An empirical and computational study of generalized adaptation to natural talker-specific VOT

Eleanor Chodroff, Alessandra Golden, and Colin Wilson

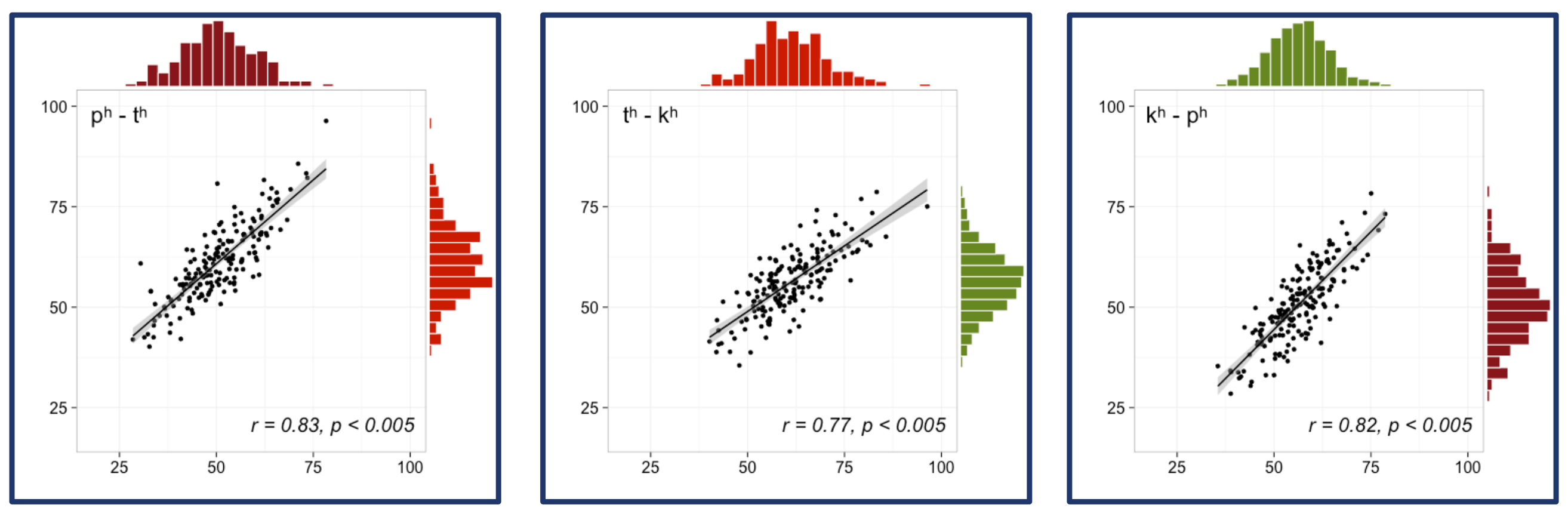
Department of Cognitive Science, Johns Hopkins University

Introduction

Substantial variability exists in the phonetic realization of speech sounds across talkers, yet listeners adapt rapidly and with ease. One source of information that could be exploited in talker adaptation is knowledge of **acoustic-phonetic covariation across phonetic categories**.

Evidence for acoustic-phonetic covariation comes from previously observed relationships among:

- F1xF2 vowel plane (e.g., Joos 1948; Nearey & Assmann 2007)
- **voice onset time (VOT**) of American English stop consonants, esp. word-initial voiceless aspirated stops (e.g., Chodroff et al. 2015)



Previous studies of **perceptual generalization** and **phonetic imitation** (e.g., Theodore & Miller 2010; Nielsen 2011) provide evidence that knowledge of VOT correlations may play a role in talker adaptation.

