Perceptual learning of multiple talkers requires additional exposure

Sahil Luthra[1,2], Hannah Mechtenberg[1], & Emily B. Myers[1,2,3]

[1] Department of Psychological Sciences, University of Connecticut

[2] The Connecticut Institute for the Brain and Cognitive Sciences

[3] Department of Speech, Language and Hearing Sciences, University of Connecticut

Author Note

This research was supported by NIH R01 DC013064 (E.B.M., PI), by NSF IGERT grant DGE-1144399, and by NSF Research Traineeship grant (NRT) IGE1747486. SL was supported by an NSF Graduate Research Fellowship. The authors thank the members of the Language and Brain Lab for their feedback throughout the project and Dr. Arthur Samuel for helpful feedback on a previous draft of this manuscript. All stimuli, data, and analysis scripts are publicly available at https://osf.io/bs7ja/.

Abstract

Because different talkers produce their speech sounds differently, listeners benefit from maintaining distinct *generative models* (sets of beliefs) about the correspondence between acoustic information and phonetic categories for different talkers. A robust literature on phonetic recalibration indicates that when listeners encounter a talker who produces their speech sounds idiosyncratically (e.g., a talker who produces their /s/ sound atypically), they can update their generative model for that talker. Such recalibration has been shown to occur in a relatively talker-specific way. Because listeners in ecological situations often meet several new talkers at once, the present study considered how the process of simultaneously updating two distinct generative models compares to updating one model at a time. Listeners were exposed to two talkers, one who produced /s/ atypically and one who produced /ʃ/ atypically. Critically, these talkers only produced these sounds in contexts where lexical information disambiguated the phoneme's identity (e.g., *epi_ode*, *flouri_ing*). When initial exposure to the two talkers was blocked by voice (Experiment 1), listeners recalibrated to these talkers after relatively little exposure to each talker (32 instances per talker, of which 16 contained ambiguous fricatives). However, when the talkers were intermixed during learning (Experiment 2), listeners required more exposure trials before they were able to adapt to the idiosyncratic productions of these talkers (64 instances per talker, of which 32 contained ambiguous fricatives). Results suggest that there is a perceptual cost to simultaneously updating multiple distinct generative models, potentially because listeners must first select which generative model to update.

Introduction

Listeners must contend with a tremendous amount of acoustic-phonetic variability as they attempt to interpret the speech signal. Much of this variability is attributable to individual differences between talkers (Peterson & Barney, 1952), whose productions may differ from each other due to physiological, dialectal and/or social reasons (Johnson, 2008). To accommodate this variability, listeners can condition their interpretation of the signal on contextual information (e.g., lexical knowledge; Ganong, 1980) as well as on their beliefs about how a particular talker speaks (Kleinschmidt, 2019). Critically, any acoustic variability that is unexplained by the context can serve as a learning signal to the listener, allowing individuals to update their *generative model* of the talker – that is, their set of beliefs about how that particular talker produces their speech sounds (Davis & Sohoglu, 2020; Kleinschmidt & Jaeger, 2015). For example, if a listener encounters an ambiguous speech sound (*?*) in the context of the word *diver_ity*, they can leverage lexical knowledge to infer that the intended phoneme was /s/. As a result, the listener can update their generative model of how that particular talker produces the /s/ sound, and consequently, they will be more likely to map a similar ambiguous sound from that talker to the /s/ category (Norris, McQueen, & Cutler, 2003). The particular case where listeners leverage lexical knowledge to guide perceptual learning for speech is often referred to as *lexically guided perceptual learning*. Recent work in this domain suggests that listeners continually update their generative models, with perceptual learning reflecting aggregated experience with a talker's voice (Saltzman & Myers, in press; Tzeng, Nygaard, & Theodore, in press).

Because perceptual learning for speech constitutes a potential mechanism for listeners to accommodate phonetic variability between talkers, it is important to consider the extent to which such learning is talker-specific. A lexically guided perceptual learning study by Eisner and

McQueen (2005) is informative in this regard. In that study, participants were initially exposed to a /s/-/f/ blend, with some subjects hearing this ambiguous sound in /s/-biased lexical frames and others hearing the ambiguous sound in /f/-biased frames. Subsequently, listeners categorized sounds on a continuum from [ɛs] to [ɛf]; note that in this test phase, lexical context no longer served as a cue to the identity of the fricative, as both [ɛs] and [ɛf] are nonwords. Critically, the fricatives encountered at test were either produced by the same talker who had produced the ambiguous sound during exposure or by a different talker. When the talker was consistent between exposure and test, listeners showed evidence of perceptual learning, categorizing the ambiguous test fricatives in line with their previous exposure. However, if the talker producing the fricative segment was different from the talker heard during exposure, then listeners did not show evidence of learning. Because learning did not generalize to a second talker, these data are consistent with the notion that perceptual learning constitutes a way for listeners to learn how a *particular* talker produces their speech sounds.

Nevertheless, perceptual learning for speech has been shown to generalize across talkers in some situations, as found, for instance, by Kraljic and Samuel (2005). In their study, listeners were first exposed to /s/-/ʃ/ blends in lexically-biased contexts. During a subsequent phonetic categorization phase, listeners categorized stimuli from two [asi]-[aʃi] continua, one from a male talker and one from a female talker. Critically, one of the talkers was the same talker that listeners had heard during exposure, and one was different. Across several experiments, the authors found that listeners generalized if they were exposed to the female talker and later tested on the male talker, but they did not generalize from the male talker to the female talker. Further analyses indicated that this asymmetry in generalization was likely driven by an asymmetry in acoustic properties of the stimuli. In particular, the spectral mean of the fricatives produced by the female

talker during exposure was similar to the spectral mean of the fricatives produced by the male talker during test; however, the spectral mean of the male talker's speech heard during exposure was different from the mean of the female talker's speech heard during test. In other words, listeners generalized from one talker's voice to another when the two talkers were similar acoustically, but not when the spectral means of the two talkers were distinct. These data suggest that listeners can learn about the characteristic way a talker produces their speech sounds, but in certain environments (e.g., acoustic similarity between the key tokens), listeners may nonetheless generalize what they have learned to other similar talkers. Thus, generative models may not be specific to a particular talker but rather may be applicable to an entire group of similar talkers. If true, this would make it difficult for listeners to learn the specific idiolects of talkers who had otherwise similar voices.

