



Distributional learning for speech reflects cumulative exposure to a talker's phonetic distributions

Rachel M. Theodore^{1,2} · Nicholas R. Monto^{1,2}

Published online: 2 January 2019
© Psychonomic Society, Inc. 2018

Abstract

Efficient speech perception requires listeners to maintain an exquisite tension between stability of the language architecture and flexibility to accommodate variation in the input, such as that associated with individual talker differences in speech production. Achieving this tension can be guided by top-down learning mechanisms, wherein lexical information constrains interpretation of speech input, and by bottom-up learning mechanisms, in which distributional information in the speech signal is used to optimize the mapping to speech sound categories. An open question for theories of perceptual learning concerns the nature of the representations that are built for individual talkers: do these representations reflect long-term, global exposure to a talker or rather only short-term, local exposure? Recent research suggests that when lexical knowledge is used to resolve a talker's ambiguous productions, listeners disregard previous experience with a talker and instead rely on only recent experience, a finding that is contrary to predictions of Bayesian belief-updating accounts of perceptual adaptation. Here we use a distributional learning paradigm in which lexical information is not explicitly required to resolve ambiguous input to provide an additional test of global versus local exposure accounts. Listeners completed two blocks of phonetic categorization for stimuli that differed in voice-onset-time, a probabilistic cue to the voicing contrast in English stop consonants. In each block, two distributions were presented, one specifying /g/ and one specifying /k/. Across the two blocks, variance of the distributions was manipulated to be either narrow or wide. The critical manipulation was order of the two blocks; half of the listeners were first exposed to the narrow distributions followed by the wide distributions, with the order reversed for the other half of the listeners. The results showed that for earlier trials, the identification slope was steeper for the narrow-wide group compared to the wide-narrow group, but this difference was attenuated for later trials. The between-group convergence was driven by an asymmetry in learning between the two orders such that only those in the narrow-wide group showed slope movement during exposure, a pattern that was mirrored by computational simulations in which the distributional statistics of the present talker were integrated with prior experience with English. This pattern of results suggests that listeners did not disregard all prior experience with the talker, and instead used cumulative exposure to guide phonetic decisions, which raises the possibility that accommodating a talker's phonetic signature entails maintaining representations that reflect global experience.

Keywords Speech perception · Perceptual learning · Computational models · Distributional learning

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13423-018-1551-5>) contains supplementary material, which is available to authorized users.

✉ Rachel M. Theodore
rachel.theodore@uconn.edu

¹ Department of Speech, Language, and Hearing Sciences, University of Connecticut, 850 Bolton Road, Unit 1085, Storrs, CT 06269, USA

² Connecticut Institute for the Brain and Cognitive Sciences, University of Connecticut, 337 Mansfield Road, Unit 1872, Storrs, CT 06269, USA

Introduction

Variability in speech acoustics is highly structured, including variability associated with individual talkers' idiolects. For a given acoustic-phonetic property specifying a given speech sound contrast, talkers systematically differ with respect to the mode and variance of their distributions, and the degree to which their distributions are separated in acoustic-phonetic space (e.g., Hillenbrand et al., 1995; Newman et al., 2001; Theodore et al., 2009). There is a rich evidence base indicating that listeners use structured phonetic variation to optimize the mapping to linguistic representations on a talker-contingent

basis (e.g., Nygaard & Pisoni, 1998; Theodore & Miller, 2010; Theodore et al., 2015), even when a talker's input is potentially ambiguous (e.g., Norris et al., 2003). Theories of perceptual learning posit that listeners track phonetic cues with respect to a higher-order structure, such as a particular talker, to derive a structure-specific probabilistic mapping that optimizes phonetic categorization (e.g., Kleinschmidt & Jaeger, 2015; McMurray et al., 2009). On this view, adaptation reflects the process of modifying the mapping to speech sounds in line with statistical distributions in the input. Support for this theory comes from Clayards et al. (2008), who examined stop consonant identification between listeners who were exposed to either narrow or wide variances of voice-onset-time (VOT) distributions. Consistent with predictions generated by the ideal observer computational framework, the slope of the identification function was steeper (i.e., more categorical) for listeners who heard narrow compared to wide distributions, indicating greater uncertainty for more variable input (Clayards et al., 2008).