Limitations of previous VOT adaptation experiments:

- Exaggerated VOT manipulation (e.g., short: 88 ms vs. long: 183 ms)
- Extensive exposure to the novel talker (e.g., 120 exposures before testing)
- Limited stimulus variability (e.g., just two VOT values per stop and talker)

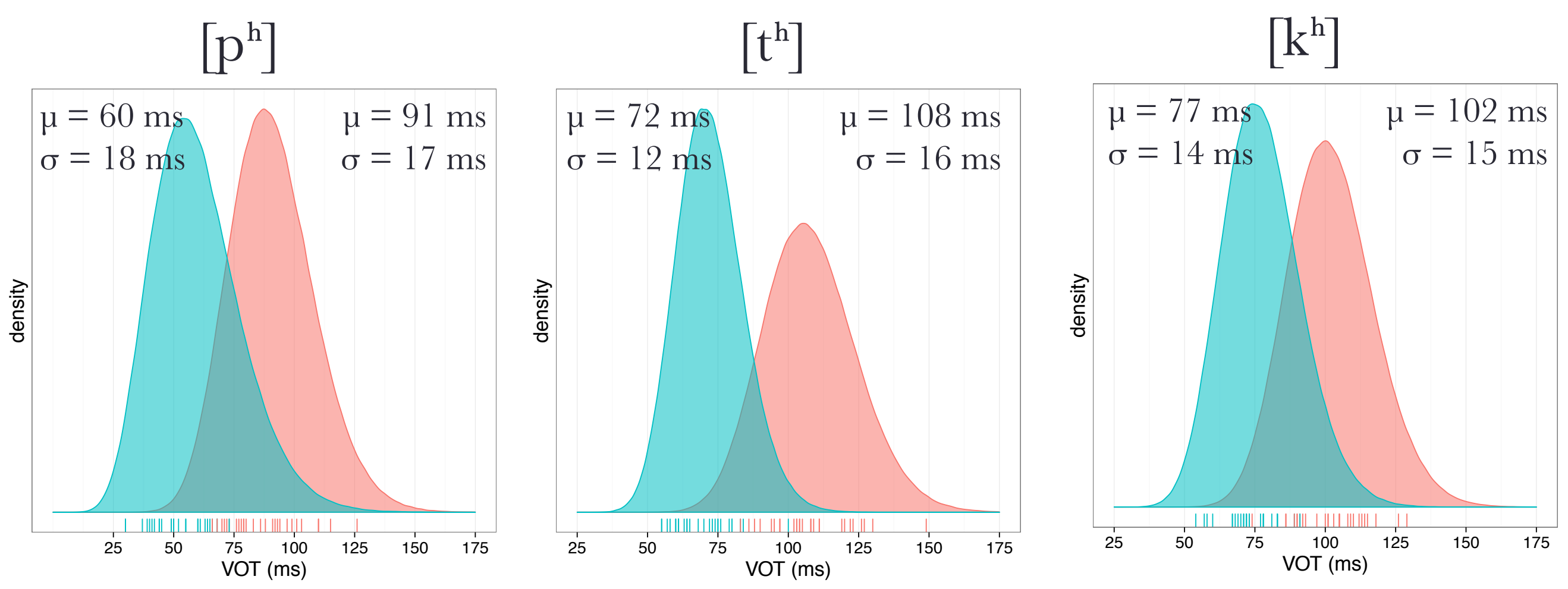
Objectives:

- Employ more natural and variable stimuli to investigate perceptual adaptation and generalization to an unheard place of articulation
- Examine effects of adaptation after minimal exposure
- Examine effectiveness of VOT covariation in accounting for adaptation

Methods

Stimuli

- Stimuli created from careful-speech CVC syllables (Chodroff & Wilson 2014): [pʰ tʰ kʰ] × [i eɪ ɛ æ ʌ ɑ ɔ oʊ u] ×[t]
- Gamma distributions fit to stop categories from two male talkers: one with naturally long VOTs and one with naturally short VOTs
- Manipulated VOT of a single male talker to match randomly generated values (3 stops × 9 vowels × 3 repetitions × 2 VOT lengths = 162 stimuli)