An ecologically-valid model of talker-specific phonetic learning would require the listener to simultaneously maintain and update different generative models for different talkers (or sets of talkers). The observation that perceptual learning for speech does not necessarily generalize across talkers is consistent with the proposal that listeners can simultaneously maintain talker-specific generative models, but it is not diagnostic of it—a lack of generalization across talkers would also be observed if listeners were quickly updating a single set of beliefs about how all individuals produce their speech sounds. Under this latter view, listeners might instead maintain a single generative model of how talkers produce their speech sounds, and they would update this model any time talkers switch. To adjudicate between these possibilities, Kraljic and Samuel (2007) examined whether individuals were able to simultaneously maintain differing sets of beliefs about two different talkers. During an initial exposure phase, listeners were exposed to a male talker and a female talker; for one talker, lexical information biased interpretation of an ambiguous fricative

toward /s/, and for the other talker, lexical information biased interpretation of the ambiguous sound toward /ʃ/. To maximize the possibility of learning both talkers, the authors blocked the exposure phase by voice, such that listeners heard all the exposure stimuli from one talker before hearing the exposure stimuli from the second talker. After being exposed to both talkers, participants completed a phonetic categorization task. Results indicated that listeners were able to maintain separate generative models for how the two talkers produced these fricative consonants. Overall, these results suggest that listeners can learn to map the same phonetic information (e.g., the phonetic features corresponding to the ambiguous fricative) onto different perceptual categories (/s/ or /ʃ/) depending on who the talker is, at least for some types of speech sounds.

The studies discussed thus far show that listeners who are exposed to one talker at a time in an unbroken block can maintain separate phonetic mappings for two different talkers. Yet in natural conversation, talkers will alternate in rapid succession. How might listeners learn the differing phonetic characteristics of, for example, Julie Andrews' RP British accent in "Mary Poppins" compared to her castmate Dick Van Dyke's faux Cockney accent, when these voices alternate in dialogue? Should we expect similar degrees of perceptual learning when listeners are tasked with learning two talkers encountered in a blocked fashion (i.e., sequentially) versus when the two talkers are encountered in an interleaved manner (i.e., simultaneously)? In other words,

how does updating one generative model at a time compare to the challenge of updating multiple generative models in rapid alternation?[1]

Domain-general theories of learning offer potential clues as to how perceptual learning might differ across these two situations. For instance, research on motor skill learning suggests when there is trial-to-trial variability during learning, individuals face a high degree of contextual interference, and this interference leads to poorer performance during the practice phase (Magill & Hall, 1990; Shea & Morgan, 1979). When applied to perceptual learning for speech, such a result might imply that interleaved exposure to multiple talkers will create a high degree of contextual interference, such that updating one generative model may interfere with the process of updating another. As such, listeners tasked with maintaining distinct generative models for two talkers may therefore require additional support if the two talkers are initially encountered in an interleaved manner as compared to a blocked one.

In the current set of experiments, we investigated how simultaneously updating two distinct generative models compares to the process of updating one set of beliefs at a time. In Experiment 1, listeners completed exposure and test blocks for one voice before completing exposure and test blocks for the other voice. Because all testing with the first talker is completed before listeners encounter the second talker, this design does not require listeners to simultaneously maintain multiple generative models; as such, any learning observed in this experiment represents a

[1] Here, we emphasize that the question of interest relates to variability in the *structure* of exposure (i.e., whether exposure to multiple talkers is blocked by talker or interleaved). This is distinct from investigations of how perceptual learning is affected by variability in the *stimuli* presented during exposure (e.g., whether stimuli are produced by a single talker or by multiple talkers; c.f., Bradlow & Bent, 2008, who compared how perceptual adaptation to foreign-accented speech differs when listeners are exposed to multiple accented talkers compared to a single accented talker). In the current study, we considered how the structure of the exposure phase (i.e., whether trials are blocked by talker) influences perceptual learning, while controlling for the degree of stimulus variability (i.e., listeners always heard two talkers over the course of the experiment).

theoretical upper limit on the amount of perceptual learning we should expect to see when listeners track two talkers simultaneously.

In Experiment 2, our goal was to describe the sufficient conditions for maintenance of two distinct talker models during interleaved learning. Critically, interleaved exposure to two talkers requires listeners to continually adjust which generative model they are updating every time there is a talker switch. Because this switch may pose additional cognitive demands, perceptual learning may not be observed as readily when talkers are interleaved as compared to when they are blocked. Motivated by findings from the category learning literature, we also manipulated two key factors that could theoretically boost simultaneous learning of two talkers. First, we manipulated overall amount of exposure to the talkers' voices—with the prediction that interleaved learning may be successful but require increased exposure. Second, we manipulated the presence or absence of explicit feedback during the training task, with the prediction that feedback might increase motivation and/or direct attention to relevant aspects of the signal for learning (in this case, the identity of the talker).

