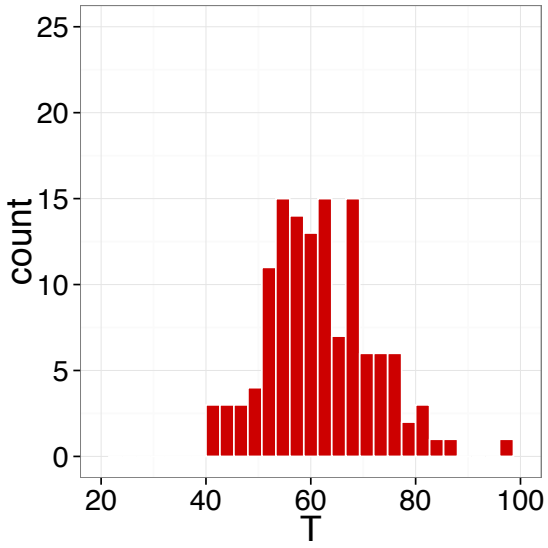
Word types

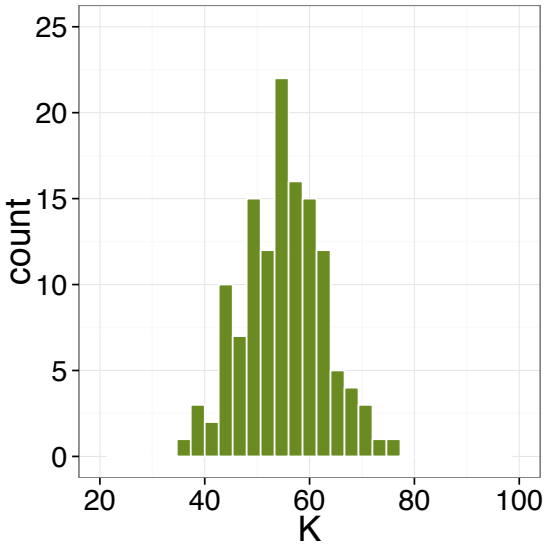
P : 17 T : 14 K : 22 B : 18 D : 16 G : 12

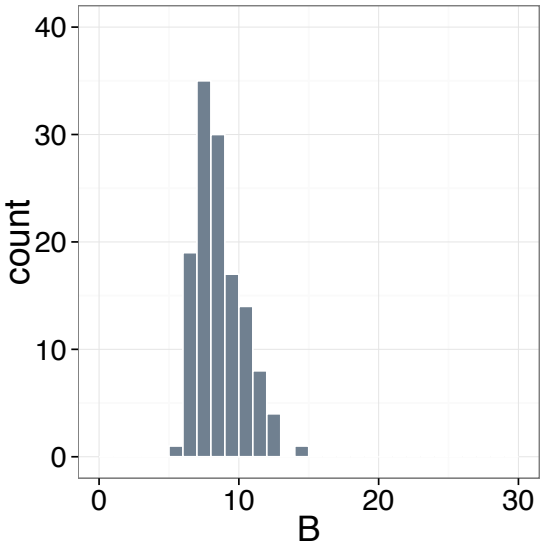
*Function words except "to" retained in the analysis