What has yet to be confirmed in the literature on statistical-based accounts of talker adaptation (and the literature on probabilistic inference in speech more generally) is the time course of experience that is used to guide processing. Does online comprehension of a talker's speech rely on recent experience, or rather does it reflect global experience? There is some evidence suggesting that adaptation to a talker's idiosyncratic input reflects cumulative experience. Kraljic et al. (2008) exposed listeners to a talker who produced a fricative with both canonical and acoustically ambiguous forms. The ambiguous form was embedded in an informative lexical context, and order in which listeners heard the two forms was manipulated. Hearing the ambiguous form first promoted perceptual learning to incorporate the ambiguity into the fricative category. However, learning was not observed given initial exposure to the canonical form, suggesting that listeners aggregated experience with the talker's productions, heavily weighting initial experience (see also Kraljic & Samuel, 2005). Additional evidence in support of cumulative experience with input distributions guiding speech perception comes from studies showing that a complete re-weighting of an atypical cue relationship is resistant to long-term training, suggesting that while listeners are sensitive to distributional shifts in cue relationships, acceptance of the new relationship reflects long-term exposure (e.g., Idemaru & Holt, 2011).

The Bayesian belief-updating model of speech adaptation (Kleinschmidt & Jaeger, 2015) posits that tension between flexibility and stability in the language architecture is achieved through cumulative tracking of a talker's input, where perceptual categories are modified to the degree that the input provides evidence that distributional beliefs formed through long-term experience with a language are not optimal for that talker. This claim was tested using the lexically guided perceptual learning paradigm (Saltzman & Myers, 2018). Two groups

of listeners heard a talker produce distributions of /s/ and // across two blocks. In one block, listeners heard clear /s/ and acoustically ambiguous // tokens; in the other block, listeners heard clear // and ambiguous /s/ tokens. Ambiguous tokens were embedded in informative lexical contexts, block order was manipulated between the two groups, and all listeners completed a phonetic categorization task for a *shine-sign* continuum following each block. In each block, listeners who heard modified /s/ showed more *sign* responses than those who heard modified //, and the learning effect was equivalent between blocks. These results were interpreted as evidence that listeners had used only the most recent statistical experience to guide phonetic decisions, suggesting that talker adaptation cannot be characterized by cumulative tracking of statistical input, contrary to the predictions of the Bayesian belief-updating model of speech adaptation.

Here we provide an additional test of local versus global tracking accounts using the distributional learning paradigm of Clayards et al. (2008). This paradigm was selected because lexical information is not required to resolve ambiguity in the input and predictions for local versus global statistical tracking can be quantitatively derived. Listeners completed two blocks of phonetic identification where a single probabilistic cue to the stop voicing contrast, VOT, was manipulated. Each block contained two VOT distributions (Fig. 1a). Variance was manipulated across blocks to be either narrow or wide, and block order was manipulated between two groups. Thus, local statistics differed between the groups in each block, and global statistical experience was equivalent between groups at the end of the exposure period. Predictions (Fig. 1b) were generated using a modified version of Bayes theorem (Clayards et al., 2008) shown in equation (1). The prediction for the local statistics hypothesis considered input within each block, whereas the prediction for the global statistics hypothesis considered the combined input, which is conceptually identical to Saltzman and Myers (2018).

$$p(k|VOT) = \frac{p(VOT|k)}{p(VOT|k) + p(VOT|g)} \quad (1)$$

This framework predicts that the slope of the identification function will differ between the two groups in block one. Predictions for the local versus global hypotheses are dissociated in block two. If listeners disregard previous experience with the talker (during block one) and instead use only local experience with the talker (provided in block two) to guide phonetic decisions, then the slope of the identification function will again differ between the two order groups. However, if phonetic decisions reflect global experience, then the slope of the identification functions will converge for the two groups in block two.