General Method

Stimuli

Stimuli were constructed and recorded for a previous study (Luthra, Magnuson, & Myers, in press) and were repurposed for the current study. Thirty-two total words, 16 with a "medial" /s/ and 16 with a "medial" /ʃ/, with the same parameters as Kraljic and Samuel (2005), were created for the exposure phase of the experiment. The sets of medial /s/ and /ʃ/ words were equated on frequency (Kučera & Francis, 1967, $t(28) = 1.2$, $p = 0.24$), number of syllables prior to the medial fricative ($t(30) = -1.23$, $p = 0.23$) and total syllable number ($t(30) = 0.46$, $p = 0.65$).

Table 1. Words presented during the exposure phase

/s/-biased words		**/ʃ/-biased words**	
absent	accent	adoption	brochure
answer	Arkansas	definition	efficient
colosseum	currency	friendship	graduation
dinosaur	diversity	handshake	impatient
episode	eraser	invitation	ocean
insane	parasite	parachute	pediatrician
peninsula	pregnancy	permission	pressure
receipt	rehearsal	professional	vacation

A female speaker of North American English produced the lexical (e.g., *colosseum*) and non-lexical counterpart (e.g., *colosheum*) for each token (see Table 1 for full list). Recording occurred in a sound-isolated booth with a RØDE NT-1 condenser microphone and a Focusrite Scarlet 6i6 digital audio interface. Each token was produced twice, and the first author chose the best production of the pair. All recordings were passed though the native noise reduction filter in Audacity (http://audacityteam.org/). For further details about stimuli creation, please see Luthra et al. (in press).

These tokens were normalized to an amplitude of 70dB SPL. Then, an 11-step continuum between the recorded lexical and non-lexical tokens was generated (e.g., *colosseum – colosheum*) using STRAIGHT (Kawahara et al., 2008). The STRAIGHT software supports auditory morphing between two chosen endpoints; each endpoint is aligned temporally and spectrally before

interpolation. The medial fricative at step 7 was judged to be sufficiently ambiguous between /s/ and /ʃ/. From the original 11-step continuum, step 4 was chosen as the clear /s/ token (e.g., *colosseum*) while step 10 served as the clear /ʃ/ token (e.g., *colosheum*). Thus, all presented tokens were taken from the morphed recordings.

Finally, to generate a set of male-spoken tokens, the finalized female tokens were passed through the "Change Gender" tool in Praat (Boersma & Weenik, 2017). Transformation parameters included setting the formant shift ratio to 0.8 and the new pitch median to 100 Hz. All other parameters were held at default values. In this way, we were able to match the acoustic details of the male tokens as closely as possible to the female tokens. The subjective judgement of the authors and an informal survey of people familiar with the original talker indicated that this manipulation yielded a convincing shift in both gender (male from female voice) and talker (the original female talker was unidentifiable after the Change Gender tool was applied).

For the phonetic categorization task stimuli, a 7-step continuum was constructed following the same procedure as for the exposure stimuli. The same female speaker recorded clear productions of *sign* and *shine*, which were then subject to interpolation with STRAIGHT. We took steps 4 through 10 for the 7-step continuum, with step 4 as the clear production of *sign*, step 7 as ambiguous between *sign* and *shine,* and step 10 as an unambiguous *shine*. These morphed female-spoken tokens were also transformed into a male-spoken continuum using the same parameters described above using the "Change Gender" tool in Praat (Boersma & Weenik, 2017).

Procedure

Experiments were designed and hosted using the Gorilla Software Builder (www.gorilla.sc). Participants were recruited using Prolific (www.prolific.co). Participation

screening parameters (as implemented through Prolific) were as follows: between 18 and 90 years of age, normal or corrected-to-normal vision and hearing, current country of residence was the United States, English-speaking monolingual, no language-related disorders, approval rating in Prolific above 90 (max 100), and use of a desktop computer. Participation in one experiment also disqualified them from entrance into all subsequent experiments. All participants indicated their consent via digital information sheet per the guidelines of the University of Connecticut's Institutional Review Board. Payment was set at $3.33, as completion of the study took approximately 20 minutes (rate of $10 per hour).

After providing informed consent, participants completed a screener task to ensure they were wearing headphones (Woods, Siegel, Traer, & McDermott, 2017). Performance typically differs on this task when participants wear headphones compared to when listening over loudspeakers due to the nature of the stimuli. Participants who failed the screener were given a second chance to pass. Following the screening task, participants completed a short demographics questionnaire that asked about regional accent, sex, ethnicity, and race.

The experiment proper had two distinct phases—an exposure phase and a phonetic categorization phase (see Figure 1A for schematic). All experiments (1 and 2A-D) utilized this two-part design—differing only in the structure of the first exposure phrase. During the exposure phase (see Figure 1B), participants listened to the male and female talkers produce 32 tokens each (16 ambiguous and 16 unambiguous) while looking at a fixation cross in the center of the screen. Lexical context biased the interpretation of the ambiguous stimuli, with each talker having a distinct bias—the specific bias direction of each talker was counterbalanced across participants. For example, some participants heard the male talker produce ambiguous tokens in an /ʃ/-biased context (e.g., *ambi_?_ion*) and unambiguous /s/ tokens (e.g., *Arkansas*), while the female talker

produced the ambiguous tokens in an /s/-biased context (e.g., *colo_?_eum)* and unambiguous /ʃ/ tokens (e.g., *friendship*). As a cover task to keep participants' attention on the auditory stimuli, we instructed participants to indicate the talker's sex via keyboard press after each production. The "s" key corresponded to the left label (e.g., MALE) while the "k" key corresponded to the right label (e.g., FEMALE). Labels remained on the screen for 4000 ms, after which (in the absence of a response), the next stimuli played. We counterbalanced the position of the male and female labels (right or left) across participants. The structure of the exposure phase was altered in each experiment, but the content (i.e., stimuli, talker decision) did not change across experiments. As discussed further below, this means that in Experiment 1 (where trials were blocked by talker), listeners simply pressed the same button throughout the entire exposure phase.

