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)

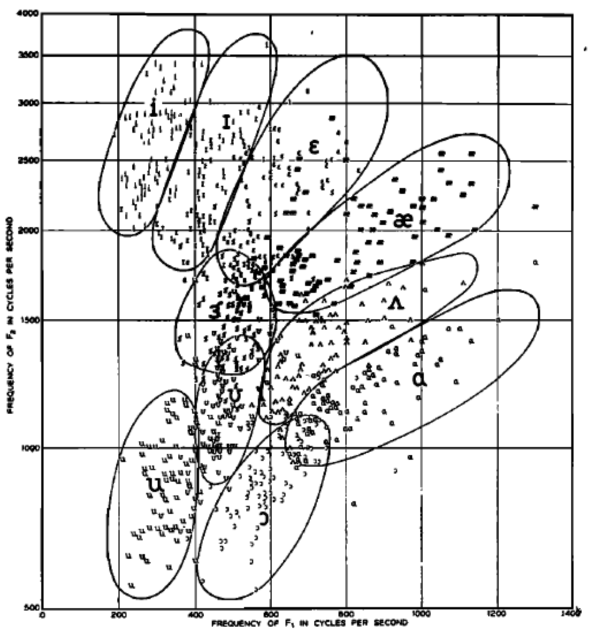
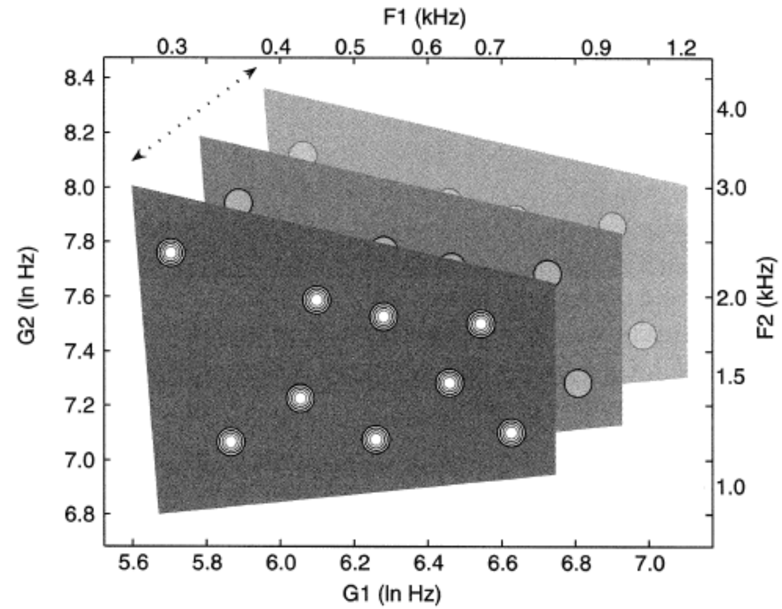


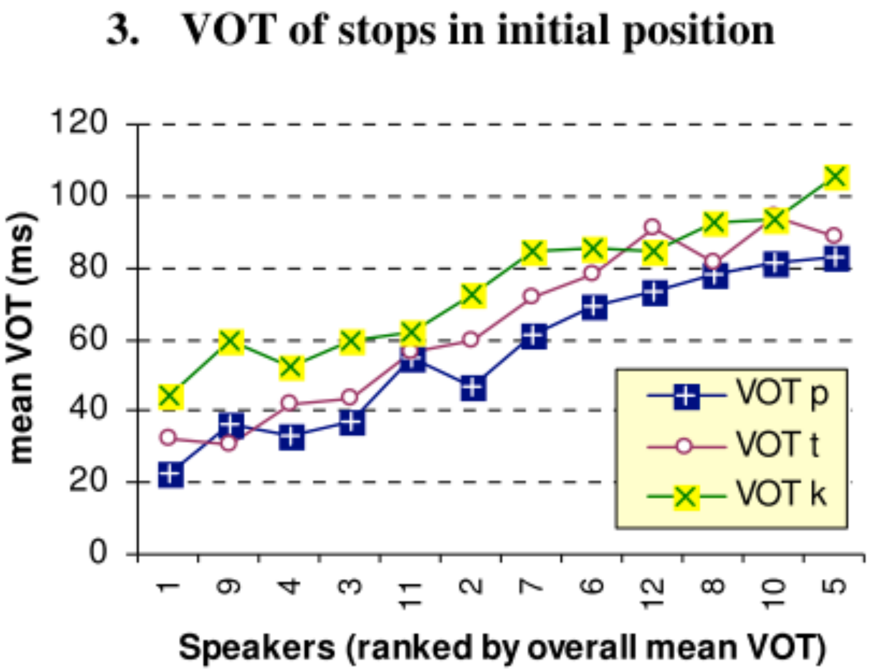
FIG. 8. Frequency of second formant *versus* frequency of first formant for ten vowels by 76 speakers.



Joos, 1948; Peterson & Barney, 1952;
Nearey, 1989; Nearey & Assmann, 2007

Evidence for structured variability

Covariance of talker means across stops



Scobbie, 2008; Theodore et al., 2009; Yao, 2009

Structured variability in stop consonants

	$[p^h]$ t1	$[p^h]$ t2	$[p^h]$ …	$[t^h]$ t1	$[t^h]$ t2	$[t^h]$ …	$[k^h]$ t1	$[k^h]$ t2	$[k^h]$ …
VOT$^+$	64	41	…	70	56	…	65	46	…
f0	213	191	…	210	190	…	222	203	…
rel. amplitude	16	16	…	15	13	…	16	15	…
mean frequency	2087	1600	…	4053	3376	…	2103	1930	…
F1 onset*	485	495	…	510	520	…	500	510	…
vowel duration	113	101	…	89	79	…	96	68	…
…	…	…	…	…	…	…	…	…	…

* = hypothetical values

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

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

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: ±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)