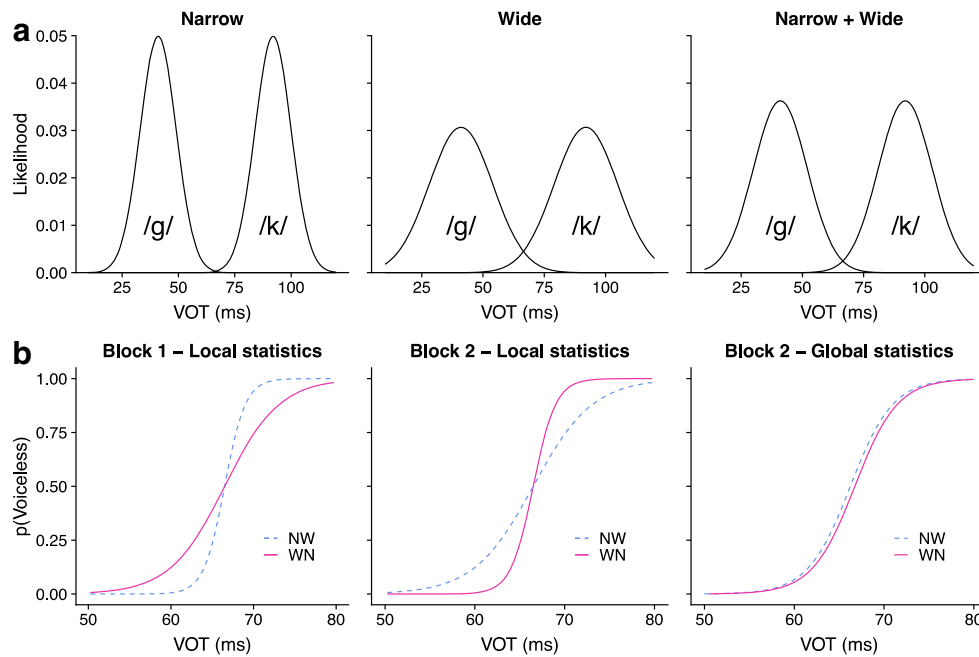


Fig. 1 Histograms of the input distributions and predicted identification functions for the local versus global tracking hypotheses. **Panel a** shows the input distributions for the narrow and wide blocks, and the distributions formed by aggregating distributions across the two blocks. **Panel b** shows the categorization functions predicted by equation (1) for each order group in block one (left), for the local statistics in block two

(middle), and for the global statistics in block two (right). The local statistics predictions were formed based on the input presented in each block; the global statistics predictions were formed considering the distributional information that was presented across the two blocks combined

Experiment 1

Method

Participants The participants ($n = 79$) were monolingual speakers of American English, 18–22 years of age; one additional participant was tested but excluded due to being bilingual. All passed a pure tone hearing screen administered at 25 dB for octave frequencies between 500 and 8,000 Hz. Participants were randomly assigned to either the narrow-wide order condition ($n = 39$) or the wide-narrow order condition ($n = 40$) and received either monetary compensation or partial course credit. All participants provided informed consent following procedures approved by the University of Connecticut Institutional Review Board.

Stimuli Stimuli were a subset of those used in Theodore and Miller (2010) and consisted of auditory tokens of *goal*, *coal*, *gain*, and *cane* produced by a female speaker that varied in word-initial VOT. Stimuli were drawn from two VOT continua, *goal-coal* and *gain-cane*. Each continuum was created using a naturally produced token as the voiced-initial endpoint (i.e., *goal*, *gain*). The LPC-based speech synthesizer in the ASL software package (Kay Elemetrics) was used to successively increase word-initial VOT in 4- to 5-ms increments, resulting in VOTs that perceptually ranged from /g/ to /k/. Twelve tokens from each continuum were selected for further

use; VOTs ranged from 11 ms to 119 ms in approximately 10-ms increments.

Tokens were arranged into two sets, one for the narrow block and one the wide block. Each set consisted of 236 tokens and contained equal numbers of each of the four words. Figure 1 shows histograms of the two sets. For the narrow set, mean VOT for the /g/ and /k/ distributions was 40 ms and 92 ms, respectively; the standard deviation for both distributions was 8 ms. Mean VOTs in the wide set were the same, but the standard deviation for both distributions was 13 ms.

Procedure Participants completed two blocks of phonetic categorization (472 trials in total), with block order determined by their experimental assignment. Testing took place in a sound-attenuated booth. Auditory stimuli were presented via headphones (Sony MDR-7506) at a comfortable listening level that was held constant across participants. Stimulus presentation and response collection were controlled using SuperLab 4.5 running on a Mac OS X system.