After the exposure phase, participants completed a phonetic categorization task with tokens spoken by the same talkers as the previous exposure phase (see Figure 1C). Participants heard the seven randomized steps of each *sign-shine* continuum ten times each, resulting in a total of 70 trials for a given talker. The phonetic categorization phase was grouped by talker, such that participants heard all 70 tokens from one talker, and then, in a separate block after a short break, heard the 70 tokens from the other talker. The order of talker blocks (MALE/FEMALE, FEMALE/MALE) was counterbalanced across participants. After each production, participants indicated whether each production sounded more like *sign,* by pressing the key corresponding to an image of a street sign, or more like *shine*, with a key press corresponding to an image of a sun. The position of the sign and shine images on the screen were counterbalanced across participants. The order of stimuli within a block was also randomized. The structure and content of the phonetic

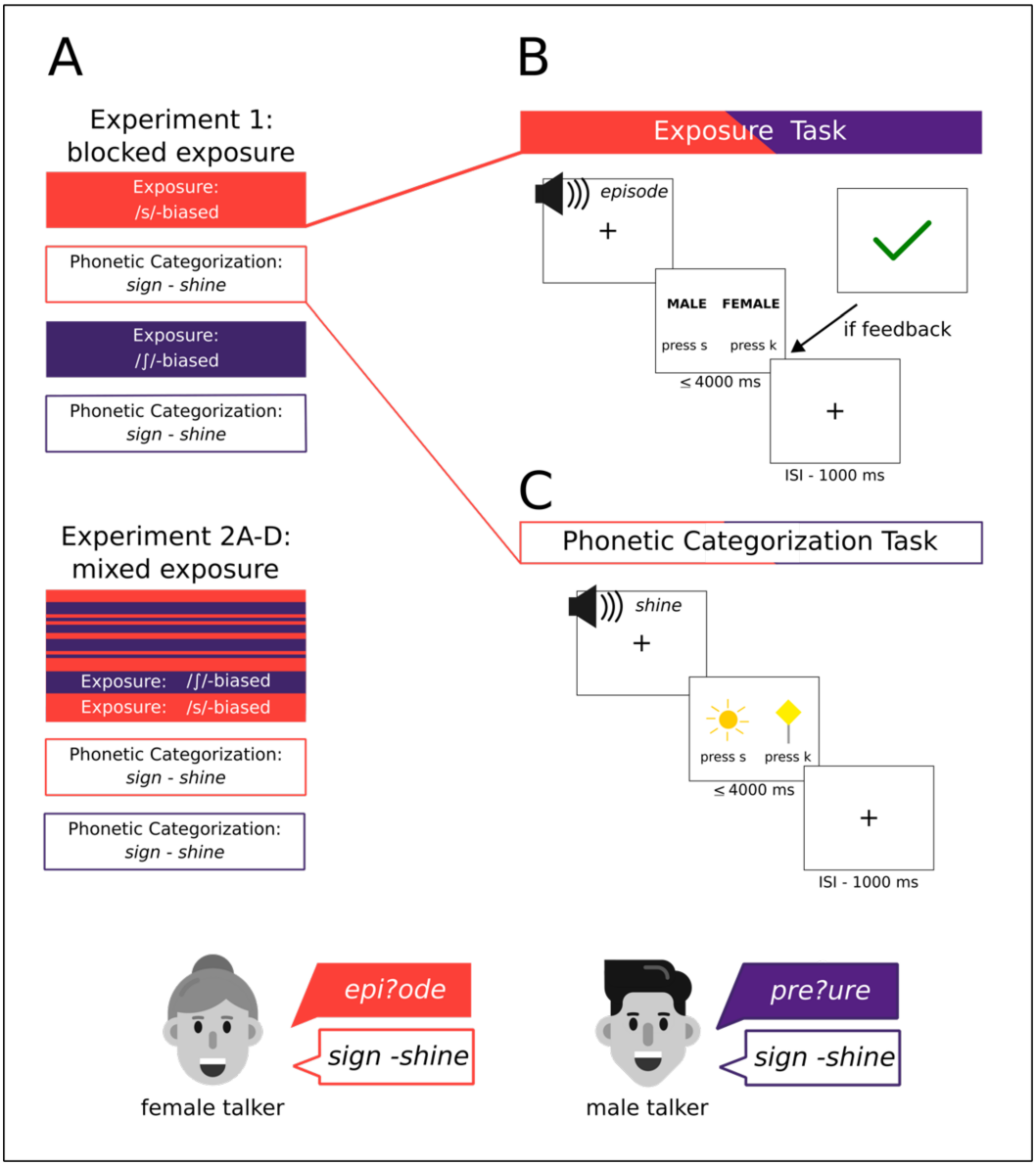

Figure 1. (A) General schematic for experiment task structure. For Experiment 1 (blocked), listeners were exposed to one talker and then immediately completed a phonetic categorization task for the same talker; this process was then repeated for the second talker. For Experiment 2 (mixed), listeners were exposed to the two talkers in an intermixed fashion. This exposure phase was followed by two blocks of phonetic categorization (grouped by talker). (B) Exposure task schematic. Participants first listened to the entire word, and were then allotted 4000 ms to complete the talker decision (indicate whether the word was spoken by the male or female talker). If feedback was a part of the experiment (Experiments 2B and 2D), then feedback followed immediately after the key press. After the key press and/or feedback display, there was an inter-stimulus interval of 1000 ms. (C) Phonetic categorization task schematic. Participants listened to a token from a 7-step continuum from *sign* to *shine*. They then indicated via key press whether the word sounded more like *sign* (sign image) or more like *shine* (sun image). There was no feedback. After the keypress, there was an ISI or 1000ms.

categorization task was held constant across all experiments, with only a small adjustment in Experiment 1 (detailed below). After completing the experiment, participants were debriefed and compensated through Prolific.

