Generalization in VOT imitation: Feature adaptation or acoustic covariation?

Colin Wilson[1], Eleanor Chodroff[1], Kuniko Nielsen[2]
[1]Johns Hopkins University, [2]Oakland University

4pSC68

Generalized adaptation to novel talkers

Talkers vary considerably in the phonetic realization of speech sounds (e.g., Peterson & Barney, 1952; Newman *et al.*, 2001; Allen et al., 2003; Chodroff & Wilson, under review)

Listeners readily adapt to novel talker phonetics in a way that ***generalizes*** across words and sound categories

- Generalization across ***words***
  (e.g., Nygaard et al., 1994; Norris et al., 2003; Allen & Miller, 2004; McQueen et al., 2006; Nielsen, 2011)
- Generalization across ***sounds***
  (e.g., vowels: Ladefoged & Broadbent, 1957; Maye et al., 2008; stops: Eimas & Corbit, 1973; Kraljic & Samuel, 2006; Theodore & Miller, 2010; Nielsen, 2011; but cf. Cooper, 1979; Clarke & Luce, 2005)

Generalized talker adaptation is observed in speech perception and in **phonetic imitation/convergence** (e.g., Nielsen, 2011)

- What is the rational basis for generalization across sounds?
  - Talker-specific phonetic realizations of different sounds are *mutually predictable* (i.e., not independent)
  - Covariation of talker-specific phonetics results from many anatomical and (socio-)linguistic factors (e.g., differences in vocal tract length, speaking style)
- How do listeners represent covariation across talkers?
  - In Bayesian models of speech perception/adaptation, listeners have a *prior* distribution on talker phonetics
    (e.g., Nielsen & Wilson, 2008; Feldman et al., 2009; Pająk *et al.*, 2013; Kleinschmidt & Jaeger 2015, 2016)
  - Listener's prior might encode covariation relations among sound categories *directly* or via *features/gestures*