In each block, one randomization of the 236 tokens that formed the /g/ and /k/ distributions was presented. On each trial, participants identified each token as either *gain*, *cane*, *goal*, or *coal* by pressing the appropriately labeled button on a response box. Participants were instructed to make their decision as quickly as possible without sacrificing accuracy and to guess if they were unsure. The interstimulus interval

was 2,000 ms, timed from the participant's response. Participants were given a brief break between the two blocks. Prior to the start of the first block, participants completed 12 practice trials consisting of three repetitions of *gain*, *cane*, *goal*, and *coal* with VOTs matching the mean of the respective /g/ and /k/ distributions. The entire procedure lasted approximately 30 min.

Results

The raw data and analysis script can be retrieved at <https://osf.io/38h47>. Three participants were removed from the analysis, one due to failure to respond to 97% of the trials in one block and two because their identification responses did not show the expected categorical relationship between voiceless responses and VOT (due to responding *gain* and *goal* for most of the trials). Trial-level identification responses (0 = voiced, 1 = voiceless) were fit to a generalized linear mixed-effects model (GLMM) with the binomial response family using the `glmer()` function from the `lme4` package in R. The model specified VOT, order, block, and their interactions as fixed effects; the random effects structure specified random intercepts by subject and random slopes by subject for VOT and block. VOT was centered around the mean; order and block were contrast-coded (narrow-wide = -0.5, wide-narrow = 0.5; block 1 = -0.5, block 2 = 0.5). The results of the full model are shown in Table 1, which revealed an interaction between VOT, order, and block ($\beta = 0.554$, $SE = 0.154$, $z = 3.581$, $p < 0.001$).

The three-way interaction is visualized in Fig. 2a. The slope of VOT as a predictor of voiceless responses in each block for each order group was determined using simple slopes analysis with the `jtools` package in R (Fig. 2b). In block one, the narrow-wide group showed a steeper categorization slope compared to the wide-narrow group, but this difference was attenuated in block two. To determine whether the interaction reflected slope movement of only one group, two

follow-up models were constructed, one for each group, following the fixed and random-effects structure of the full model (removing order as a fixed effect). The interaction between VOT and block was significant for the narrow-wide group ($\beta = -0.375$, $SE = 0.124$, $z = -3.025$, $p = 0.002$), with the direction of the beta estimate indicating that the VOT slope decreased from block one to block two. No interaction between VOT and block was observed for the wide-narrow group ($\beta = 0.061$, $SE = 0.111$, $z = 0.550$, $p = 0.582$).

A parallel analysis was performed using trial number (centered around the mean) as the measure of time (instead of block) to examine whether the same pattern would be observed on a finer-grained scale (Table 2). There was a significant interaction between VOT, order, and trial number ($\beta = 0.279$, $SE = 0.078$, $z = 3.565$, $p < 0.001$). A simple slopes analysis was performed to extract the beta coefficient (and corresponding standard error) for the fixed effect of VOT for each order group at three trials: trial 200 (the previous 200 trials are statistically distinct between the two orders), trial 325 (cumulative statistics begin to merge between the two orders), and trial 450 (the previous 200 trials are distinct given local statistics, but near equivalent given cumulative statistics). The three-way interaction (Fig. 2c) emerges because the VOT slope is steeper for those in the narrow-wide order compared to the wide-narrow order at earlier but not later trials. As for the by-block analysis, follow-up models showed an interaction between VOT and trial number for the narrow-wide group ($\beta = -0.227$, $SE = 0.062$, $z = -3.678$, $p < 0.001$) but not the wide-narrow group ($\beta = -0.002$, $SE = 0.055$, $z = -0.043$, $p = 0.966$). Full statistics for all follow-up models are provided in the [Supplementary Material](#).

Experiment 2

The results of Experiment 1 are consistent with predictions for the global statistics hypothesis in terms of the between-group difference over time. However, the between-group pattern reflected by-trial movement of only the narrow-wide group, contrary to the predictions shown in Fig. 1, suggesting that the implementation of Bayes' rule is not sufficient to characterize performance. The framework outlined in Kleinschmidt and Jaeger (2015) provides an alternative. Their model predicts that exposure in block one will be weighted with respect to overall experience with English VOT distributions, and that exposure in block two will be weighted with respect to exposure in the first block. We performed computational simulations to test whether the Bayesian belief-updating model of speech adaptation would predict the asymmetry in learning as a function of block order.