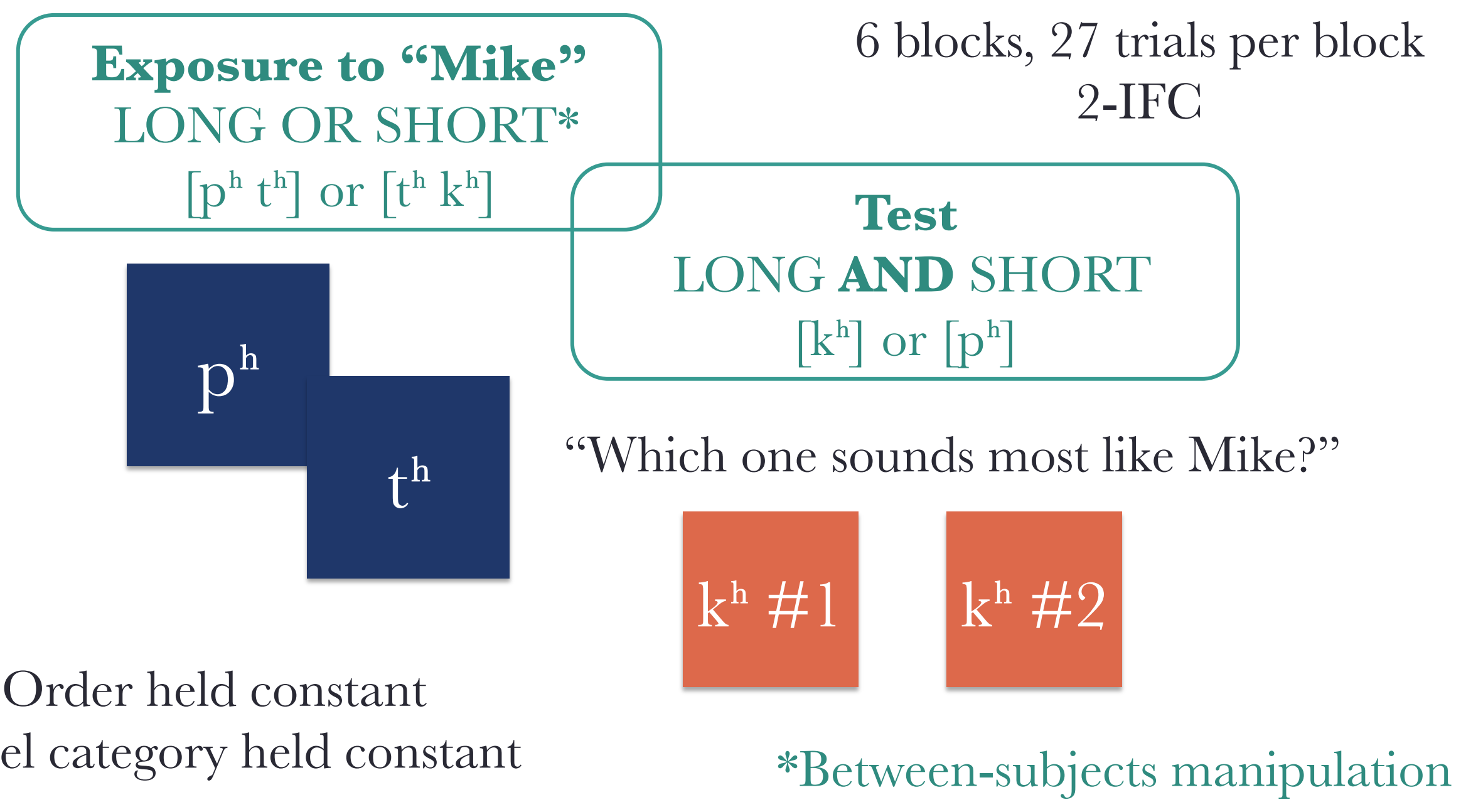
long p^h: $\alpha = 26.96$, $\beta = 0.297$
short p^h: $\alpha = 10.66$, $\beta = 0.179$

long t^h: $\alpha = 43.08$, $\beta = 0.400$
short t^h: $\alpha = 37.87$, $\beta = 0.525$

long k^h: $\alpha = 44.86$, $\beta = 0.439$
short k^h: $\alpha = 31.28$, $\beta = 0.404$

Procedure

Trial structure



Exposure to "Mike"
LONG OR SHORT*
[pʰ tʰ] or [tʰ kʰ]

6 blocks, 27 trials per block
2-IFC

Test
LONG **AND** SHORT
[kʰ] or [pʰ]

"Which one sounds most like Mike?"