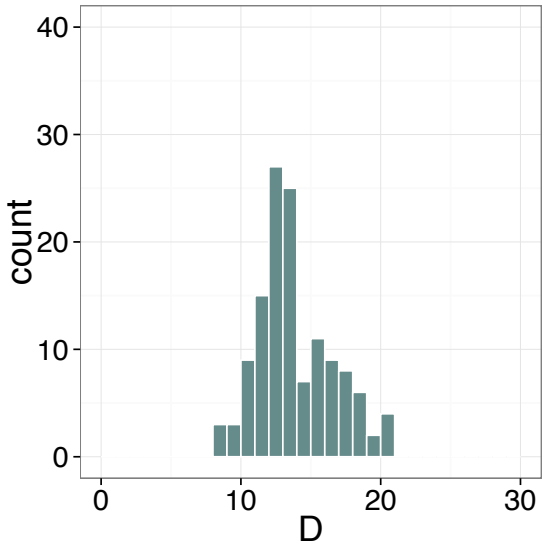
Extensive Variation in Talker Means

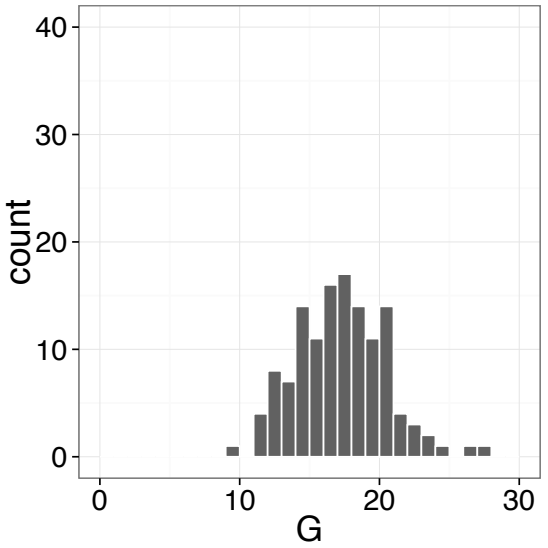




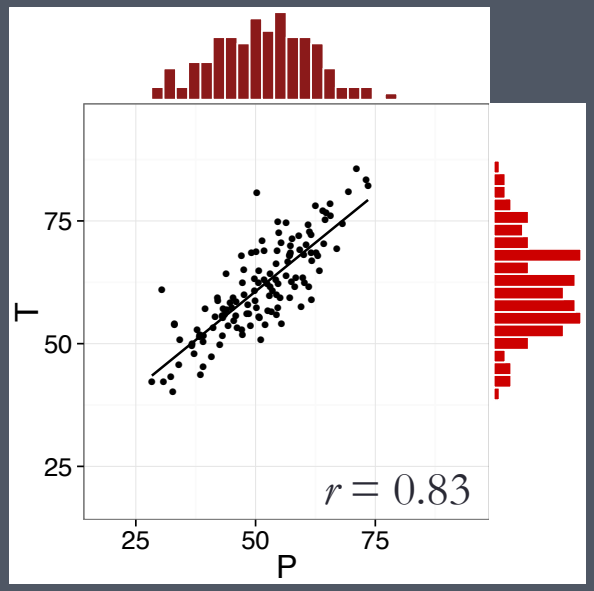






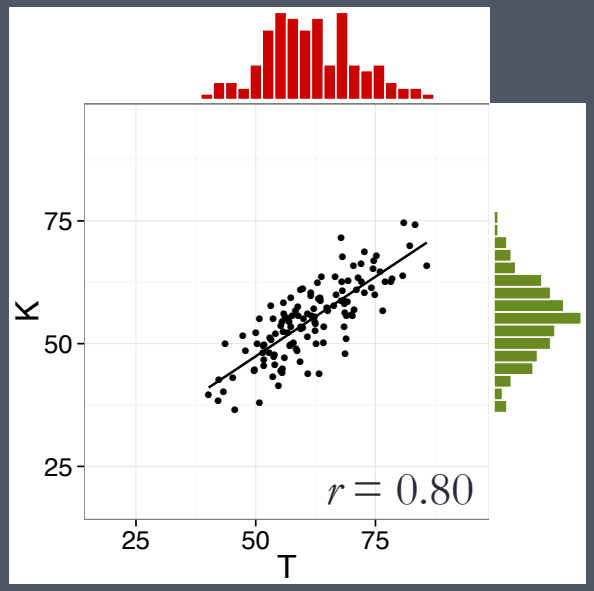