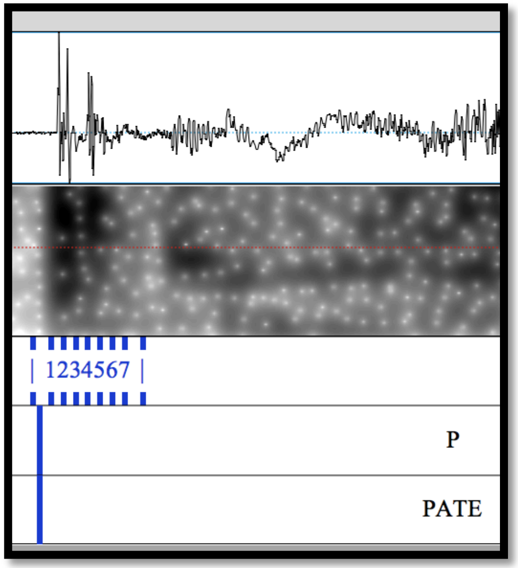
* Secondary cue to stop voicing

(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 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 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

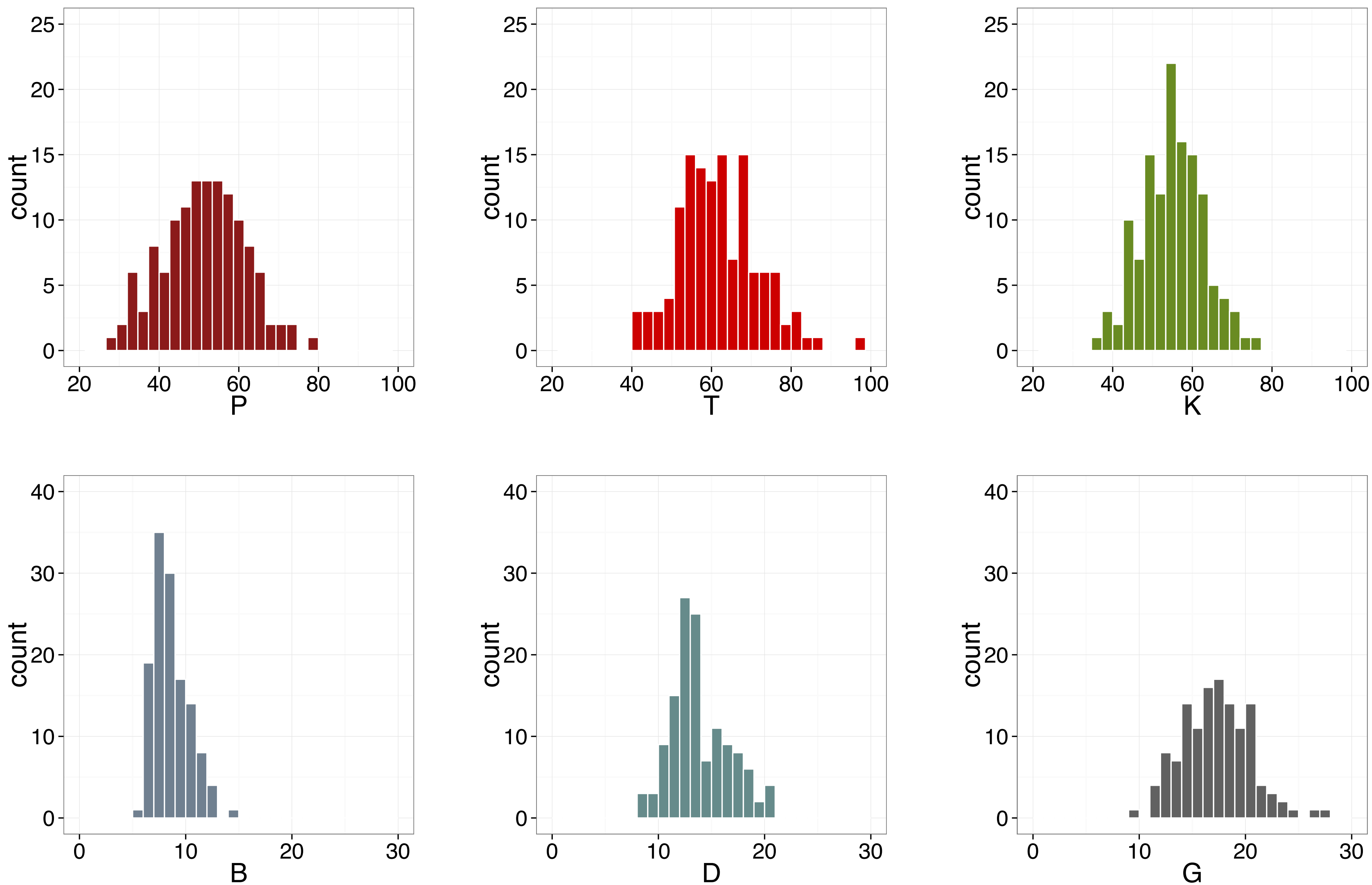
Stop	Range	Median	Total
P	47 – 98	77	9,686
T	17 – 77	46	5,906
K	55 – 114	93	11,765
B	70 – 138	99	12,681
D	70 – 192	140	17,441
G	59 – 122	91	11,591

Word types

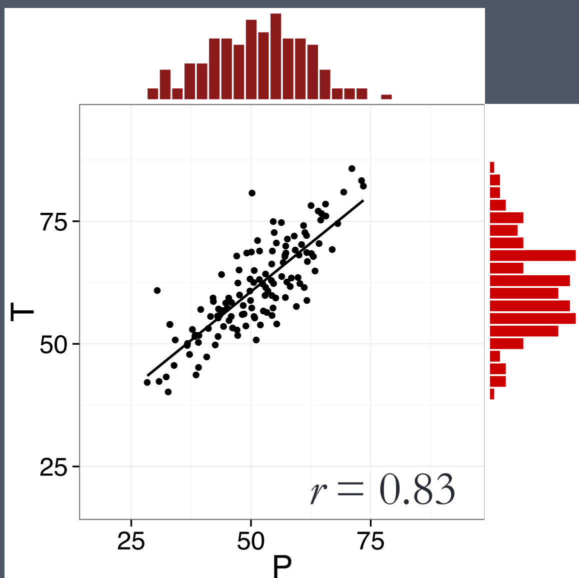
P : 17 T : 14 K : 22 B : 18 D : 16 G : 12

