An empirical and computational study of generalized adaptation to natural talker-specific VOT
Eleanor Chodroff (chodroff@cogsci.jhu.edu), Alessandra Golden, and Colin Wilson
Johns Hopkins University

Substantial variability exists in the phonetic realization of speech sounds across talkers, yet listeners
adapt with relative ease. One source of information that could be exploited in adapting to novel talkers is
knowledge of acoustic covariance across phonetic categories. This type of covariation has most notably
been found for vowels, with talkers having congruent vowel spaces that are shifted relative to one another
in the F1xF2 plane (e.g., Peterson & Barney, 1952; Nearey, 1989). Strong correlations have also been
found for voice onset time (VOT) of word-initial aspirated stop consonants: for example, talkers with
longer mean VOT values for [pʰ] tend to have longer values for [kʰ] (Theodore et al., 2009). More
recently, VOT correlations at or above \( r = .80 \) have been identified for all three aspirated stops of
American English in a large read speech corpus containing more than 100 talkers (Chodroff et al., 2015).

Some evidence that knowledge of VOT correlations plays a role in talker adaptation has been
provided by previous studies of perceptual generalization and phonetic imitation (e.g., Theodore &
Miller, 2010; Nielsen, 2011). Existing findings are limited, however, because the relevant talker
differences involved highly exaggerated VOT values (e.g., 88 ms vs. 183 ms), extensive exposure to the
novel talker (e.g., 120 exposure trials before testing), or limited stimulus variability (e.g., only two VOTs
per stop and talker). The present study employed more natural and variable stimuli, with alternating
exposure and testing, to investigate the use of covariation in rapid perceptual adaptation.

**Perceptual generalization.** Stimuli: To address limitations of previous generalization studies, we
generated stimuli with substantial variability from more natural VOT distributions. All stimuli were
created from careful-speech productions of CVC syllables in a previously collected laboratory corpus
(Chodroff & Wilson, 2014). The syllables were composed of six stop consonants /p t k b d g/ crossed with
nine vowels /i e æ ʌ a o u/ and a final /t/. Gamma densities, which closely approximate the shapes of
natural VOT distributions, were fit to tokens for each aspirated stop from two of the male talkers: one
with naturally long VOT values (long VOT talker: [pʰ] mean = 91 ms (sd = 17), [tʰ] 108 (16), [kʰ] 102
(15)), and one with naturally short values (short VOT talker: [pʰ] 60 (18), [tʰ] 72 (12), [kʰ] 77 (14)). For
each aspirated stop and each vowel, three repetitions of the first male talker’s productions were selected
and their VOTs were manipulated to match values randomly generated from the long and short
distributions; the result was 162 stimuli (3 stops x 9 vowels x 3 repetitions x 2 VOTs).

Procedure: There were four conditions in the experiment. For each condition, two of the aspirated
stops ([pʰ tʰ] or [kʰ]) were selected as training categories and assigned either relatively long or relatively
short VOT levels. Within a trial, one instance of each of the training stops with the appropriate VOT level
was presented, and then the participant performed a two-alternative forced-choice task for the untrained
stop (i.e., [kʰ] or [pʰ]). The choices differed only in VOT (long vs. short), and participants were asked to
select the one that "sounded most like the talker." Vowel category was held constant within a trial, and
testing stimuli were counterbalanced such that the long VOT option was presented first in exactly half of
the trials. Each participant (N = 10-11 per condition) completed 6 blocks with 27 unique trials per block.

Results: Inspection of the results revealed a strong bias to respond with the first stimulus regardless of
choice order or condition. We performed two analyses to determine whether, despite this bias,
participants showed generalization of the VOT level to the untrained stop. A logistic mixed-effects
analysis included an intercept (representing the bias to choose the first option) and a binary predictor that
indicated whether the first option was congruent (+1) or incongruent (−1) with the training stops (e.g.,
when exposure involved long VOT [pʰ tʰ], the longer VOT [kʰ] is the congruent option). Listeners
generalized talker VOT significantly in all conditions (e.g., Fig. 1a) except when exposed to short values
and tested on [pʰ] (Expose Long, Test [kʰ]: \( \beta_1 = 0.21 \ p < 0.01 \); Expose Short, Test [kʰ]: \( \beta_1 = 0.31 \ p < 0.01 \);
Expose Long, Test [pʰ]: \( \beta_1 = 0.40 \ p < 0.001 \); Expose Short, Test [pʰ]: \( \beta_1 = 0.16 \ p = 0.05 \)).