Cross-Place Correlations of Talker Means:
Voiceless (long-lag) Stops



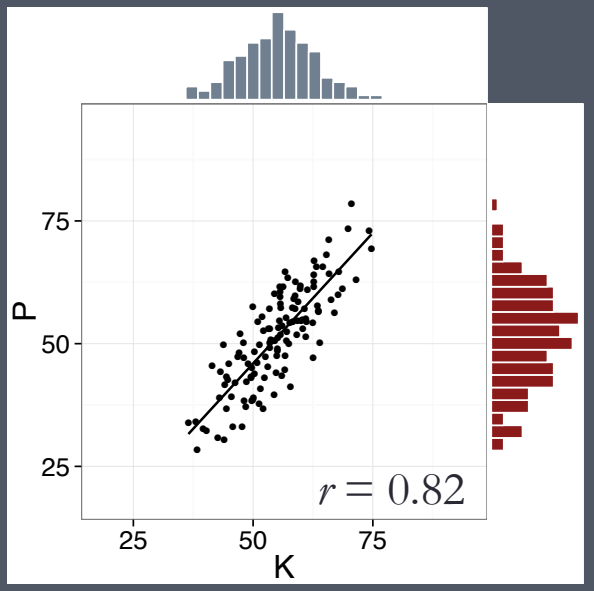
P – T

95% CI: [0.76, 0.88]



T – K

95% CI: [0.74, 0.85]



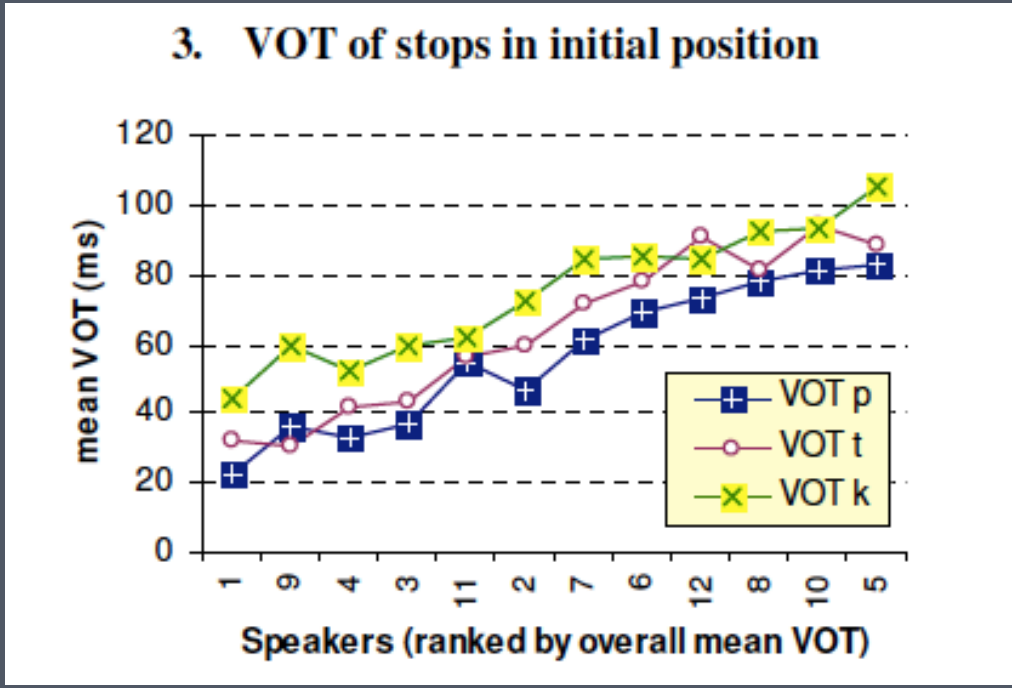
K – P

95% CI: [0.77, 0.87]