*Function words except "to" retained in the analysis

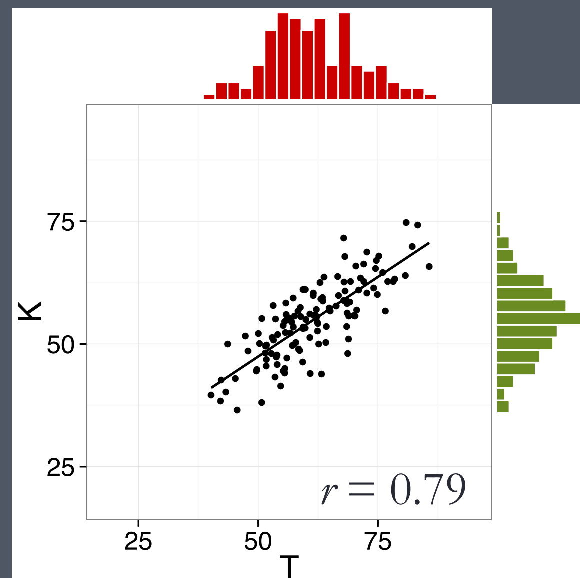
Variation in Talker Means for VOT+ (ms)



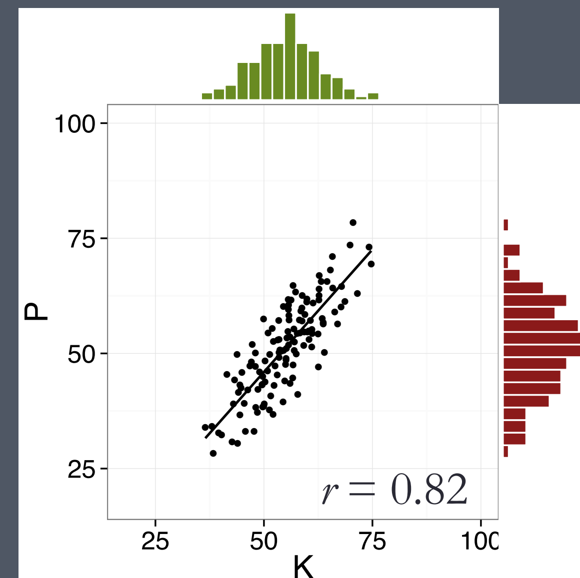
Cross-Place Correlations of Talker Means:
Voiceless (long-lag) Stops



P – T
95% CI: [0.75, 0.88]



T – K
95% CI: [0.72, 0.84]

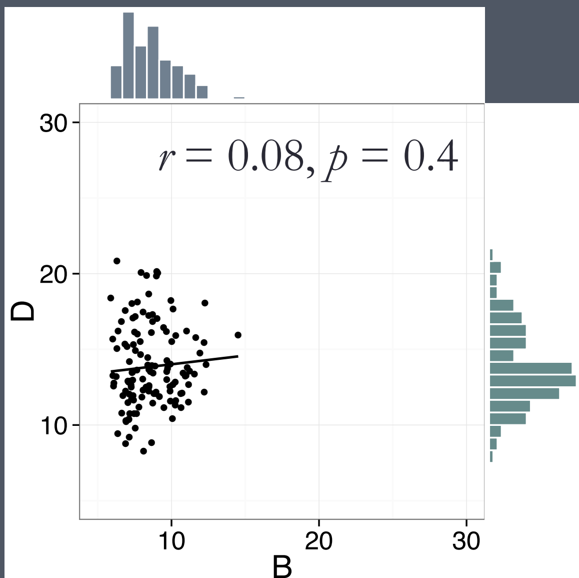


K – P
95% CI: [0.76, 0.87]

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

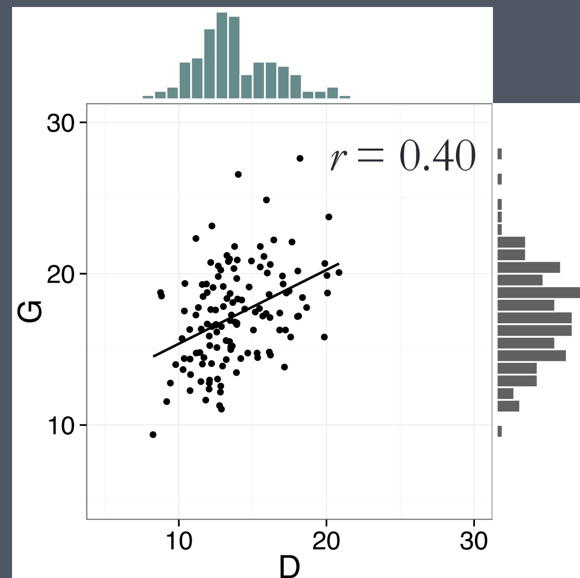
Cross-Place Correlations of Talker Means:

Voiced (short-lag) Stops



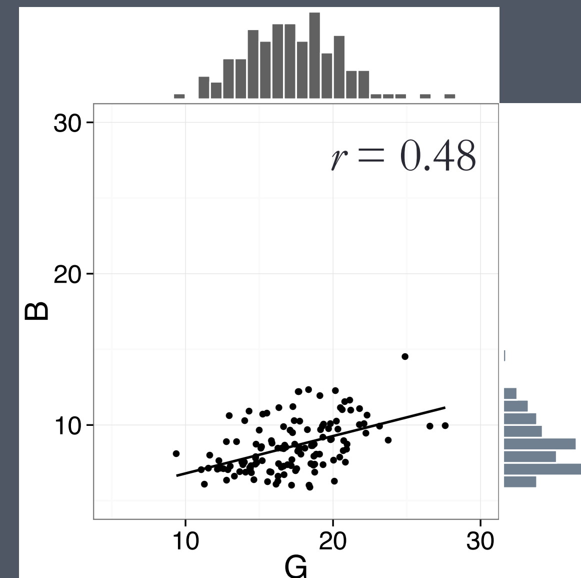
B – D

95% CI: [-0.08, 0.23]



D – G

95% CI: [0.26, 0.53]