These findings were supported by a second analysis assessing sensitivity (\( d' \)) and response bias (log
\( \beta \)) for each participant (e.g., Wickens, 2001). Sensitivity to the difference between long and short VOTs
for the test stop was significantly different from chance (\( d' = 0 \) in all but one condition (Expose Long,
Test \([k]^3\): \(d' = 0.26 p < 0.05, \log \beta = 0.48\); Expose Short, Test \([k]^3\): \(d' = 0.38 p < 0.001, \log \beta = 0.42\); Expose Long, Test \([p]^3\): \(d' = 0.50 p < 0.001, \log \beta = 0.42\); Expose Short, Test \([p]^3\): \(d' = 0.20 p = 0.10, \log \beta = 0.29\). These results indicate that listeners exploit knowledge of VOT correlations in the population to generalize relatively small talker-specific effects from two of the aspirated stops to the third one.

A central finding is that perceptual adaptation occurs rapidly, with listeners tuning expectations about a talker’s voice after brief exposure (e.g., Morton et al., 2015). Accordingly, we found that the same pattern of results emerged within the first two blocks (Expose Long, Test \([k]^3\]): \(\hat{\beta}_1 = 0.25 p < 0.01, \log \beta = 0.32 p < 0.05\); Expose Short, Test \([k]^3\): \(\hat{\beta}_1 = 0.26 p < 0.05, \log \beta = 0.34 p < 0.05\); Expose Long, Test \([p]^3\): \(\hat{\beta}_1 = 0.58 p < 0.001, \log \beta = 0.58 p < 0.001\); Expose Short, Test \([p]^3\): \(\hat{\beta}_1 = 0.94, \log \beta = 0.01 p = 0.93\).

**Figure 1.** Proportion long VOT selected in the two Test \([k]^3\) conditions of Experiment 1 (a) and by the adaptation model (b). Red bars: Expose Long condition. Blue bars: Expose Short condition.

**Adaptation model.** We developed a computational account of performance in the experiment that extends previous models of speech perception and adaptation (e.g., Nielsen & Wilson, 2008; Feldman et al., 2009). For concreteness, we focus on a single trial in the Expose Long, Test \([k]^3\) condition. The listener is first exposed to instances of \([p]^3\) and \([k]^3\) with relatively long VOT values \(y_p\) and \(y_k\). Because all sensory perception is subject to noise (e.g., Green & Swets, 1966), the listener perceives VOT values \(x_p \sim N(y_p, \sigma^2)\) and \(x_k \sim N(y_k, \sigma^2)\) where \(N\) denotes a Gaussian distribution and \(\sigma^2\) is the perceptual noise variance. After each exposure, the listener updates a posterior distribution on the talker’s mean VOTs for all three stops. The update is performed with Bayes' theorem, e.g., \(p(\mu | x_k) \propto p(x_k | \mu) \cdot N(\mu | \mu_{prior}, \Sigma_{prior})\) where \(\mu = [\mu_p, \mu_m, \mu_k]^T\) is the estimate of the talker’s VOT means and \(\mu_{prior}\), \(\Sigma_{prior}\) are the mean and covariance matrix of the Gaussian prior distribution. The prior parameters were initially inferred from the production study described earlier and then updated incrementally (i.e., the posterior for one exposure stimulus is the prior for the next). The covariance matrix \(\Sigma_{prior}\), a novel feature of our model, encodes the generalization that talker-specific VOT means are correlated and is critical for predicting performance on the test stop.

The trial continues with the presentation of long and short VOT choices for \([k]^3\), \(y_k\), and \(y_k\). The listener perceives noisy versions of the choices, \(x_k\), and \(x_k\), and computes the probability density of each one, \(p(x_k | \mu_k)\) and \(p(x_k | \mu_k)\), according to the current talker \(\mu\) estimate. With these computations, the model matched the qualitative finding that listeners generalize the expectation of longer VOT to \([k]^3\). However, the model overestimated the extent of adaptation even with a non-optimal, probabilistic response rule (i.e., \(p(\text{respond long}) \propto p(x_k | \mu_k)\)). Adding a bias to choose the first stimulus brought the models predictions much closer to human performance across the four conditions (e.g., Fig. 1b).