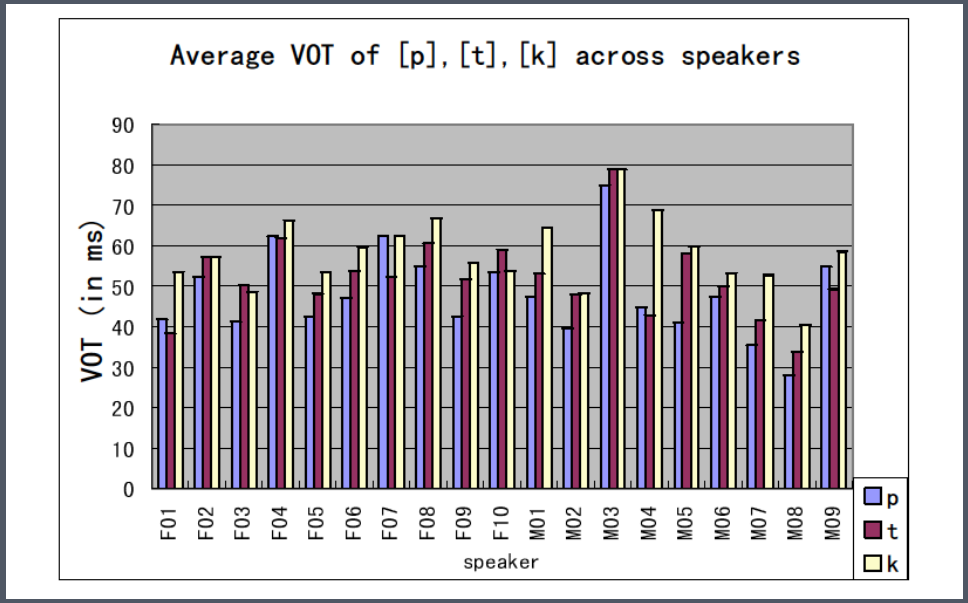
Each point = talker mean

In brackets: 95% CIs based on 1000 bootstrap replicates

All *p*s < 0.0003 (alpha-corrected) unless otherwise indicated

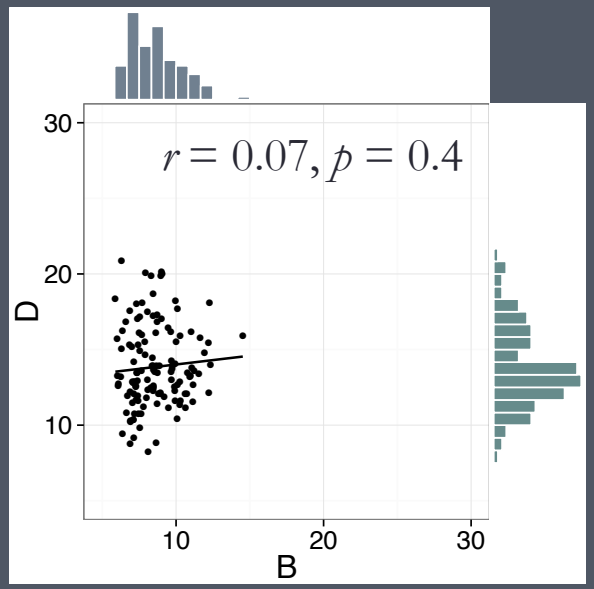


Scobbie, 2005



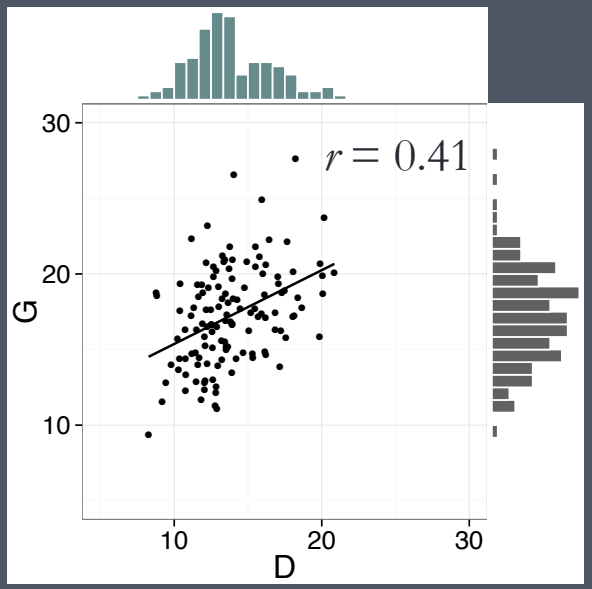
Yao, 2007

Cross-Place Correlations of Talker Means:
Voiced (short-lag) Stops