G – B

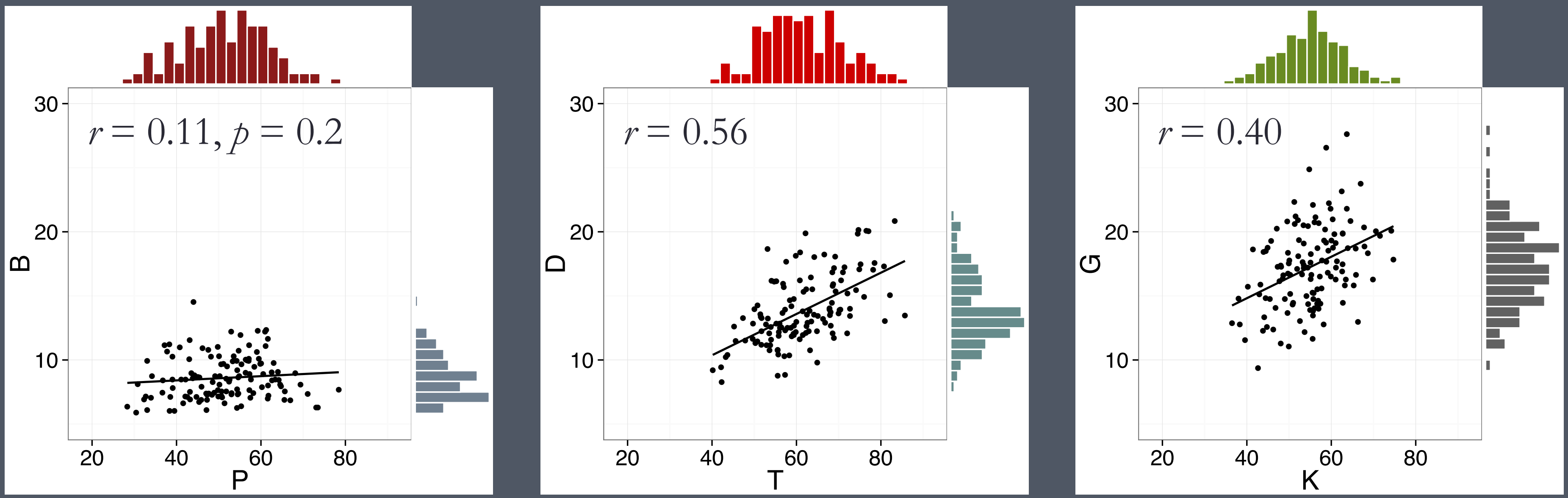
95% CI: [0.34, 0.59]

Each point = talker mean

In brackets: 95% CIs based on 1000 bootstrap replicates

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

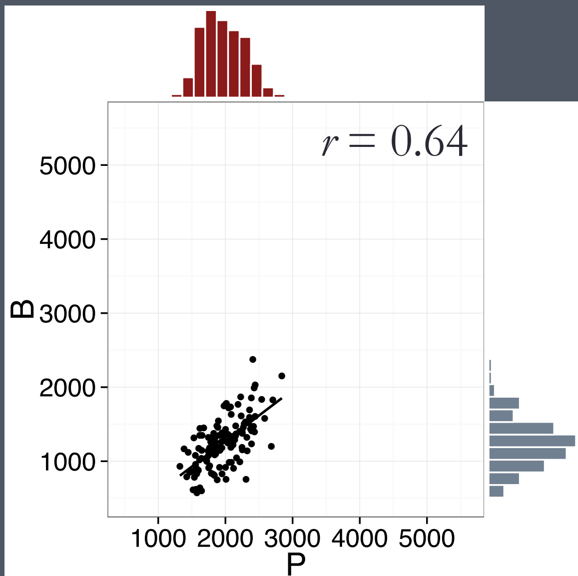
Cross-Voice Correlations of Talker Means:
Cross-Voice



P – B	T – D	K – G
95% CI: [-0.08, 0.28]	95% CI: [0.44, 0.68]	95% CI: [0.26, 0.53]

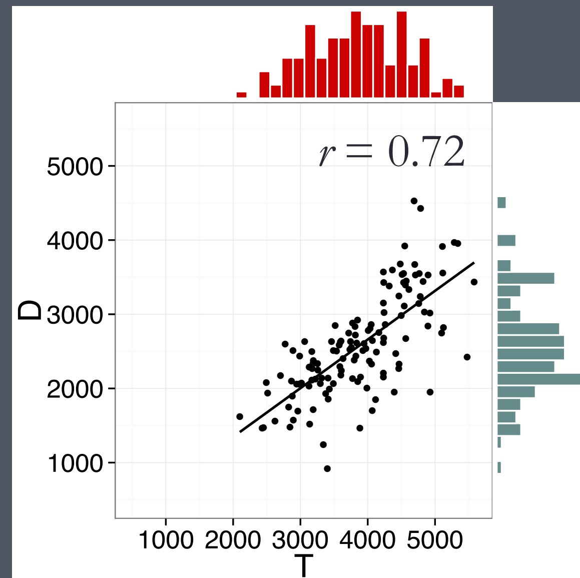
Each point = talker mean
In brackets: 95% CIs based on 1000 bootstrap replicates
All *ps* < 0.0003 (alpha-corrected) unless otherwise indicated

Spectral Center of Gravity (Hz)



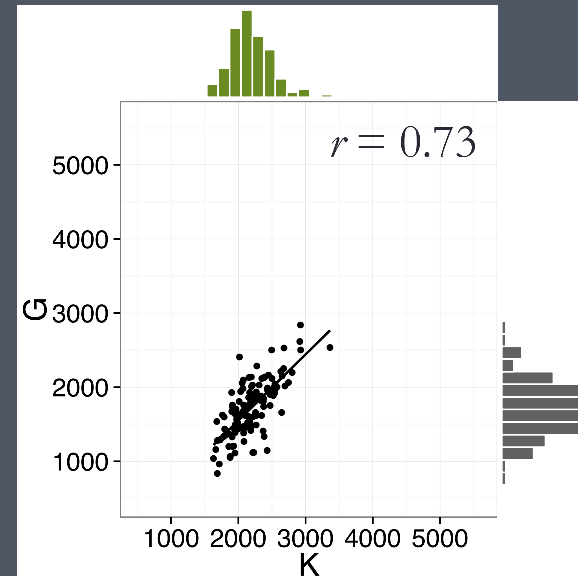
P – B

95% CI: [0.52, 0.73]



T – D

95% CI: [0.61, 0.78]