Order held constant
Vowel category held constant

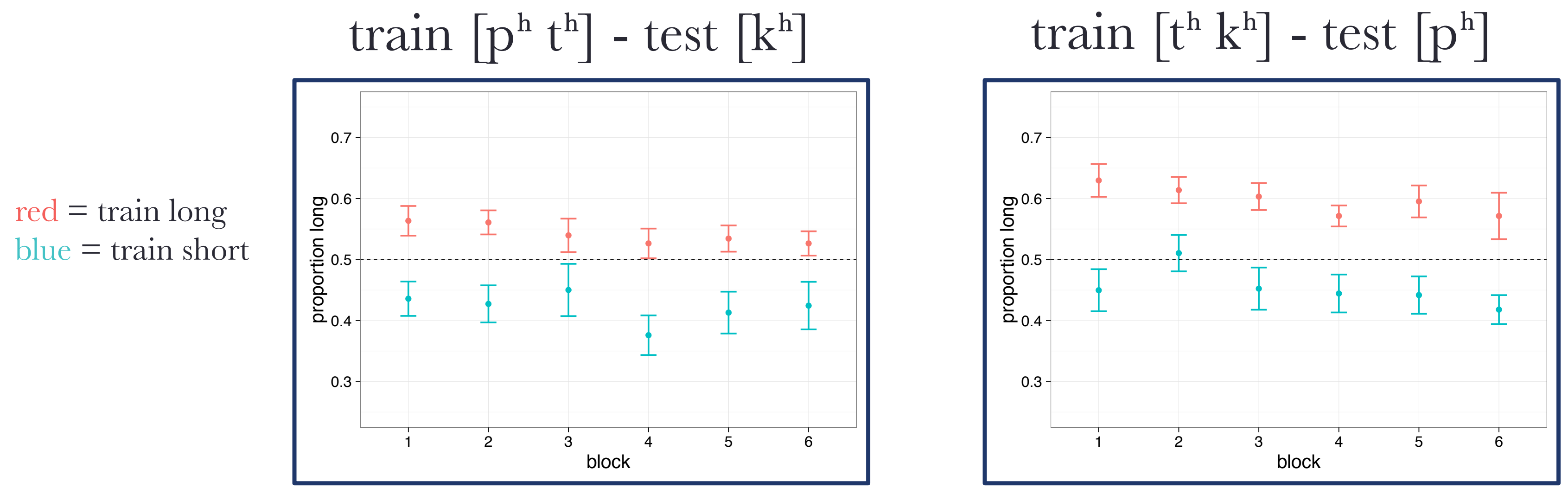
*Between-subjects manipulation
(four separate conditions)

Participants

- 48 participants (12 per condition, 34 female)

Perceptual Adaptation: Results

- Listeners showed generalized adaptation to both long and short VOTs across place of articulation
  - Significant generalization observed in all four conditions
  - Moderately lower sensitivity between long and short stimuli in the 'train long [pʰ tʰ]– test [kʰ]' condition
  - Difference of log VOT values (= log of VOT ratio) provided best quantitative account of congruency effect on choice responses (see condition x vot.ratio in logistic mixed-effects model), but similar results with VOT difference
- Evidence for early adaptation
  - Found significant interaction between condition and vot.ratio in the first block (27×2 exposure stimuli)
  - Bias to select first choice in all conditions (e.g., Yeshurun et al. 2008; Garcia-Perez & Alcala-Quintana 2011)



Logistic Mixed Effects Model

response #1 ~ 1 + condition*vot.ratio + (1 + vot.ratio | subj) + (1 | base.word)
condition = long (+1) or short (-1) | vot.ratio = log(VOT #1 / VOT #2)

Test [kʰ] All blocks	$\beta_0 = 0.36$, $p < 0.001$ $\beta_{\text{cond×vot.ratio}} = 0.28$, $p < .001$	Test [pʰ] All blocks	$\beta_0 = 0.27$, $p < 0.05$ $\beta_{\text{cond×vot.ratio}} = 0.36$, $p < .001$
Test [kʰ] First block	$\beta_0 = 0.42$, $p < 0.001$ $\beta_{\text{cond×vot.ratio}} = 0.25$, $p < .01$	Test [pʰ] First block	$\beta_0 = 0.06$, $p = 0.06$ $\beta_{\text{cond×vot.ratio}} = 0.43$, $p < .001$