Table 1 Results of the generalized linear mixed-effects model for Experiment 1, with time measured by the fixed effect of Block. The model contained 35,797 observations totaled across 76 participants. All test statistics represent those reported by the `glmer()` function

Fixed effects	β	SE	z	p
Intercept	-0.556	0.094	-5.884	<0.001
VOT	4.446	0.160	27.800	<0.001
Block	-0.125	0.075	-1.659	0.097
Order	-0.076	0.188	-0.406	0.685
VOT * Block	-0.164	0.083	-1.976	0.048
VOT * Order	-0.390	0.314	-1.242	0.214
Block * Order	-0.144	0.146	-0.985	0.325
VOT * Block * Order	0.554	0.155	3.581	<0.001

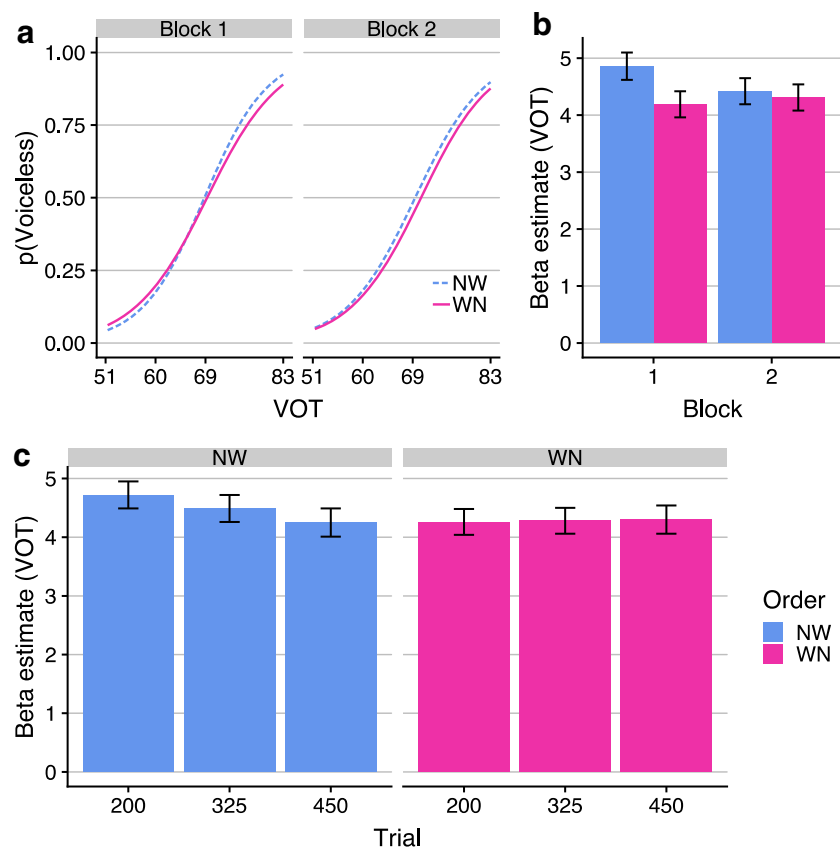


Fig. 2 **Panel a** shows the predicted effect of voice-onset-time (VOT) on voiceless responses in each block for the narrow-wide (NW) and wide-narrow (WN) order groups in terms of the fixed-effects of the generalized linear mixed-effects model (GLMM) described in the main text. To promote visualization, the abscissa range spans the four most intermediate

VOTs of the input distributions. **Panel b** shows the simple slope (beta estimate) for VOT in each block for each order group; error bars indicate the standard error of the beta estimate. **Panel c** shows the simple slope (beta estimate) for VOT at trials 200, 325, and 450 for each order group; error bars indicate the standard error of the beta estimate

Method

Simulations were implemented using the *beliefupdatr* package in R (Kleinschmidt, 2017). Parameter specification consists of prior distributions in terms of mean and variance of two categories (/g/ and /k/), and a confidence

Table 2 Results of the generalized linear mixed-effects model for Experiment 1, with time measured by the fixed effect of Trial Number. The model contained 35,797 observations totaled across 76 participants. All test statistics represent those reported by the *glmer()* function