K – G

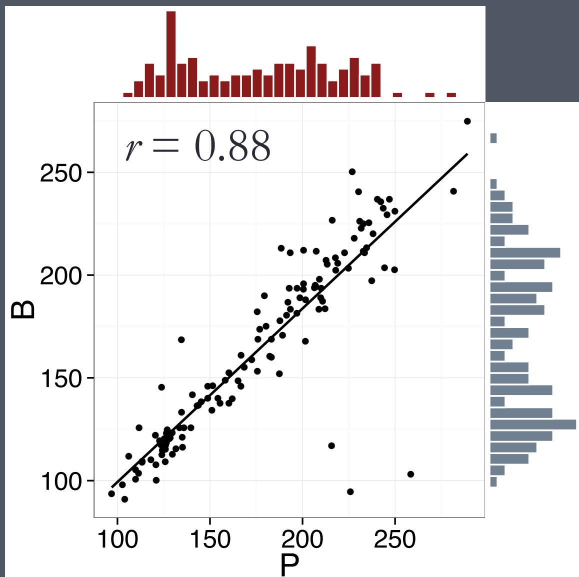
95% CI: [0.59, 0.80]

P-T	T-K	K-P	B-D	D-G	B-G
0.44	0.52	0.57	0.55	0.68	0.61
[0.29, 0.56]	[0.38, 0.63]	[0.44, 0.66]	[0.33, 0.68]	[0.58, 0.77]	[0.48, 0.72]

Each point = talker mean

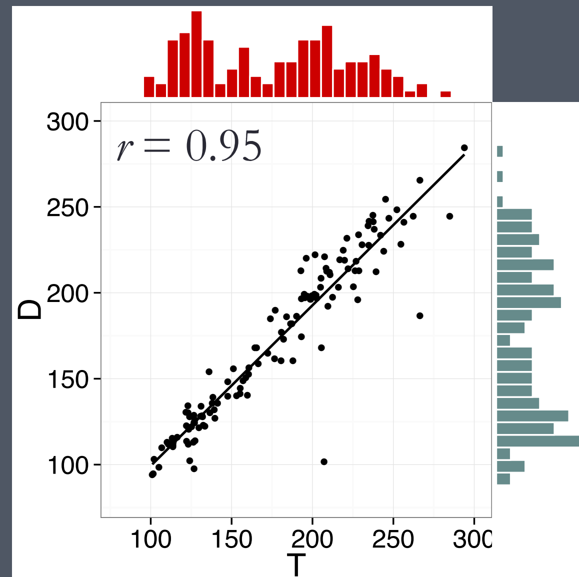
All *ps* < 0.0003 (alpha-corrected) unless otherwise indicated

f0 (Hz)



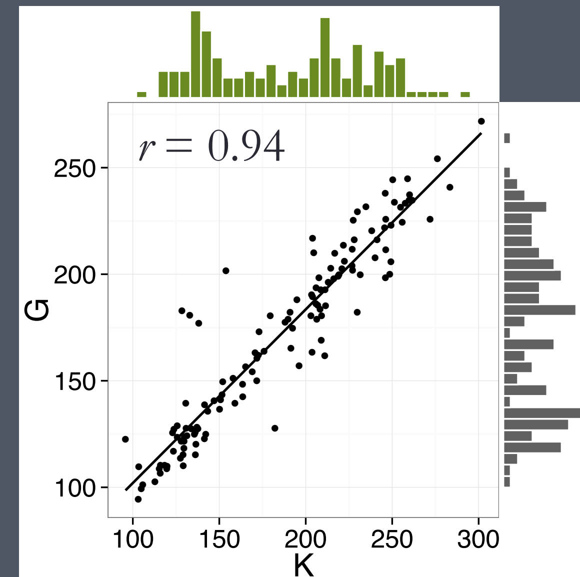
P – B

95% CI: [0.71, 0.95]



T – D

95% CI: [1]



K – G

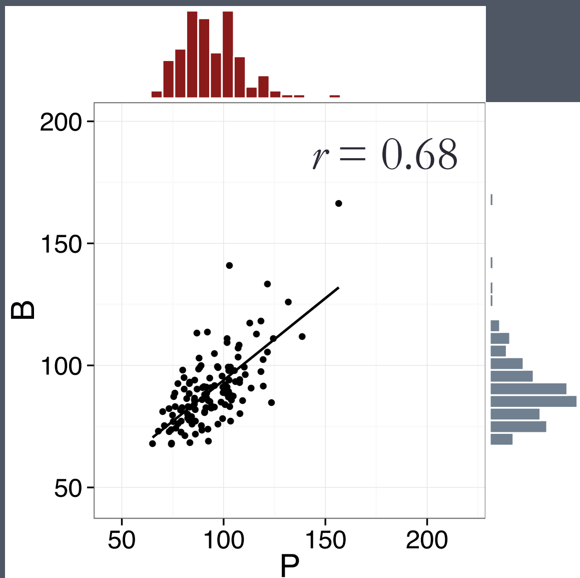
95% CI: [0.89, 0.96]

P-T	T-K	K-P	B-D	D-G	B-G
0.89	0.95	0.92	0.98	0.95	0.96
[1]	[1]	[0.80, 0.96]	[0.96, 0.98]	[0.91, 0.97]	[0.92, 0.97]

Each point = talker mean

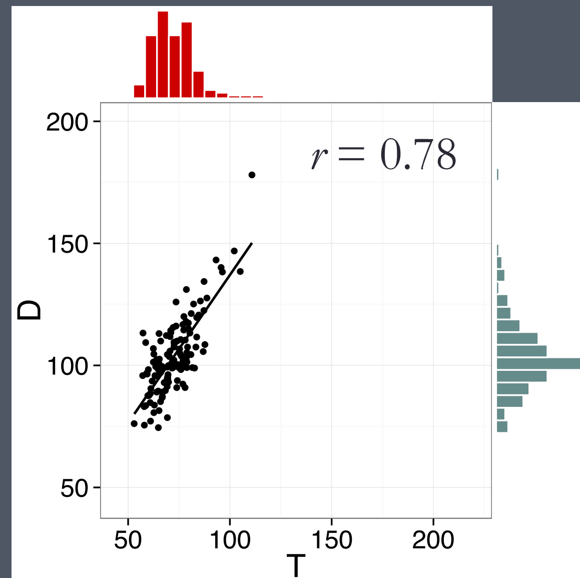
All *p*s < 0.0003 (alpha-corrected) unless otherwise indicated

