Generalization in VOT imitation: Feature adaptation or acoustic covariation?

Colin Wilson¹, Eleanor Chodroff¹, Kuniko Nielsen²

¹Johns Hopkins University, ²Oakland University

Abstract

Listeners readily adapt to novel talker phonetics in a way that generalizes across words and sound categories. We examined this adaptation in a paradigm that varied the acoustic stimulus while the talker's phonetic realization remained the same. Listeners were exposed to novel talker productions of p- and b- stops: listeners did not imitate reduced VOT for unheard [p] and [b] words, nor did they generalize their adaptation to unseen novel talker productions. However, listeners showed feature adaptation in the reduced VOT condition, which we interpret as evidence for covariation-based adaptation.

Introduction

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; Nielsen & Chodroff, 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, 2013)

• Generalization across sounds
  (e.g., vowels: Ladefoged & Broadbent, 1957; Maye et al., 2008; stops: Eimas & Corbit, 1973; Kraije & Saelou, 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., 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, 2008; Feldman et al., 2009; Pajk et al., 2013; Kleinschmidt & Jaeger, 2015, 2016)
  - Listener’s prior might encode covariation relations among sound categories directly or via features/gestures

Generalization in VOT imitation: Feature adaptation or acoustic covariation?

Extended VOT condition (N = 27 AE participants)

• Pre-exposure production of 120 critical stop-initial words 100 [p]-initial / 20 [b]-initial & 30 sonorant-initial fillers

• Listeners to 80 familiarization items, a subset of the [p]-
  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 [p]-words, and crucially unheard [b]-words

Mixed-effects model with random intercept and slopes $\beta_{\text{pre-exposure}} = 3.46$ (p = 4.61); $\beta_{\text{post-exposure}} = 4.43$ (p = 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 [p]-words, let alone for unheard [b]-words

Mixed-effects model with random intercept and slopes $\beta_{\text{post-exposure}} = 0.00$ (p = 5.01); $\beta_{\text{post-exposure}} = 5.26$ (p = 5.01)

Interaction between pre-vs-post and stop n.s. $\beta = -0.31$

See Nielsen (2011) for additional analyses and discussion

Discussion

• AE talks vary substantially in their mean VOT values for word-initial aspirated stops (as for other aspects of phonetic realization)
  - Pre-exposure: [p]-range = 39ms – 92ms; [b]-range = 40ms – 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

• Both covariation models predict generalization of talker adaptation from heard [p] to unheard [b] (and unheard words, unheard [b])
  - Category-based model allows inferred VOT of [p] to_superscript_{b} (to surpass that of [b], reversing typical order, if target for [p] is sufficiently long)
  - Feature-based model predicts inferred VOT([p]) < VOT([b]), 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.

Generalization in VOT imitation: Feature adaptation or acoustic covariation?

Adaptation models

Independence model

• Listeners have knowledge of how the VOT distributions of [p] and [b] vary across talkers, but do not represent category orientation

$\alpha_{\text{p}}^2 \sim N(0,\sigma^2_{\text{p}})$ / novel talker

$\alpha_{\text{b}}^2 \sim N(0,\sigma^2_{\text{b}})$ / novel talker

• Predicts parochial adaptation: no generalization from one phonetic category to another, even for the same acoustic property (VOT)

Category-based covariation model

• Listeners represent covariation of VOT distributions for [p] and [b] directly, with a correlation coefficient ($\rho$) relating the two categories

$\alpha_{\text{p}}^2 \sim N(\mu_{\text{p}},\sigma^2_{\text{p}})$ / novel talker

$\alpha_{\text{b}}^2 \sim N(\mu_{\text{b}},\sigma^2_{\text{b}})$ / novel talker

$\rho = \frac{\alpha_{\text{p}} \cdot \alpha_{\text{b}}}{\sigma_{\text{p}} \cdot \sigma_{\text{b}}}$

• Predicts generalized adaptation, but does not enforce the empirical relation $VOT([p]) < VOT([b])$ even in the absence of [b] exposure

$\alpha_{\text{p}}^2 \sim N(\mu_{\text{p}},\sigma^2_{\text{p}})$ / novel talker

$\alpha_{\text{b}}^2 \sim N(\mu_{\text{b}},\sigma^2_{\text{b}})$ / novel talker

• 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?

Feature-gesture-based covariation model

• Listeners represent covariation of VOT distributions for [p] and [b] indirectly, via decomposition into speech gestures (and [b]) properties

• Predicts generalized adaptation, and enforces the empirical relation $VOT([p]) < VOT([b])$ in the absence of evidence to the contrary

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?

Feature-gesture-based covariation model

• Listeners represent covariation of VOT distributions for [p] and [b] indirectly, via decomposition into speech gestures (and [b]) properties

• Predicts generalized adaptation, and enforces the empirical relation $VOT([p]) < VOT([b])$ in the absence of evidence to the contrary

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?

Feature-gesture-based covariation model

• Listeners represent covariation of VOT distributions for [p] and [b] indirectly, via decomposition into speech gestures (and [b]) properties

• Predicts generalized adaptation, and enforces the empirical relation $VOT([p]) < VOT([b])$ in the absence of evidence to the contrary

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?

Feature-gesture-based covariation model

• Listeners represent covariation of VOT distributions for [p] and [b] indirectly, via decomposition into speech gestures (and [b]) properties

• Predicts generalized adaptation, and enforces the empirical relation $VOT([p]) < VOT([b])$ in the absence of evidence to the contrary

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?