We excluded data from participants if they: failed the headphone screener twice, did not respond to more than 10% of trials during the exposure or phonetic categorization phase, and/or were less than 70% accurate in classifying the unambiguous endpoints during the phonetic categorization task (following Kleinschmidt & Jaeger, 2015; Luthra et al., in press). Additionally, the number of participants was not always equated across counterbalancing conditions; while Gorilla attempts to distribute participants evenly across conditions, the fact that not every person completed the experiment occasionally led to imbalanced counterbalancing. In those cases, we randomly excluded some participants prior to data analysis, allowing us to have an equal number of participants in each counterbalancing condition.

Data Analysis

We analyzed the phonetic categorization data for all experiments using mixed effects models implemented through R (R Core Team, 2019) with the *mixed* function in the "afex" package (Singmann, Bolker, Westfall, & Aust, 2018). This function is a wrapper to the *glmer* function in the "lme4" package (Bates, Maechler, Bolker, & Walker, 2015) and reports results in an ANOVA-like format, with chi-square tests determining significance. Fixed factors included Step (seven steps: *sign-shine*, centered using the *scale* function) and Bias (deviation coded [-1, 1], /ʃ/-bias, /s/-bias). Following the recommendation of Matuschek, Kliegl, Vasishth, Baayen, and Bates (2017), we employed a backward-stepping procedure to identify the simplest random effect structure without sacrificing model fit. Specifically, we started with the maximal random effects

structure (random by-subject slopes and intercepts for Step and Bias and their interaction) and iteratively compared to a simplified model (first eliminating the random by-subject interaction of Step and Bias, then random by-subject slopes for Step) using the *anova* function. If the more complex model was a significantly better fit than the simpler model, we selected the more complex model. If there was no difference in fit between either model, we continued the stepping procedure. The final model syntax for each experiment is reported in Tables 2 and 3.

Experiment 1: Blocked Talker Exposure

In Experiment 1, we examined listeners' ability to adapt to two talkers who differed in how they produced the fricatives /s/ and /ʃ/ when exposed and tested in sequential fashion. Critically, listeners were exposed to and tested on one talker's voice before hearing the second talker; that is, they were only tasked with updating one generative model at a time. To maximize the likelihood that listeners would adopt different generative models for the two talkers they heard in the current study, we required listeners to make explicit decisions about talker identity during the initial exposure phase—note that because exposure in this experiment was blocked by talker, listeners simply had to press the same button repeatedly for each exposure phase. However, previous work has suggested that listeners may be more likely to show talker-specific effects when they actively attend to talker identity during encoding of the talkers' voices (Goldinger, 1996; Luthra, Fox, & Blumstein, 2018; Theodore, Blumstein, & Luthra, 2015). Notably, lexically guided perceptual learning has been shown to occur robustly following a variety of exposure tasks (Clarke-Davidson, Luce, & Sawusch, 2008; Drouin & Theodore, 2018; Eisner & McQueen, 2006; Leach & Samuel, 2007; Luthra et al., in press; Maye, Aslin, & Tanenhaus, 2008; McQueen, Norris, & Cutler, 2006;

White & Aslin, 2011), though to our knowledge, no previous studies have used a talker identification task specifically.

METHOD

Procedure

Participants completed the exposure and test block for a given talker (e.g., MALE exposure, MALE phonetic categorization) before completing both blocks for the other talker (e.g., FEMALE exposure, FEMALE phonetic categorization). The talker order (MALE-FEMALE / FEMALE-MALE) was counterbalanced across participants.

Participants

We recruited 52 participants for Experiment 1. Based on the exclusion criteria established above, we excluded 13 participants for data quality issues and 7 for failing the headphone screener, leaving 32 participants (22 female, 10 male) for analyses. Age of participants in this final sample ranged from 19 to 68 (mean: 34).

RESULTS

The phonetic categorization results of Experiment 1 are shown in Figure 2A. Phonetic categorization for the /s/-biased talker is shown in purple while phonetic categorization for the /ʃ/-biased talker is in red. Evidence for talker-specific phonetic recalibration is indicated by separation between the red and purple lines for the middle (more ambiguous) steps. For Experiment 1, there is a clear difference between the percent shine responses (0-100%, along y-axis) for the /s/ versus the /ʃ/-biased talkers for those middle steps.

Results of statistical analyses are shown in Table 2. There were significant main effects of Step ($p < 0.001$) and of Bias ($p < 0.01$), and no interaction of Step and Bias ($p = 0.76$). A main effect of Step was anticipated—participants rated tokens as being more "shine-like" as the continuum went from "*sign*" to "*shine*." The effect of Bias indicates that participants were more likely to judge ambiguous tokens along the *sign-shine* as being more "shine-like" following exposure to the /ʃ/-biased talker, compared to the /s/-biased talker; thus, the effect of Bias provides evidence of perceptual learning. While the separation between the /ʃ/-biased and /s/-biased categorization functions is particularly clear in the middle of the continuum (especially compared to the continuum endpoints), we did not observe a statistically significant interaction between Step and Bias—it is possible that our analysis was underpowered in its ability to detect an interaction, so we encourage caution in overinterpreting the lack of an interaction. Note also that the effect of greatest theoretical interest is the main effect of Bias, not the interaction between Bias and Step.

Table 2 Phonetic categorization results from Experiment 1 (blocked talker exposure)

SH_resp ~ step*bias + (step * bias | Subject) + (step | Subject) + (bias | Subject)

Fixed Effects	**df**	**chi square**	**p value**	**significance**
Step	1	61.53	< 0.001	***
Bias	1	9.35	0.002	**
Step x Bias	1	0.09	0.76	

Model syntax on line 2. Significance is indicated by asterisk number adhering to classic convention ($p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), $p > 0.05$ (+).