Vowel Duration (ms)



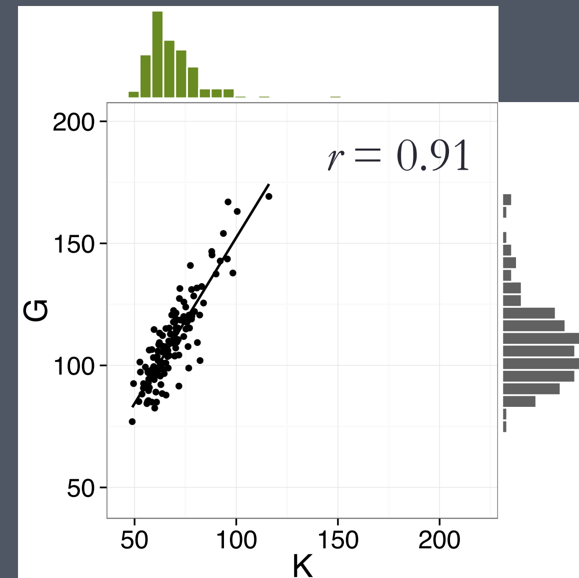
P – B

95% CI: [0.56, 0.81]



T – D

95% CI: [0.67, 0.86]



K – G

95% CI: [0.85, 0.95]

P-T	T-K	K-P	B-D	D-G	B-G
0.81	0.83	0.84	0.86	0.88	0.87
[0.72, 0.88]	[0.78, 0.88]	[0.76, 0.88]	[0.79, 0.92]	[0.83, 0.94]	[0.79, 0.93]

Each point = talker mean

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

Outline

Correlations Within-Category

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

(Nearey, 1989; Nearey and Assmann, 2007;
Solé & Ohala, 2010; Beddor et al., 2013; Clayards, submitted)

Correlations Within-Category

Correlations between talker-specific means within a stop category

	VOT x COG	**VOT x f0**		**VOT x vdur**	**COG x f0**		**COG x vdur**	**f0 x vdur**	
p	**0.32***	-0.02	-0.19	-0.07	-0.01	0.14	-0.05	0.13	0.13
t	**0.34***	0.04	-0.09	0.08	0.07	0.17	0.00	0.20	0.06
k	0.25	0.18	-0.17	0.15	0.07	0.05	-0.02	0.19	0.07
b	**0.33***	-0.21	-0.14	0.10	-0.12	-0.08	0.05	0.16	-0.16
d	**0.70***	-0.11	-0.05	**0.38***	-0.07	-0.16	0.09	0.08	-0.12
g	**0.50***	0.01	-0.25	**0.33***	0.06	-0.23	0.10	0.08	-0.15
		F	M		F	M		F	M

* $p < 0.001$

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:

$$p(\boldsymbol{m}) = \mathcal{N}(\boldsymbol{m}; \mu_{pop}, \Sigma_{pop})$$

Complete Covariance Model

Independence Model

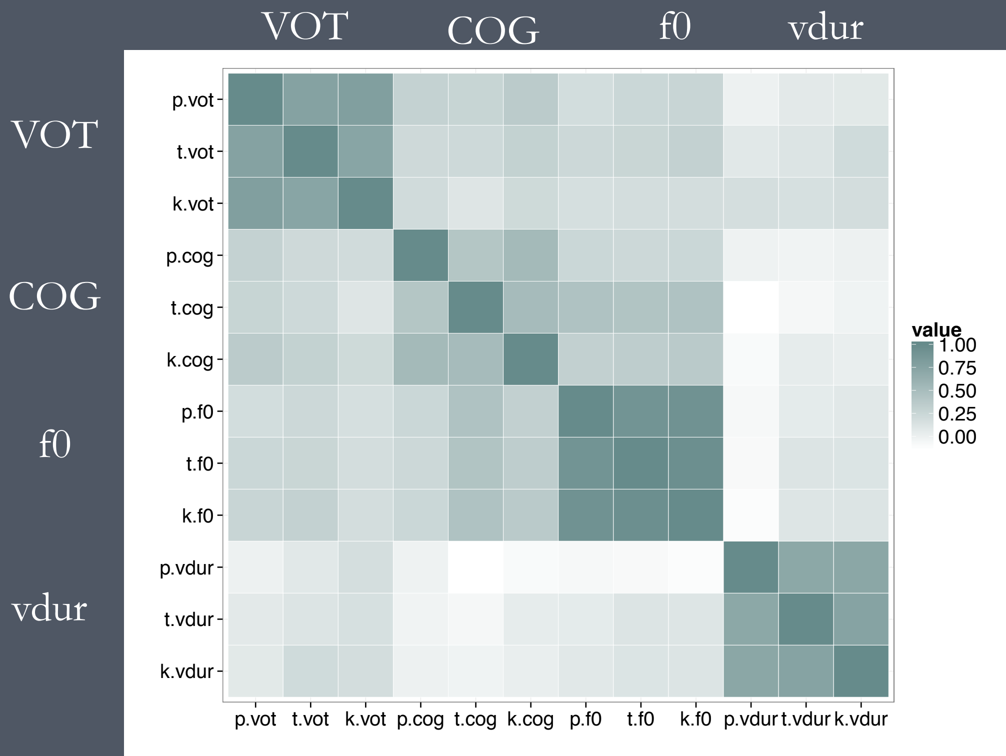
Adaptation to a novel talker:
Estimate posterior probability over talker means for each cue and stop