Adaptation models

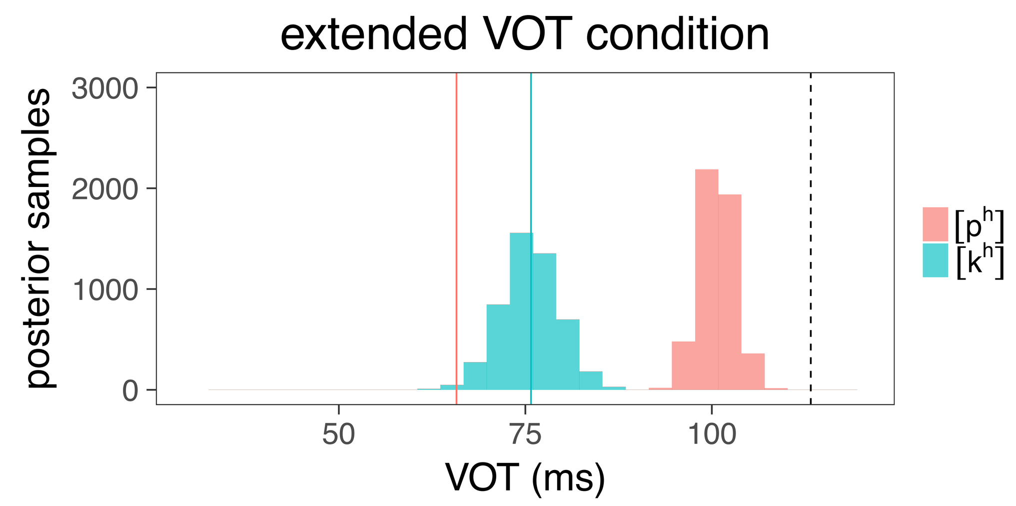
Independence model

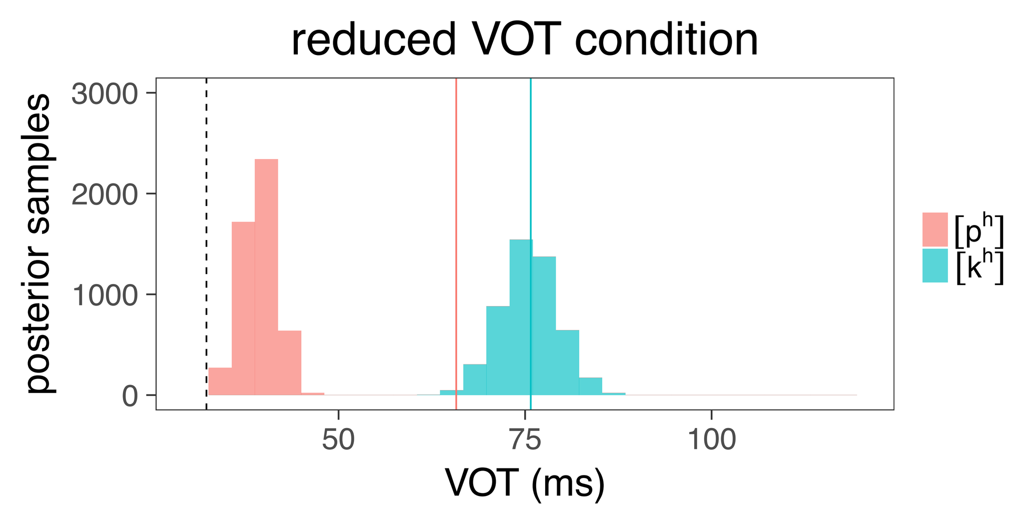
- Listeners have knowledge of how the VOT distributions of [pʰ] and [kʰ] *vary* across talkers, but do not represent category *covariation*
- Predicts parochial adaptation: no generalization from one phonetic category to another, even for the same acoustic property (VOT)

$$\begin{bmatrix}\alpha_p^*\\ \alpha_k^*\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_p\\ \mu_k\end{bmatrix}, \begin{bmatrix}\sigma_p^2 & 0\\ 0 & \sigma_k^2\end{bmatrix}\right) \quad \text{// novel talker}$$

$$x_i \sim \text{Gamma}(\alpha_p^*, \beta) \quad \text{// novel talker productions}$$

$$y_i \sim \mathcal{N}(x_i, \sigma_{noise}^2) \quad \text{// perceived VOT values}$$





$\mu_p = 32.6$ $\mu_k = 37.8$ $\sigma_p = 7.3$ $\sigma_k = 6.0$ $\beta = 0.5$ $\sigma_{noise} = 10.0$