Fixed effects	β	SE	z	p
Intercept	-0.557	0.095	-5.875	<0.001
VOT	4.460	0.160	27.905	<0.001
Trial number	-0.064	0.042	-1.526	0.127
Order	-0.073	0.189	-0.387	0.699
VOT * Trial number	-0.117	0.041	-2.826	0.005
VOT * Order	-0.385	0.315	-1.222	0.222
Trial number * Order	-0.042	0.082	-0.510	0.610
VOT * Trial number * Order	0.279	0.078	3.565	<0.001

parameter that estimates number of direct observations of the prior specification. The model input is trial-by-trial observations of the perceptual parameter (VOT) and the response category (/g/ or /k/). With this input, the learning algorithm (explicated in Kleinschmidt & Jaeger, 2015) updates the category-specific distributions on each trial by integrating the observed VOT and response with the prior distribution, weighted by confidence. The output is the posterior distribution on each trial, reflecting the likelihood of the prior distribution (formed by global experience with English) given the observed evidence (from the specific talker). The algorithm is iterative at each trial, and thus simulates how beliefs in priors change across trials.

Our simulation procedure was as follows. First, we simulated 80 lists specifying trial-level VOT presentation for 472 trials. Forty lists simulated trial-level VOT presentation for the narrow-wide group; the first 236 trials were a unique randomization of VOTs presented during the narrow block, and the second 236 trials were a unique randomization of VOTs presented during the wide block. Forty lists simulated trial-level VOT presentation for the wide-narrow group, randomizing trial-level VOTs in the reverse order. Response patterns for

the 80 simulated listeners matched the intended category for all VOTs except the four most intermediate. For the two most intermediate VOTs (60 ms, 69 ms), a random 50% of the responses were set to match the opposite category. For the next two intermediate VOTs (51 ms, 83 ms), a random 25% of the responses were set to match the opposite category. This procedure added a degree of response noise to simulate the imperfect categorization of midpoint stimuli observed in the behavioral test.

Second, nine simulations were performed for these lists, representing three prior specifications crossed with three confidence specifications. For all prior specifications, the standard deviation of priors was set to match that of a “typical talker” ($SD = 8.3$ for /g/ and 18.9 for /k/, Kleinschmidt & Jaeger, 2016). Across the three prior specifications (Fig. 3), means for /g/ and /k/ were set to be consistent with those presented in Experiment 1, shifted down ~ 10 ms, or shifted up ~ 10 ms (further information on prior specification is available in the [Supplementary Material](#)). For each specification, confidence was set to 200, 400, and 800, values that represent relatively less to relatively more confidence in the prior specification, respectively, spanning the range of inferred confidence reported previously (Kleinschmidt & Jaeger, 2016). Third, for each simulation, predicted categorization slopes were calculated for each simulated participant based on the inferred posterior distribution at trials 200, 325, and 450, by first calculating the

identification function for the inferred posterior distributions at these trials and then taking the derivative of the identification function at the category boundary.

Results

The analysis scripts, including code to execute the simulations, can be retrieved at <https://osf.io/38h47>. Figure 3 shows the predicted slope in each order condition for the nine simulations. Three patterns can be observed. First, movement in the predicted slope across trials is attenuated as confidence in the prior specification increases. Second, in all simulations, the predicted slope differs between the two orders at the early trial (200), with a steeper slope for the narrow-wide compared to the wide-narrow order, but converges between the two orders at the late trial (450). Third, in all simulations, the magnitude of the change between the early and late trials shows an asymmetry for the two orders; there is greater change in slope across trials for the narrow-wide compared to the wide-narrow order, a pattern that mirrors the behavioral results shown in Fig. 2.