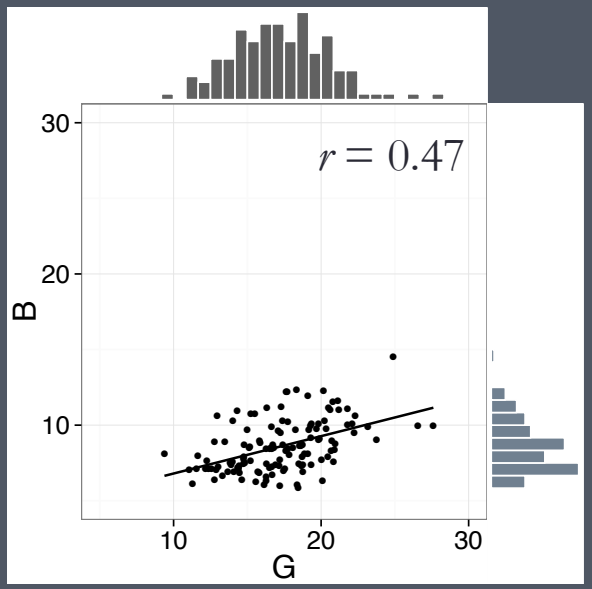
B – D

95% CI: [-0.10, 0.22]



D – G

95% CI: [0.25, 0.54]



G – B

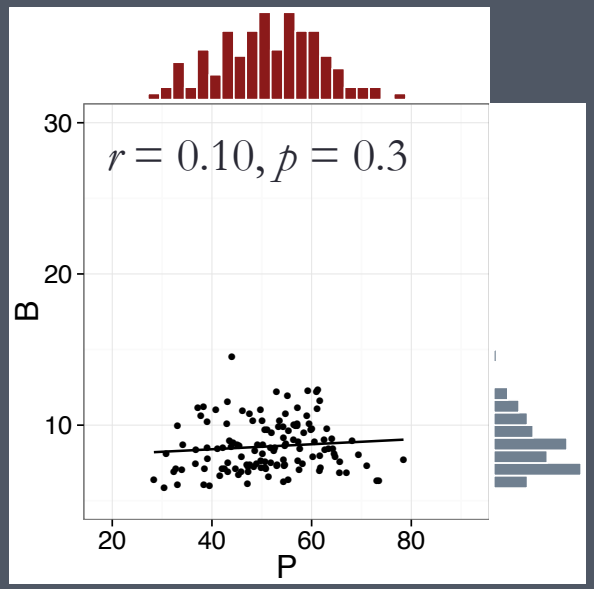
95% CI: [0.35, 0.59]

Each point = talker mean

In brackets: 95% CIs based on 1000 bootstrap replicates

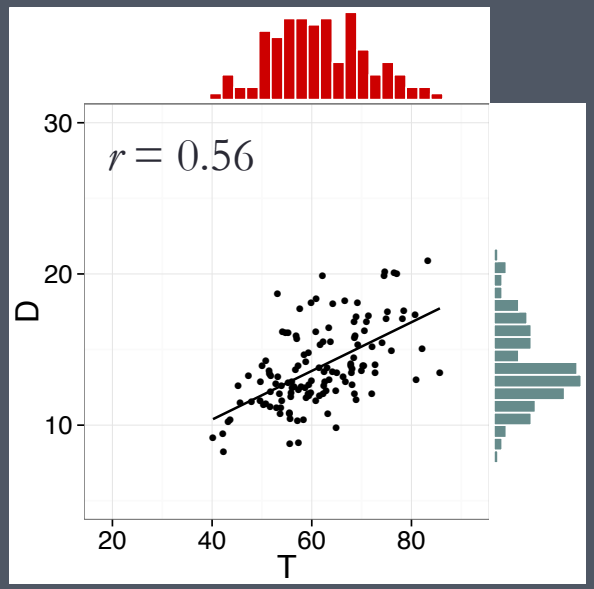
All *p*s < 0.0003 (alpha-corrected) unless otherwise indicated