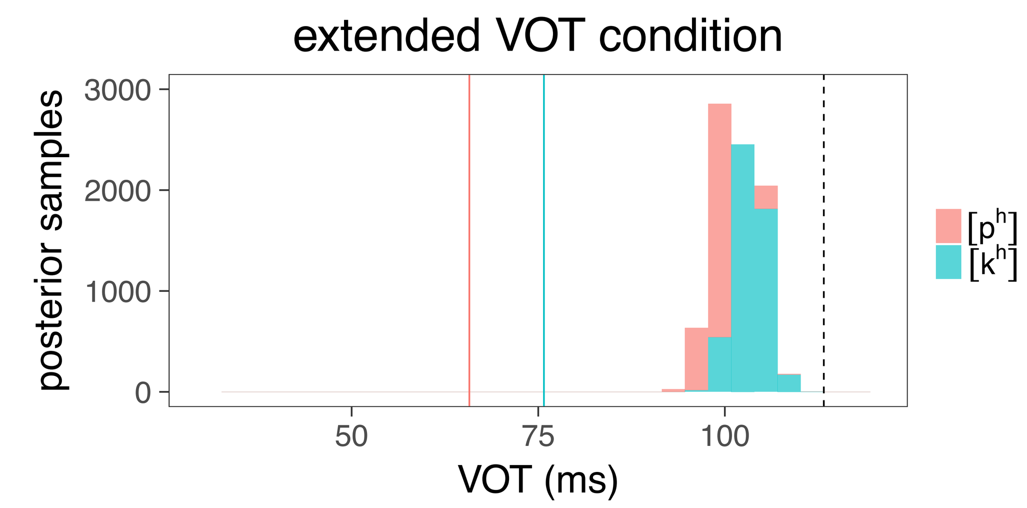
Category-based covariation model

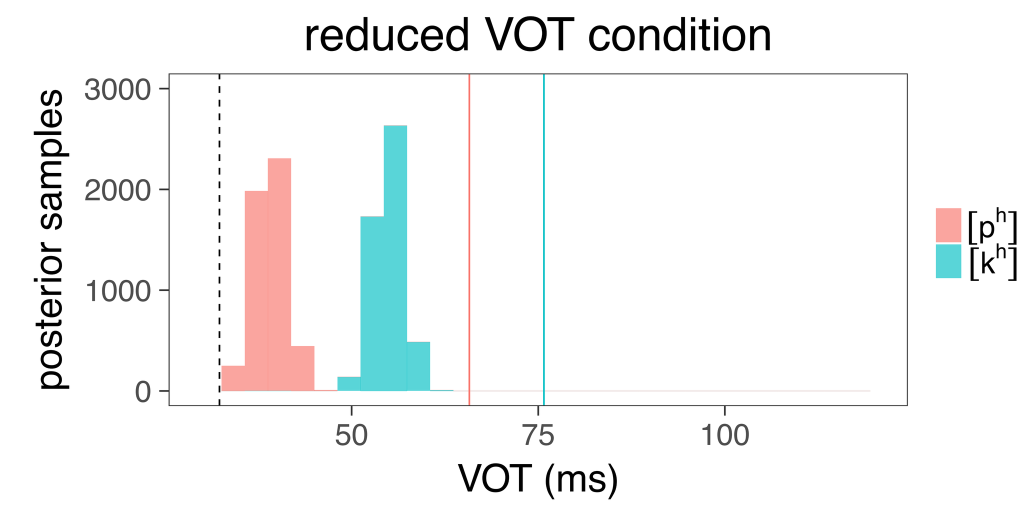
- Listeners represent *covariation* of VOT distributions for [pʰ] and [kʰ] directly, with a correlation coefficient (ρ) relating the two categories
- Predicts generalized adaptation, but does not enforce the empirical relation VOT([pʰ]) < VOT([kʰ]) even in the absence of [kʰ] exposure

$$\begin{bmatrix}\alpha_p^*\\ \alpha_k^*\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_p\\ \mu_k\end{bmatrix}, \begin{bmatrix}\sigma_p^2 & \rho\sigma_p\sigma_k\\ \rho\sigma_k\sigma_p & \sigma_k^2\end{bmatrix}\right) \quad \text{// novel talker}$$

$$x_i \sim \text{Gamma}(\alpha_p^*, \beta) \quad \text{// novel talker productions}$$

$$y_i \sim \mathcal{N}(x_i, \sigma_{noise}^2) \quad \text{// perceived VOT values}$$





$\mu_p = 32.6$ $\mu_k = 37.8$ $\sigma_p = 7.3$ $\sigma_k = 6.0$ $\rho = 0.96$ $\beta = 0.5$ $\sigma_{noise} = 10.0$