Discussion

According to distributional learning accounts, listeners generate probabilistic mappings to speech sounds that are

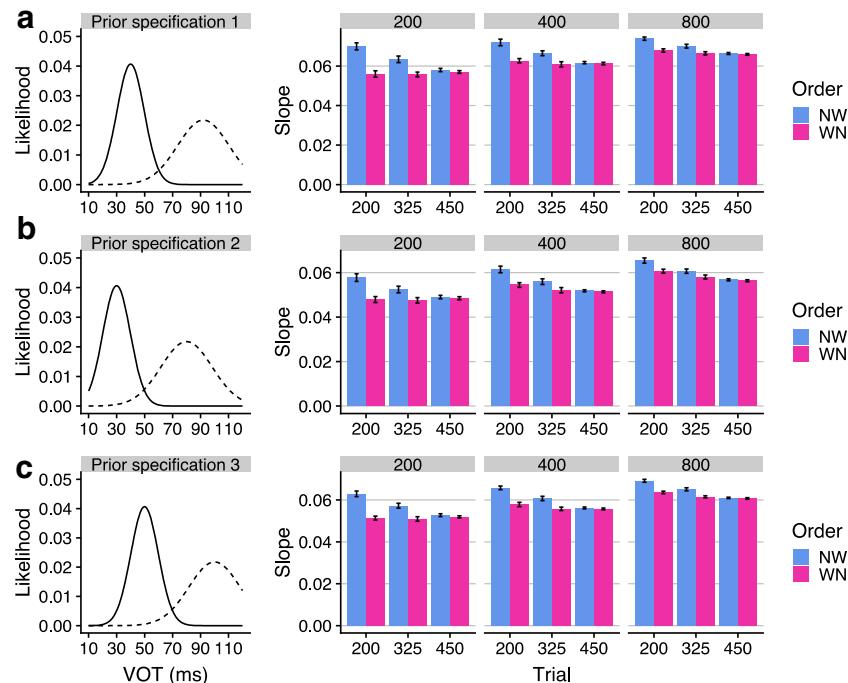


Fig. 3 Predicted categorization slopes from the computational simulations in Experiment 2. The three panels show simulation results for the three unique prior specifications (shown at left in each panel). The means of the distributions were manipulated across the prior specifications to be consistent with those presented in the behavioral test (/g/ = 40 ms, /k/ = 92 ms), shifted down ~ 10 ms (/g/ = 30 ms, /k/ =

80 ms), or shifted up ~ 10 ms (/g/ = 50 ms, /k/ = 100 ms). At right in each panel are the predicted slopes for the narrow-wide (NW) and wide-narrow (WN) order groups at three trials (trial 200, trial 325, and trial 450) for each of the three confidence parameters (200, 400, and 800). Error bars indicate standard deviation of the predicted slope for the 40 simulated listeners in each group

optimized for talker-specific input distributions (e.g., Kleinschmidt & Jaeger, 2015). The time course of experience used to guide optimization is underspecified in these accounts. In the current study, listeners who had initially experienced consistent input showed steeper identification functions compared to those who had initially received less consistent input. However, this difference became attenuated for later trials, when the cumulative statistical experience between groups was equivalent. The between-group convergence was driven by an asymmetry in learning between the two orders. The predictions derived solely with respect to the distributional statistics presented in the current study (Fig. 1b) did not adequately characterize the observed pattern of learning in Experiment 1. However, simulations (Fig. 3) in which the distributional statistics of the present talker were integrated with prior experience with English yielded the same qualitative pattern that was observed behaviorally.

These results are consistent with the theory that talker-specific phonetic adaptation reflects a process in which talker-specific input is used to modify language-general beliefs of cue distributions via a learning mechanism that cumulatively updates to accommodate the observed evidence. In this framework, the observed asymmetry in learning can be explained as the consequence of integrating the input distributions with prior knowledge and by-block exposure. VOTs in the wide distribution have a lower likelihood of occurrence compared to the narrow distribution. When a listener encounters VOTs from the wide distribution following the narrow distribution, there is greater prediction error, leading to greater slope movement. In the reverse order, listeners' beliefs from initial exposure (wide) are reaffirmed given secondary (narrow) exposure, as VOTs in the narrow distributions are present in the wide distributions, resulting in minimal prediction error and thus minimal change in beliefs.