Cross-Voice Correlations of Talker Means



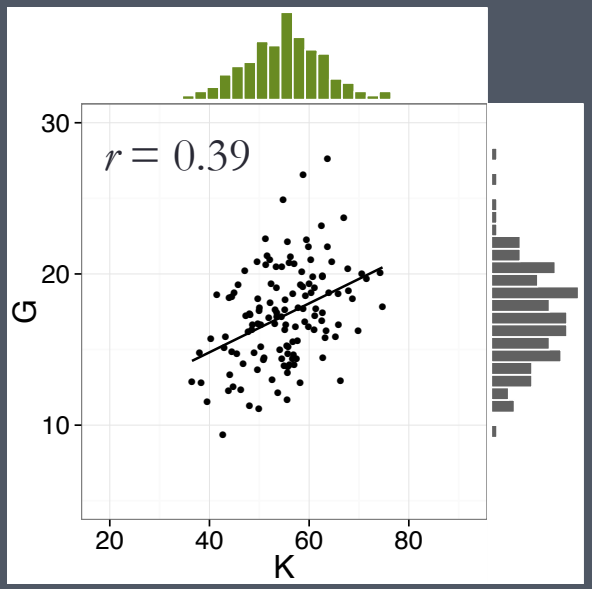
P – B

95% CI: [-0.10, 0.26]



T – D

95% CI: [0.42, 0.67]



K – G

95% CI: [0.24, 0.50]

Each point = talker mean

In brackets: 95% CIs based on 1000 bootstrap replicates

All *p*s < 0.0003 (alpha-corrected) unless otherwise indicated

linear mixed effects model predicting voice onset time

population model: vot ~ 1 + poa*voice + spk_rate + (1|word)
(β_0: 24.0 | β_{voice}: 21.4 | β_{poa1}: 1.2 | β_{poa2}: 3.8 | $\beta_{spkrate}$: 42.0)

place of articulation (sum-coded, labial baseline)
voice (sum-coded, voiceless = +1)
speaking rate in *seconds*

population model: vot ~ 1 + poa*voice + spk_rate + (1 | word)
(β_0: 24.0 | β_{voice}: 21.4 | β_{poa1}: 1.2 | β_{poa2}: 3.8 | $\beta_{spkrate}$: 42)

Random effect structure	AIC	BIC	LRT	*p* Value
population + 0	551,006	551,089		
population + (1 | talker)	546,666	546,757	4342.6	$p < 0.001$

voice (sum-coded, voiceless = +1)
place of articulation (sum-coded, labial baseline)

population model: vot ~ 1 + poa*voice + spk_rate + (1|word)
(β_0: 24.0 | β_{voice}: 21.4 | β_{poa1}: 1.2 | β_{poa2}: 3.8 | $\beta_{spkrate}$: 42)

Random effect structure	AIC	BIC	LRT	*p* Value
population + 0	551,006	551,089		
population + (1|talker)	546,666	546,757	4342.6	$p < 0.001$
population + (1 + voice|talker)	541,351	541,461	5318.9	$p < 0.001$

voice (sum-coded, voiceless = +1)
place of articulation (sum-coded, labial baseline)

population model: vot ~ 1 + poa*voice + spk_rate + (1 | word)
(β_0: 24.0 | β_{voice}: 21.4 | β_{poa1}: 1.2 | β_{poa2}: 3.8 | $\beta_{spkrate}$: 42)

Random effect structure	AIC	BIC	LRT	*p* Value
population + 0	551,006	551,089		
population + (1 | talker)	546,666	546,757	4342.6	$p < 0.001$
population + (1 + voice | talker)	541,351	541,461	5318.9	$p < 0.001$
population + (1 + poa*voice | talker)	540,575	540,749	789.57	$p < 0.001$

voice (sum-coded, voiceless = +1)
place of articulation (sum-coded, labial baseline)

Discussion

Talkers vary significantly in realization of stop consonant VOT across categories;
however, there are strong correlations of most cross-category means.
Talkers do vary but their stops covary *(to a significant degree).*

Listeners could exploit structured variation to extrapolate from limited talker-specific evidence and refine a talker-specific model with further exposure.