Feature-/gesture- based covariation model

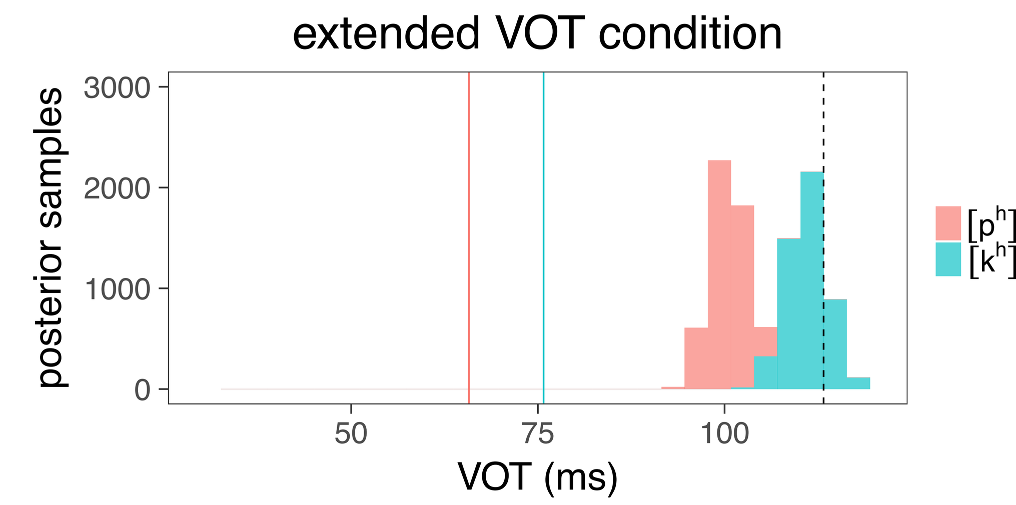
- Listeners represent *covariation* of VOT distributions for [pʰ] and [kʰ] indirectly, via decomposition into [spread glottis] and [dorsal] properties
- Predicts generalized adaptation, and enforces the empirical relation VOT([pʰ]) < VOT([kʰ]) in the absence of evidence to the contrary

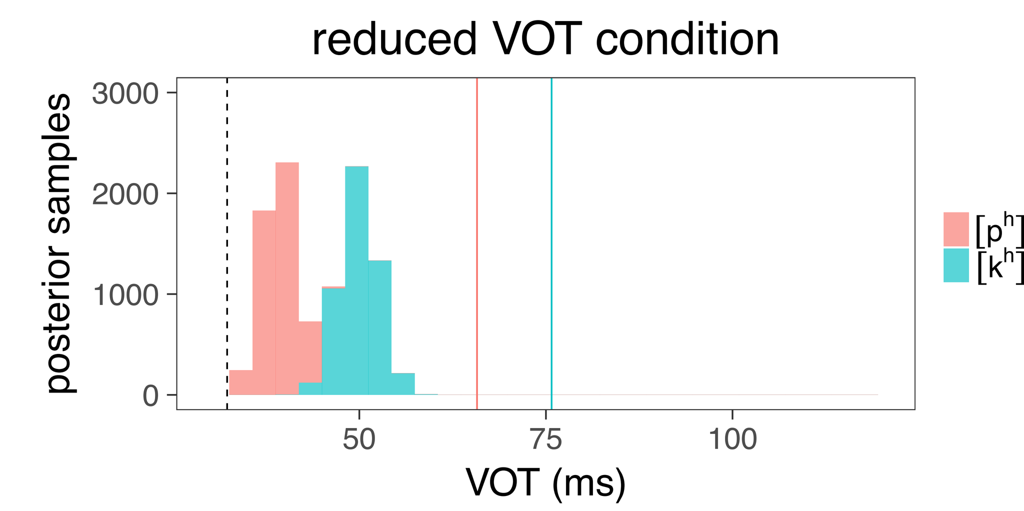
$$\begin{bmatrix}b_{[s.g.]}^*\\ b_{[dor]}^*\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_{[s.g.]}\\ \mu_{[dor]}\end{bmatrix}, \begin{bmatrix}\sigma_{[s.g.]}^2 & 0\\ 0 & \sigma_{[dor]}^2\end{bmatrix}\right) \quad \text{// novel talker}$$

$$\begin{bmatrix}\alpha_p^*\\ \alpha_k^*\end{bmatrix} \leftarrow \begin{bmatrix}1 & 0\\ 1 & 1\end{bmatrix}\begin{bmatrix}b_{[s.g.]}^*\\ b_{[dor]}^*\end{bmatrix} = \begin{bmatrix}b_{[s.g.]}^*\\ b_{[s.g.]}^* + b_{[dor]}^*\end{bmatrix}$$

$$x_i \sim \text{Gamma}(\alpha_p^*, \beta) \quad \text{// novel talker productions}$$

$$y_i \sim \mathcal{N}(x_i, \sigma_{noise}^2) \quad \text{// perceived VOT values}$$