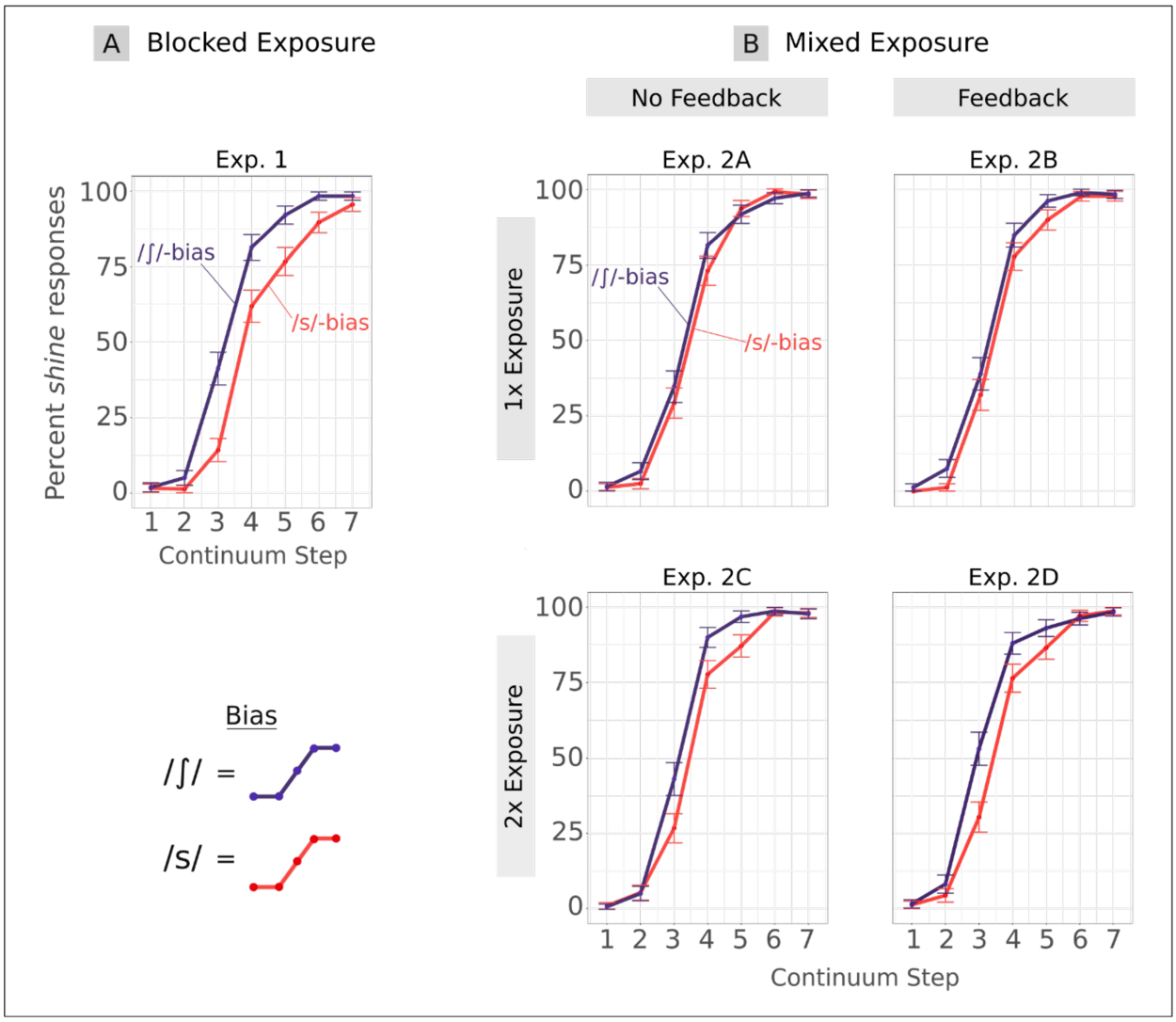

Figure 2. Results from the phonetic categorization task from Experiment 1 and 2A-D. Purple lines indicate categorization of the /ʃ/-biased talker while the red lines are for the /s/-biased talker. Listed along the x-axis are each step along a continuum from a clear "*sign*" production (step 1) and a clear "*shine*" production (step 7). Percent "*shine*" responses are indicated along the y-axis, from low to high. (A) Blocked talker exposure (Exp 1). (B) Experiment 2. Upper left: mixed talker exposure, low-exposure, no feedback. Upper right: mixed talker exposure, low-exposure, with feedback. Lower left: mixed talker exposure, high-exposure, no feedback. Lower right: mixed talker exposure, high exposure, with feedback. Error bars represent a 95% confidence interval.

Experiment 1 Discussion

Results of Experiment 1 indicated that participants were able to adapt to the idiosyncratic fricative productions of two talkers who were encountered sequentially. These data are consistent with previous lexically guided perceptual learning studies and most directly relate to the work of Kraljic and Samuel (2007), who also tested listeners' ability to update distinct generative models for two different talkers. In both studies, an ambiguous fricative /*?*/ corresponded to /s/ for one talker and to /ʃ/ for the other. However, the present experiment differs from the previous work of Kraljic and Samuel in at least two key ways.

First, Experiment 1 used a talker identification task during exposure, and participants only heard real words (no nonwords) during exposure blocks; this is in contrast to seminal lexically guided perceptual learning studies (e.g., Kraljic & Samuel, 2007; Norris et al., 2003), which have traditionally used a lexical decision task instead. The effectiveness of a talker identification task is striking particularly because in the current experiment, talker identity was blocked at exposure, such that participants were simply required to press the same button throughout each exposure block (i.e., to press the "male" button for the entirety of one exposure block and the "female" button for the entirety of the other). As such, the present results build on previous work showing that lexically guided perceptual learning can be elicited by relatively shallow exposure tasks that do not require participants to make explicit lexical judgments (Drouin & Theodore, 2018; Eisner & McQueen, 2006; Maye et al., 2008; White & Aslin, 2011).