Signal Detection Analysis (Wickens 2001)

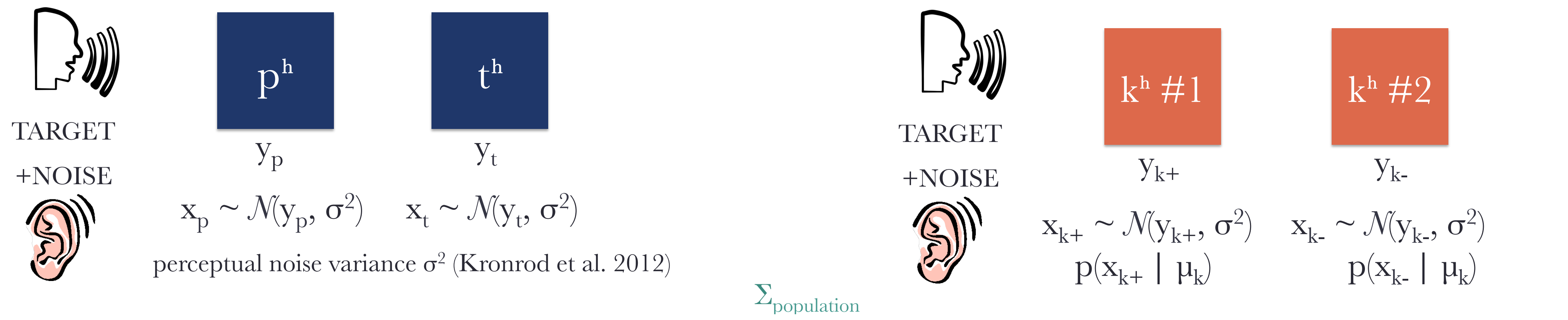
Sensitivity (*d'*) to difference between long and short VOTs
Response bias (log β) in selecting the first choice

Test [kʰ]	d'	log β	Test [pʰ]	d'	log β
Long	0.22 $p < .01$	0.35 $p < .05$	Long	0.53 $p < .001$	0.60 $p < .01$
Short	0.41 $p < .01$	0.41 $p < .01$	Short	0.26 $p = .06$	0.34 $p < .01$

Computational model

Exposure and Adaptation

Response Selection (long #1 - short #2 trial)



y_p $\quad$ y_t

$x_p \sim \mathcal{N}(y_p, \sigma^2)$ $\quad$ $x_t \sim \mathcal{N}(y_t, \sigma^2)$

perceptual noise variance σ^2 (Kronrod et al. 2012)

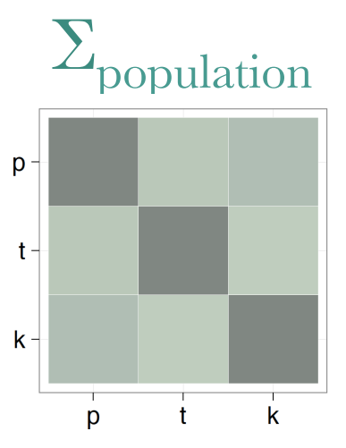
y_{k+} $\quad$ y_{k-}

$x_{k+} \sim \mathcal{N}(y_{k+}, \sigma^2)$ $\quad$ $x_{k-} \sim \mathcal{N}(y_{k-}, \sigma^2)$

$p(x_{k+} \mid \mu_k)$ $\quad$ $p(x_{k-} \mid \mu_k)$

Adaptation to novel talker

Update posterior distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ on talker mean ('target') VOT values $\boldsymbol{\mu} = [\mu_p\, \mu_t\, \mu_k]^T$ by sequential application of Bayes' Theorem to exposure stimuli. Initial prior is $\mathcal{N}(\boldsymbol{\mu}_{\text{population}}, \Sigma_{\text{population}})$.