$\mu_{[s.g.]} = 32.6$ $\mu_{[dor]} = 5.3$ $\sigma_{[s.g.]} = 7.3$ $\sigma_{[dor]} = 2.0$ $\beta = 0.5$ $\sigma_{noise} = 10.0$

Generalized adaptation and phonetic covariation in Nielsen (2011)

Extended VOT condition (N = 27 AE participants)

- Pre-exposure production of 120 critical stop-initial words 100 [pʰ]-initial / 20 [kʰ]-initial & 30 sonorant-initial fillers
- Listening to 80 familiarization items, a subset of the **[pʰ]-initial** critical words, with VOT extended by approx. +40 ms
- Post-exposure production of critical words & fillers

Generalized imitation: participants imitated extended VOT for heard and unheard [pʰ] words, and crucially unheard [kʰ] words

Mixed-effects model with random intercept and slopes
$\beta_{\text{pre-vs.-post}} = 3.46$ ($t = 4.61$), $\beta_{\text{k-vs.-p}} = 4.43$ ($t = 4.67$)
Interaction between pre-vs.-post and stop n.s. ($\beta = -0.03$)

Reduced VOT condition (N = 25 AE participants)

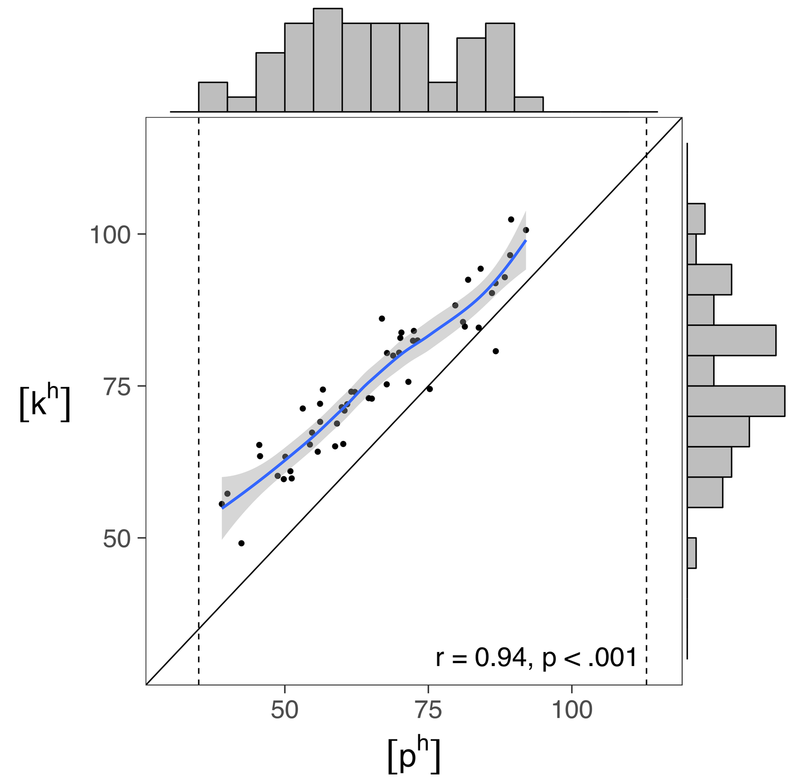
- Identical to extended condition except that VOT of familiarization items was reduced by approx. -40 ms.