Joint *(rather than independent) estimation of many talker-specific phonetic properties.*

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

Current research suggests very large scale structure to acoustic variation across talkers in AE stops

Strong correlations on other dimensions across talkers

ex.: spectral center of gravity, f0, following vowel duration, relative amplitude

Cross-dimensional correlations

Future Directions

What underlies these correlations?

- physiological factors
- dialectal/sociophonetic
- phonology-phonetics interface
- preservation of VOT^{+} cue to place

(Peterson & Lehiste, 1960; Cho & Ladefoged 1999)

Examine effect of word and prosodic positions (domain-initial strengthening, lexical frequency, neighborhood properties)

Explore cross-talker patterns in other speech sounds

Investigate cognitive status of correlations with new talker adaptation experiments

Thanks to:

Matt Maciejewski, JHU CLSP
Jan Trmal, JHU CLSP
Wade Shen, MIT

Elsheba Abraham
Alessandra Golden
Chloe Haviland
Spandana Mandaloju
Ben Wang

Emily Atkinson, JHU
Matt Goldrick, Northwestern
NYU Phonetics & Experimental Phonology Lab

Supported by:

Department of Homeland Security –
USSS Forensic Services Division

Science of Learning Institute –
Johns Hopkins University

Thank you!

Correlations 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, $p < .01$

	P-T	P-K	T-K	B-D	B-G	D-G	P-B	T-D	K-G
vot	0.83*	0.82*	0.80*	0.07	0.47*	0.41*	0.10	0.56*	0.39*
cog	0.44*	0.57*	0.52*	0.55*	0.61*	0.68*	0.64*	0.72*	0.73*
f0	0.89*	0.92*	0.95*	0.98*	0.96*	0.95*	0.88*	0.95*	0.92*
amp	0.63*	0.69*	0.69*	0.49*	0.57*	0.61*	0.07	0.52*	0.32*
vdur	0.81*	0.83*	0.84*	0.86*	0.87*	0.88*	0.68*	0.78*	0.91*

	vot-cog	vot-f0	vot-amp	vot-vdur	cog-f0	cog-amp	cog-vdur	f0-amp	f0-vdur	amp-vdur
P	0.32*	0.19	-0.12	-0.07	0.26	-0.15	-0.05	-0.23	-0.06	0.32*
T	0.34*	0.27	-0.08	0.07	0.44*	-0.13	-0.01	-0.26	-0.04	0.54*
K	0.25	0.20	-0.04	0.15	0.35*	-0.13	-0.02	-0.24	0.07	0.34*
B	0.32*	-0.43*	0.18	0.10	0.24	0.66*	0.05	0.32	0.13	0.13
D	0.70*	0.30	0.49*	0.38*	0.45*	0.36*	0.09	0.07	0.00	0.45*
G	0.48*	-0.18	0.25	0.33*	0.31*	0.49	0.10	0.28	-0.02	0.35

Variation in VOT

vot ~ 1 + poa*voice + spk_rate +
(1 + poa*voice | talker) + (1 | word)

Fixed Effects	**Beta**	**t-value**
Intercept	29.3	37.2
coronal	1.6	2.1
dorsal	3.6	4.0
vcl	21.7	30.8
speaking rate (s)*	22.3	19.4
coronal x vcl	*1.15*	*1.3*
dorsal x vcl	*-1.15*	*-1.3*

voice (sum-coded, voiceless = +1)
place of articulation (sum-coded, labial baseline)

*For every 100ms increase in average word duration, VOT increases by about 2.2ms

Variation in VOT

Model 1 vot ~ 1 + poa*voice + spk_rate +
(1 + poa*voice + spk_rate|talker) + (1|word)

Fixed Effects	**Beta**	**t-value**
Intercept	29.4	36.4
coronal	*1.6*	*1.7*
dorsal	3.6	4.0
vcl	21.7	30.8
speaking rate (s)*	21.8	13.2
coronal x vcl	*1.16*	*1.3*
dorsal x vcl	*-1.15*	*-1.3*

voice (sum-coded, voiceless = +1)
place of articulation (sum-coded, labial baseline)