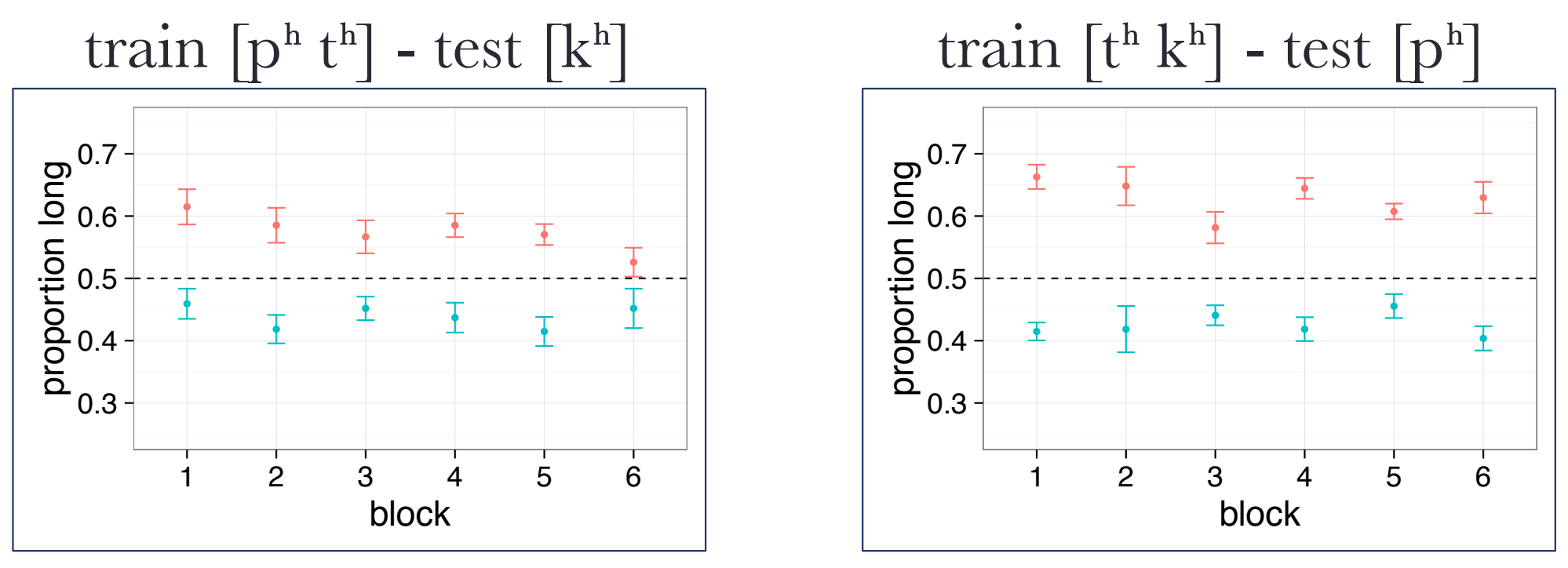
Probabilistic response rule

$$p(\text{respond } \#1) \propto \gamma_{\text{lapse}} \cdot \text{bias}(\#1) + (1 - \gamma_{\text{lapse}}) \cdot p(x_{k+} \mid \boldsymbol{\mu}) \,/\, [p(x_{k+} \mid \boldsymbol{\mu}) + p(x_{k-} \mid \boldsymbol{\mu})]$$

Population mean and covariance were inferred from the same laboratory production study used for stimulus creation (see Methods)

covariance prior + biased guessing

Listener uses population mean and covariance of stop categories for adaptation, and guesses choice 1 more often in 'lapse' trials



biased guessing

Listener responds choice 1 with stimulus-independence bias

independent preference + biased guessing

Listener has condition-independent preferences for stop-specific VOT values, and a bias for choice 1 in 'lapse' trials

BIC	**Test [pʰ]**	**Test [kʰ]**
biased guessing	6024	6059
VOT preference + biased guessing	6024	6061
covariance prior + biased guessing	5894	5954

BIC favors the covariance-based adaptation model over either biased guessing or independent preference models