$$q_0(\boldsymbol{m}) = p(\boldsymbol{m}|\mu, \Sigma)$$

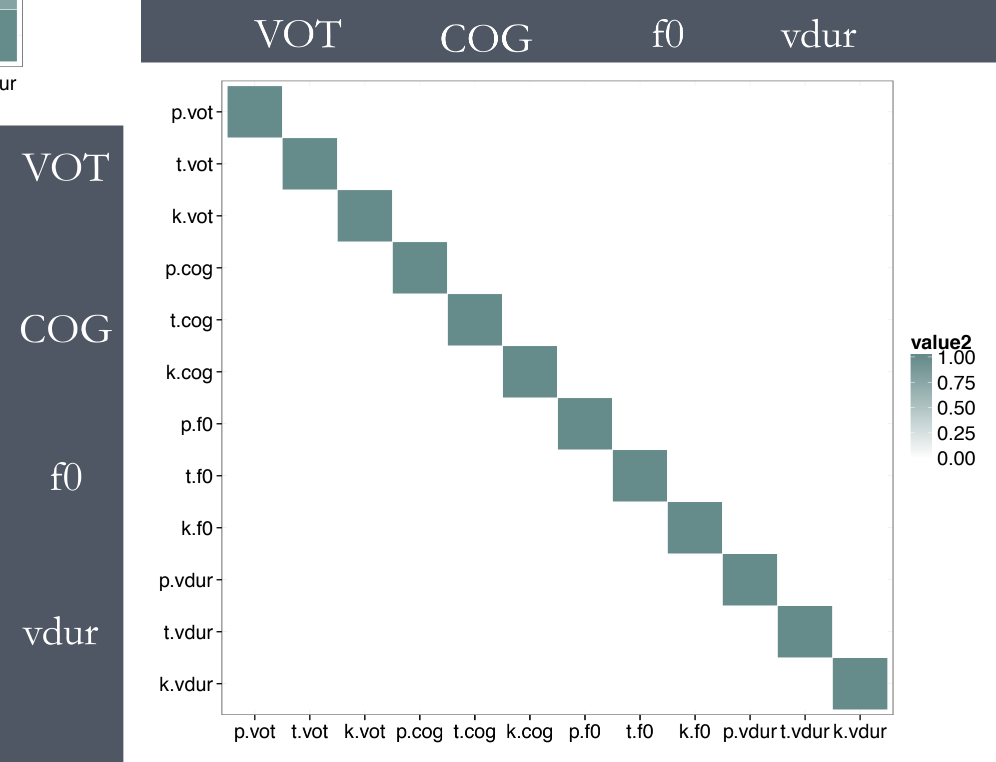
$$q_i(\boldsymbol{m}) \propto p(x_i|\boldsymbol{m}, l_i, \Sigma_{within.talker}) * q_{i-1}(\boldsymbol{m})$$

m = vector of talker-specific means (one entry per stop-cue combo)
$\boldsymbol{\mu}_{pop}$ = mean of **m** across the population,
$\boldsymbol{\Sigma}_{pop}$ = variance/covariance matrix on **m** in the population
$(\mathbf{x}_i, \mathbf{l}_i)$ = one stop production from the talker (acoustic cues, label)

Complete Covariance Model

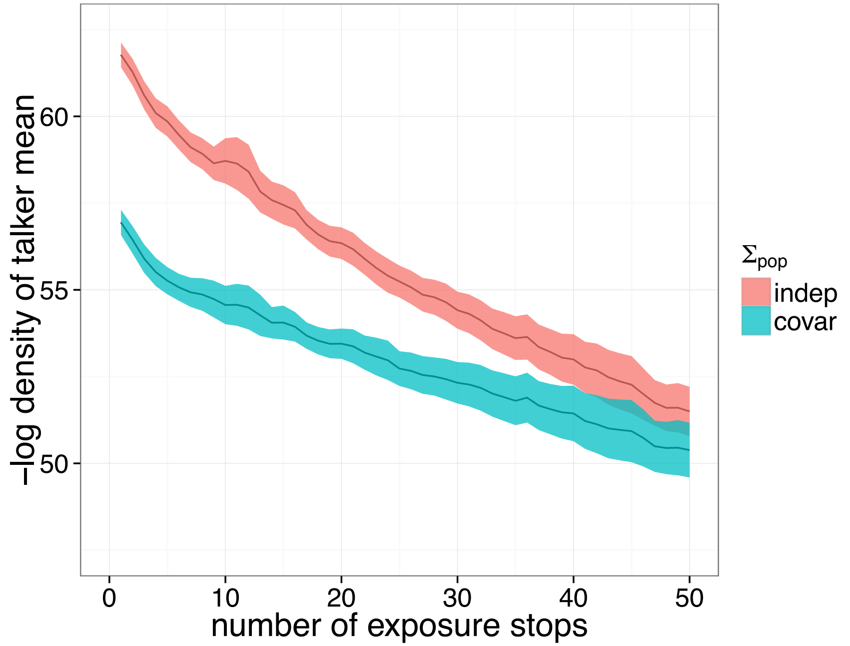


Complete Independence Model



Covariance vs Independence Models

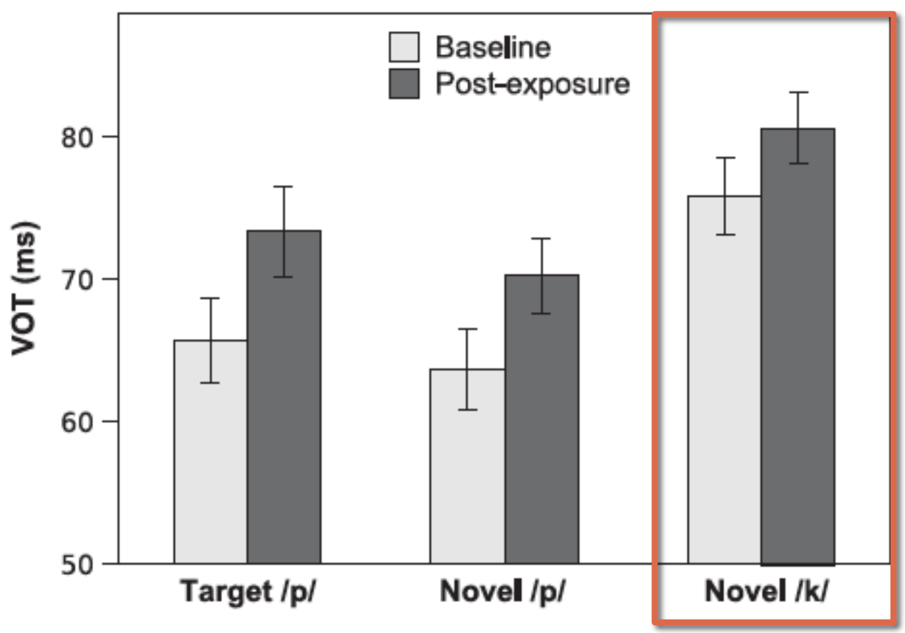
# of exposures	avg. density ratio	β	t
10	64.07	-2.08	-8.71
20	18.17	-1.45	-9.58
30	8.17	-1.05	-9.85
40	4.76	-0.78	-7.36
50	3.06	-0.56	-6.15