The current results are consistent with findings showing cumulative influences of statistical experience on online perception (e.g., Holt, 2005; Idemaru & Holt, 2011; Kraljic & Samuel, 2005; Kraljic et al., 2008), and extend them to show experience-driven changes in the consistency in which a cue is used to guide categorization. We note that the current work examined but one of many types of distributional statistics that listeners may track in order to optimize the mapping to speech sounds, that being the variability – and hence reliability – of an acoustic-phonetic cue. Accommodating VOT variation in the current study did not require adjustment of cue-weighting (e.g., Idemaru & Holt, 2011), use lexical information as a learning signal (e.g., Kraljic et al., 2008), or require changing the boundary between phonetic categories to optimally accommodate the distributional manipulation (e.g., Saltzman & Myers, 2018). Additional investigations are

needed in order to examine whether the results reported here generalize to other acoustic-phonetic properties and to other types of distributional manipulations.

The current results do not converge with Saltzman and Myers (2018), who found that perceptual learning for a talker's productions was guided by local, not cumulative, statistical experience. Reconciling these findings is important for incorporating the time course of exposure into distributional learning accounts. Here we consider three possibilities. First, the disparate results may reflect the role of lexical information in guiding interpretation of acoustic input; distributional statistics of the input may be discarded when lexical information constrains online phonetic categorization. Second, statistical sensitivity for spectral versus temporal properties of speech may differ, to the extent that spectral properties are more informative of talker identity than temporal properties (Kleinschmidt, 2018). Third, the apparent discrepancy may reflect how predictions for local versus global accounts were derived. Recall that Kraljic et al. (2008) showed evidence of perceptual learning for listeners who were exposed to ambiguous and then clear productions, but not for listeners exposed to clear followed by ambiguous productions, suggesting that secondary exposure was considered in the global exposure context. In Saltzman and Myers, initial *and* secondary exposure were to non-canonical, acoustically ambiguous forms. It is possible that the robust learning observed for the secondary exposure occurred precisely because the initial exposure did not contain canonical productions, which would have blocked learning for the secondary exposure. This would result in learning that reflects cumulative exposure, but is not predicted solely by aggregated statistical experience. Future research is needed to expand distributional learning theories to account for factors that may block or reset the cumulative integration of statistical experience, as when initial experience deviates substantially from prior knowledge or may be considered incidental to a speaker's input (e.g., Kraljic et al., 2008).

To conclude, listeners show an exquisite ability to modify the mapping to speech sounds to accommodate systematic variation in the speech stream. The current results support a theoretical account in which talker-specific adaptation reflects the cumulative integration of experience with distributional statistics that are represented at both language-general and talker-specific levels. Theoretical and computational accounts of speech processing will be advanced by further specification of the time course in which experience is aggregated to guide perception, in addition to explicating the mechanisms by which higher-order structures (e.g., lexical knowledge, talker identity) may differentially influence sensitivity to and retention of distributional input. Future work is aimed at this goal.

Acknowledgements This work was supported by NIH NIDCD grant R21DC016141 to RMT and by the Acoustical Society of America Raymond H. Stetson Scholarship in Phonetics and Speech Science to NRM. The views expressed here reflect those of the authors and not the NIH or the NIDCD. We express gratitude to Stephen Graham for his assistance with data collection and to Emily Myers for fruitful discussion and feedback on this manuscript.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108, 804–809.
- Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97(5), 3099–3111.
- Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, 16(4), 305–312.
- Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, 37(6), 1939–1956.
- Kleinschmidt, D. F. (2017). beliefupdatr: Belief updating for phonetic adaptation. R package version 0.0.3.
- Kleinschmidt, D. F. (2018). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience*. <https://doi.org/10.1080/23273798.2018.1500698>
- Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122, 148–203.
- Kleinschmidt, D. F., & Jaeger, T. F. (2016). What do you expect from an unfamiliar talker? In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 2351–2356). Austin, TX: Cognitive Science Society.
- Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, 51(2), 141–178.
- Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science*, 19(4), 332–338.
- McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, 12(3), 369–378.
- Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America*, 109, 1181–1196.
- Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47(2), 204–238.
- Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, 60(3), 355–376.
- Saltzman, D., & Myers, E. (2018). Listeners are maximally flexible in updating phonetic beliefs over time. *Psychonomic Bulletin & Review*, 25(2), 718–724.
- Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *Journal of the Acoustical Society of America*, 128, 2090–2099.
- Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *Journal of the Acoustical Society of America*, 125, 3974–3982.
- Theodore, R. M., Myers, E. B., & Lomibao, J. A. (2015). Talker-specific influences on phonetic category structure. *Journal of the Acoustical Society of America*, 138, 1068–1078.