Discussion

Found perceptual generalization of VOT across place of articulation

- Observed for both long and short VOT distributions
  - Present findings are consistent with Theodore & Miller (2010), who also observed generalization of long and short VOT values in a perceptual task with more extreme VOT manipulations
  - Previous studies have found no perceptual learning or generalization when listeners were asked to *produce* a shortened VOT (e.g., Nielsen 2011) or voiceless VOT values were more consistent with *unaspirated* stops (e.g., Clarke & Luce 2005)
- Observed with less exaggerated and more variable VOT values
  - Manipulation more representative of natural talker variation
  - Commensurately weaker adaptation effects relative to Theodore & Miller (2010)
- Generalization occurs rapidly, with minimal exposure (27 × 2 instances)

Prior knowledge of VOT covariation accounts for generalization results

- Biased guessing or condition-independent VOT preference do not account for perceptual results
- Alternative models include:
  - Prior linear relationships between stop categories (e.g., Chodroff & Wilson, submitted)
  - Estimate grand mean VOT of voiceless stops from exposure items and perform mean subtraction or z-scoring for normalization (e.g., Lobanov 1971; Nearey 1978; McMurray & Jongman 2011)
  - Dispreference for test VOTs that disobey the rank order (VOT [pʰ] < VOT [kʰ]) relative to exposure VOTs. Predicts less generalization in the 'train short – test [kʰ]' and 'train long – test [pʰ]' conditions as both test VOTs generally obey the ranking (short - k^h: $\chi^2(1) = 118.9$, long - p^h: $\chi^2(1) = 393.2$; *p*s < 0.001)
- Data indicates substantial perceptual noise (see also Kronrod et al. 2012) and strong response bias
- How much noise is present in VOT perception? How does the response bias relate to perceptual noise? Why do we find a primacy bias when some sequential two-interval experiments find recency biases?

Perception results and modeling indicate that listeners exploit covariation among stop VOTs in talker adaptation

References & Acknowledgments

References. Chodroff, E., Godfrey, J., Khudanpur, S., Wilson, C. (2015). Structured variability in acoustic realization: A corpus study of voice onset time in American English stops. *Proc. 18th ICPhS.* Glasgow, UK. • Chodroff, E., Wilson, C. (2014). Burst spectrum as a cue for the stop voicing contrast in American English. *J. Acoust. Soc. Amer.*, *136*(5), 2762–2772. • García-Pérez, M. A., Alcalá-Quintana, R. (2011). Interval bias in 2AFC detection tasks: sorting out the artifacts. *Atten. Percep. Psychophys.*, *73*(7), 2332–2352. • Joos, M. (1948). Acoustic Phonetics. *Language*, *24*(2), 5–136. • Kronrod, Y., Coppess, E., Feldman, N. H. (2012). A unified model of categorical effects in consonant and vowel perception. *Proc. 34th Ann. Conf. Cog. Sci. Soc.*, 629–634. • Lobanov, B. M. (1971). Classification of Russian Vowels Spoken by Different Speakers. *J. Acoust. Soc. Amer.*, *49*(2-2), 606–608. • McMurray, B., Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psych. Review*, *118*(2), 219–246. • Nearey, T. M., Assmann, P. F. (2007). Probabilistic "Sliding Template" Models for Indirect Vowel Normalization. In M.-J. Solé, P. S. Beddor, & M. Ohala (Eds.), *Exp. App. Phon.* (pp. 246–270). New York: Oxford University Press. • Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *J. Phon.*, 39(2):132–142. • Theodore, R., Miller, J. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *J. Acoust. Soc. Amer.*, 128(4):2090–2099. • Wickens, T. D. (2001). *Elementary signal detection theory*. Oxford University Press. • Yeshurun, Y., Carrasco, M., Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research*, *48*(17), 1837–1851.

Acknowledgments. We would like to thank Chris Kirov, the JHU Phonetics and Phonology Lab, and the audience at the JHU Center for Language and Speech Processing for helpful discussion.
This work was supported by a Distinguished Science of Learning Pre-doctoral Fellowship from the JHU Science of Learning Institute and a fellowship from the Dolores Zohrab Liebmann Fund.