Second, participants in the study by Kraljic and Samuel (2007) were exposed to both talkers before completing the first phonetic categorization block. While previous work suggests that participants are able to maintain talker-specific generative models (Eisner & McQueen, 2005), some degree of generalization from one talker to another has been observed, even with fricative

sounds (Kraljic & Samuel, 2005). It is conceivable that in the study by Kraljic and Samuel (2007), being exposed to both talkers prior to any testing may have introduced some degree of interference, potentially attenuating the amount of learning that would be observed for the two talkers. In contrast, Experiment 1 of the current study was designed such that listeners completed both exposure and test phases with one voice before hearing the second voice; under this design, there should be minimal interference from one generative model to the other. The degree of learning observed in Experiment 1 therefore constitutes a theoretical upper limit on the amount of learning that should be expected in Experiment 2, where there was a potential for interference between the two generative models.

Experiment 2: Mixed Talker Exposure

In Experiment 2, we examined how the challenge of simultaneously updating two generative models compares to the process of updating two generative models sequentially. To do so, we modified the design of Experiment 1 such that listeners were exposed to the two talkers in an interleaved fashion, with talker identity varying randomly from trial to trial. In a 2x2 between-subjects design, we also parametrically manipulated the amount of exposure during training and whether or not participants received feedback to titrate out the necessary conditions for multi-talker learning. These manipulations are described below.

Learning studies across multiple domains have suggested that when individuals encounter trial-by-trial variability during training, performance during training is hindered relative to when there is low variability (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Fuhrmeister & Myers, 2020; Magill & Hall, 1990; Shea & Morgan, 1979). We therefore hypothesized that in order to overcome the contextual interference introduced by interleaved exposure to two talkers,

individuals might need an increased number of trials relative to when exposure to the two talkers was blocked. As such, we manipulated whether participants in Experiment 2 received the same number of exposure trials as in Experiment 1 or twice as much exposure.

In addition to manipulating the degree of exposure to the two talkers, we also manipulated whether participants received feedback during the exposure task, as previous research has suggested that feedback may modulate the extent of speech category learning (e.g., tone category learning; Chandrasekaran, Koslov, & Maddox, 2014; Chandrasekaran, Yi, & Maddox, 2014). In particular, speech category learning is thought to be mediated by dual learning systems: an explicit, *reflective* system whereby individuals test whether members of two auditory categories can be distinguished by a verbalizable rule, and an implicit, *reflexive* system whereby categorization is achieved by integrating across stimulus dimensions and therefore members of two categories cannot be distinguished by an easily articulable rule (Chandrasekaran, Yi, et al., 2014). Individuals tend to rely more on the reflective system during the early stages of learning but eventually grow to rely on the reflexive system as they gain experience; this transition from reflective to reflexive processing is thought to be critical for speech category learning, since speech sound categories are not distinguished by unidimensional, easily articulable rules. These previous speech category learning studies have demonstrated that providing immediate feedback about response correctness encourages participants to rely more on the reflexive system, thus promoting speech category learning. Here, we hypothesized that providing listeners with immediate feedback about their correctness on the talker identification task used in the exposure phase might encourage them to employ optimal learning strategies, potentially promoting better learning of the specific ways that the two talkers produced their /s/ and /ʃ/ sounds. Thus, we had four groups for Experiment 2, orthogonally manipulating both the amount of exposure subjects received to each talker (32 trials

/ 64 trials) as well as whether subjects received feedback about their performance on the talker identification task during exposure.

METHOD

In Experiment 2, we intermixed the two talkers during the exposure phase (see Figure 1A, lower panel, for schematic). During exposure, participants heard each word once (with half of the words spoken by the male talker and the other half spoken by the female talker) before hearing any of those words spoken again (with each word spoken by the opposite talker from whom had said it previously). Following exposure, participants completed two phonetic categorization test blocks (blocked by talker, with the talker order counterbalanced across participants), as in Experiment 1. In Experiment 2, after completion of the exposure phase there were two blocks of phonetic categorization (grouped by talker), the order of which were counterbalanced across participants (e.g., MALE-FEMALE, FEMALE-MALE).

Exposure manipulation (Low Exposure: 2A, 2B; High Exposure: 2C, 2D)

During the exposure phase, we systematically manipulated the number of talker trials participants listened to. In the low exposure experiments (2A and 2B), participants heard 64 trials during the exposure phase (32 each, male and female). In the high exposure experiments (2C and 2D), we doubled the exposure trials to 128 by duplicating the low-exposure condition.

Feedback manipulation (No feedback: 2A, 2C; Feedback, 2B, 2D)

We also manipulated whether listeners received feedback on the talker identification task. Studies of non-native phonetic learning suggest that adding feedback during the task increases the

likelihood of learning phonetic detail (Chandrasekaran, Yi, & Maddox, 2014). Notably, in these cases, feedback is provided on the to-be-learned phonetic information (i.e., whether the talker produced one speech sound or another speech sound). In the current study, however, we provided feedback on the talker decision, based on previous work that talker-specific phonetic effects are more pronounced when listeners attend to talker information (e.g., Goldinger, 1996). A green check mark ("correct) or a red "X" ("incorrect") appeared on the screen after each talker decision. Feedback appeared immediately after a response, and remained on the screen for 1000 ms. Experiments that included feedback were Experiments 2B and 2D, and there was no feedback included for Experiments 2A and 2C.