*For every 100ms increase in average word duration, VOT increases by about 2.2ms

Automatic pre-processing

Reading and recording errors removed via automatic and manual pre-processing

- SCLite: score for agreement btw. hypothesized and reference sentences
- Human listening for sentences with < 100% agreement

All wav files force-aligned to a "cleaned" transcript with the Penn Forced Aligner (PFA, Yuan & Liberman, 2008)

Stop consonant boundaries refined with AutoVOT (Sonderegger & Keshet, 2010)

Window of analysis

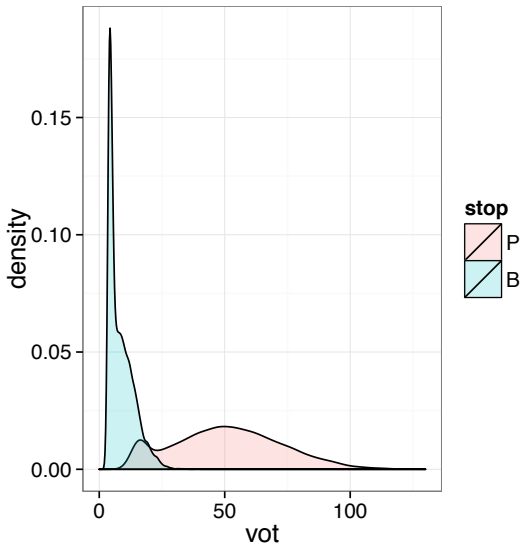
PFA interval + 30ms in both directions for voiceless stops

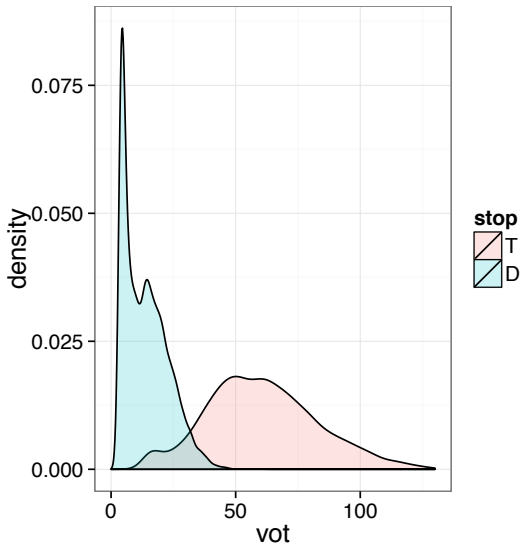
minimum VOT= 15ms

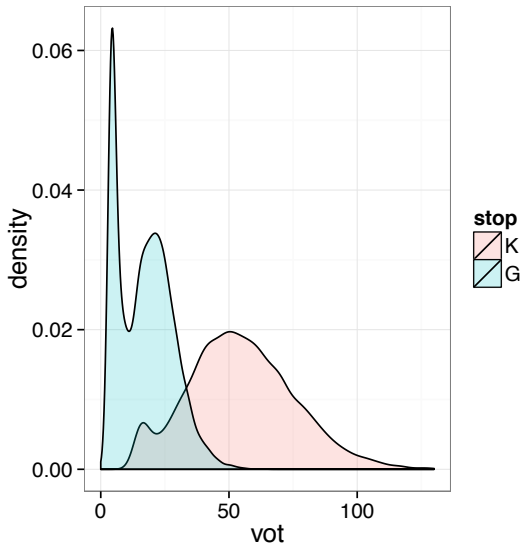
PFA interval + 10ms in both directions for voiced stops

minimum VOT = 4ms

Population VOT







B < D < G << P < K < T

Stop	Mean (ms)	SD (ms)	Mean (ms)	SD (ms)	Mean (ms)	Range (ms)
P	51	22	44	22	58	20:120
T	61	22	49	24	70	30:105
K	55	21	52	24	80	50:135
B	9	5	18	7	1	0:5
D	14	9	24	14	5	0:25
G	17	10	27	11	21	0:35
		Present study		Byrd (1993)	Lisker & Abramson (1964)	