No sig. imitation: participants did not imitate reduced VOT for heard or unheard [pʰ] words, let alone for unheard [kʰ] words

Mixed-effects model with random intercept and slopes
$\beta_{\text{pre-vs.-post}} = 0.00$ ($t = 0.01$), $\beta_{\text{k-vs.-p}} = 5.26$ ($t = 5.01$)
Interaction between pre-vs.-post and stop n.s. ($\beta = -0.31$)

See Nielsen (2011) for additional analyses and discussion

- A plausible explanation for generalized adaptation in the extended VOT condition is that participants *extrapolated* from familiarization: novel talker has long [pʰ] VOT ⇝ novel talker has long [kʰ] VOT
- Generalized adaptation across aspirated stops is rational given *robust VOT covariation* across talkers (e.g., Zlatin, 1974; Koenig, 2000; Newman, 2003; Theodore et al., 2009; Chodroff & Wilson, u.r.)
- VOT covariation is evident, replicating previous findings, in *pre-exposure* productions of all participants from Nielsen (2011)



Key
- Each participant has a pair of VOT means
- Marginal histograms display cross-talker variation for each stop
- Loess fit shows strong [pʰ] - [kʰ] covariation across sampled range
- Dotted lines indicate averages of extended and reduced [pʰ]-initial familiarization stimuli

Discussion

- AE talkers vary substantially in their mean VOT values for word-initial aspirated stops (as for other aspects of phonetic realization)
  - Pre-exposure: [pʰ] range: 39ms – 92ms [kʰ] range: 49ms – 102ms
  - Importantly, VOT means tightly *covary* across talkers ($r > 0.90$)
- Generalized adaptation to extended VOT is incompatible with a model in which listeners represent variation but not covariation
- Covariation prior could be stated at two levels of representation:
  - Direct relationship of cue covariation between phonetic categories
  - Relationship between categories mediated by features / gestures (Nielsen & Wilson 2008, Pająk *et al.*, 2013)
- Both covariation models predict generalization of talker adaptation from heard [pʰ] to unheard [kʰ] (and unheard words, unheard [tʰ], …)
  - Category-based model allows inferred VOT of [pʰ] to surpass that of [kʰ], reversing typical order, if target for [pʰ] is sufficiently long
  - Feature-based model predicts inferred VOT([pʰ]) < VOT([kʰ]), and *parallel adaptation for both categories*, in line with Nielsen (2011)
- Models predict adaptation in the reduced VOT condition, but imitation was n.s. Is this a difference between *perceptual* adaptation and *production* convergence? Do listeners have more complex / asymmetric prior?

In Bayesian models of adaptation, the prior is key to understanding how listeners *generalize* from their experience with a novel talker.

Modeling details

- Multivariate Gaussian priors over talker-specific parameters were estimated from pre-exposure productions of Nielsen (2011): lab/careful-speech register
- VOT distribution for each stop category within a talker was modeled with a Gamma(α,β) distribution (e.g., Goldrick et al., 2011, Chodroff et al., 2016)
  - Asymmetric distribution with longer right tail (cf. Gaussian)
  - $E[x] = \alpha/\beta$, $Var[x] = \alpha/\beta^2$, here $\beta = 0.5 \Rightarrow$ within-category VOT variability increases with the mean (Chodroff & Wilson, under review)
- Noise in listeners' perception of VOT, and other sources of unintended variability, modeled with Gaussian distribution ($\sigma \approx 10$ms, Kronrod et al., 2016)
- Inference of talker-specific parameters conditioned on perceived exposure stimuli was performed with MCMC sampling in Stan (Carpenter et al., in press)

$$\log p(\text{talker params} \mid \text{percepts}) \propto \log p(\text{percepts} \mid \text{params}) + \lambda \log(\text{params} \mid \text{prior})$$

  - Parameter λ scales prior relative to likelihood (in figures above, $\lambda = 10.0$)
  - Experimental/talker/listener effects on adaptation can be modeled by varying λ (e.g., $\lambda \to 0$ predicts max. adaptation, $\lambda \to \infty$ no adaptation)

Acknowledgements

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.