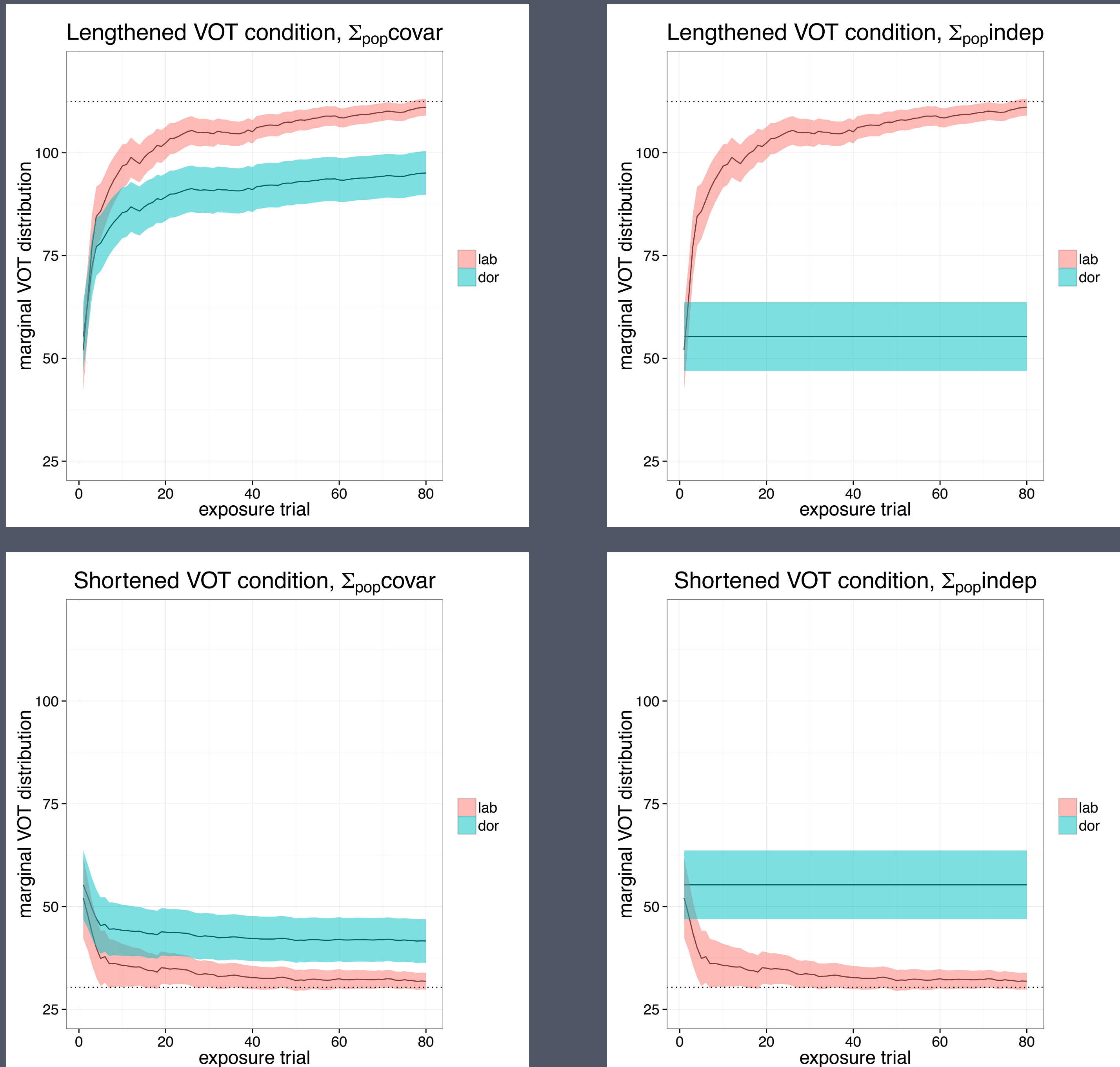
Bayesian Model of Talker Adaptation: Application

Perceptual Generalization across Phonetic Categories

Listeners generalize a talker's characteristic VOT across stop categories.
(Eimas & Corbit, 1973; Theodore & Miller, 2010; Nielsen, 2011)



Nielsen, 2011
Phonetic Imitation



Lengthened VOT condition, Σ_{pop}covar
Lengthened VOT condition, Σ_{pop}indep
Shortened VOT condition, Σ_{pop}covar
Shortened VOT condition, Σ_{pop}indep
marginal VOT distribution
exposure trial
25
50
75
100
0
20
40
60
80
lab
dor

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
169^{th} Acoustical Society of America
18^{th} 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

	B	D	G
VOT vs. COG	**-0.18 – 0.70** **mean: 0.30***	**-0.14 – 0.73** **mean: 0.46***	**-0.11 – 0.81** **mean: 0.49***
VOT vs. f0	-0.33 – 0.37 mean: -0.01	-0.38 – 0.29 mean: -0.08*	-0.47 – 0.31 mean = -0.04
VOT vs. vdur	-0.32 – 0.23 mean: -0.06*	-0.27 – 0.35 mean: 0.01	-0.20 – 0.34 mean: 0.10*
COG vs. f0	-0.40 – 0.45 mean: -0.03	-0.52 – 0.45 mean: -0.07*	-0.42 – 0.41 mean: -0.01
COG vs. vdur	-0.26 – 0.41 mean: 0.03	-0.34 – 0.31 mean: 0.00	-0.40 – 0.32 mean: 0.04
f0 vs. vdur	**-0.44 – 0.24** **mean: -0.10***	**-0.53 – 0.25** **mean: -0.19***	**-0.58 – 0.20** **mean: -0.20***

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, $p < .001$