Experiment 2A: Mixed Talker, 1X Exposure, No Feedback

This experiment consisted of low exposure (64 trials) and no talker-decision feedback. We recruited 51 participants from Prolific with the same exclusionary criteria described in experiment 1. Eight failed the headphone screener, five had poor data quality, and five were rejected to equalize counterbalancing conditions, and one was removed due to a technical error. After exclusions, a total of 32 participants (18 female, 14 male) were left for all analyses. Participants ranged in age from 20 to 66 (mean: 32).

RESULTS

Plotted results for the phonetic categorization task are in Figure 2B. Model output and syntax of the linear mixed effects regression can be seen in Table 3A. There was a main effect of Step ($p < 0.001$), no effect of Bias ($p = 0.20$), and no interaction of Step and Bias ($p = 0.98$). The lack of any main effect or interaction involving Bias indicates a lack of phonetic recalibration.

Experiment 2B: Mixed Talker, 1X Exposure, With Feedback

Participants heard 64 trials during exposure (low-exposure condition) and received feedback after the talker decision. Thirty-nine participants were recruited from Prolific. After applying the same data exclusion criteria as before (five failed the headphone screener, one had poor data quality, and one was removed to equate counterbalancing conditions), there were 32 participants (20 female, 12 male) remaining for analyses. Participants in the final sample ranged in age from 18 to 68 (mean: 32).

RESULTS

Figure 2B plots the results for Experiment 2B, and the results of the model are listed in Table 3B. We found a significant main effect of Step ($p < 0.001$), no effect of either Bias ($p = 0.91$) or an interaction ($p = 0.15$). These results indicate no effect of phonetic recalibration when feedback was introduced for the talker decision.

Experiment 2C: Mixed Talker, 2X Exposure, No Feedback

We presented a total of 128 trials (64 from each talker; high-exposure condition) and did not include feedback. Fifty participants were recruited from Prolific. Seven participants failed the headphone screener, three had poor data quality, and two more were excluded to equate counterbalancing conditions—leaving 32 (14 female, 18 male) for further analyses. These participants ranged in age from 18 to 61 (mean: 29).

RESULTS

Results are shown in in Figure 2B. There was an expected significant effect of Step ($p < 0.001$), a marginal effect of Bias ($p < 0.06$) and no interaction of Step and Bias ($p = 0.80$), as shown in Table 3C. The marginally significant effect of Bias suggests that, by doubling talker exposure, participants categorized the ambiguous stimuli between "sign" and "shine" differently depending on the bias of the talker during the exposure phase.

Experiment 2D: Mixed Talker, 2X Exposure, With Feedback

The marginally significant effect of Bias found in Experiment 2C suggests that doubling exposure may help listeners engage in phonetic recalibration processes. To extend this result, we added talker-decision feedback to the high-exposure condition. We recruited 44 participants from Prolific. Once participants were rejected for failing the headphone (seven participants) or the data quality checks (three for poor data quality and two to equalize counterbalancing conditions), we were left with 32 participants (16 female, 16 male) for all subsequent analyses. Participants ranged in age from 19 to 70 (mean: 32).

RESULTS

Plotted in Figure 2B are the results for Experiment 2D. Results of the linear mixed effects regression are given in Table 3D; in contrast to previous analyses, the optimal random effects structure did not include a random by-subject interaction for Step and Bias. There were significant effects of Step ($p < 0.001$), Bias ($p < 0.01$), and an interaction of Step and Bias ($p < 0.01$). These results indicate the presence of a phonetic recalibration effect that was specific to each talker. Additionally, the significant interaction of Step and Bias suggests that the effect of Bias was

inconsistent along the dimension of Step, likely reflecting a numerically greater effect of bias on ambiguous stimuli near the category boundary.

Experiment 2 Discussion

When participants received the same amount of exposure as in Experiment 1, but with intermixed training (Exp 2) rather than blocked training (Exp 1), we did not observe talker-specific learning, evidenced by the same pattern of phonetic categorization regardless of talker and a lack of Bias effect. However, participants who received twice as much exposure (64 trials per talker, of which 32 contained the ambiguous fricative) showed talker-specific perceptual learning. Taken together, these results suggest that when listeners are tasked with simultaneously updating two distinct generative models, additional exposure is required relative to when listeners need only update one model at a time.[2]

[2] We also ran an omnibus analysis for all mixed-exposure experiments (2A-2D) with Feedback (no/yes) and Exposure Number (1x/2x) as fixed factors. While we saw main effects of both Step ($\chi^2(1) = 283.65$, $p < 0.001$) and Bias ($\chi^2(1) = 7.77$, $p < 0.01$), there were no interactions between Bias and Feedback or between Bias and Exposure Number (Bias x Exposure Number: $\chi^2(1) = 2.48$, $p = 0.12$). We did not include this analysis in the results, as the goal of the current study was to assess the necessary conditions for multi-talker learning and not the size of the learning effect as a function of our manipulations.

Table 3 Phonetic categorization results from Experiment 2 (mixed talker exposure).

A. Experiment 2a: Mixed Exposure (1x), no Feedback

SH_resp ~ step*bias + (step * bias | Subject) + (step | Subject) + (bias | Subject)

Fixed Effects	df	chi square	p value	significance
Step	1	76.63	< 0.001	***
Bias	1	1.65	0.20	
Step x Bias	1	0.00	0.98	

B. Experiment 2b: Mixed Exposure (1x), with Feedback

SH_resp ~ step*bias + (step * bias | Subject) + (step | Subject) + (bias | Subject)

Fixed Effects	df	chi square	p value	significance
Step	1	70.08	< 0.001	***
Bias	1	0.01	0.91	
Step x Bias	1	2.12	0.15	

C. Experiment 2c: Mixed Exposure (2x), no Feedback

SH_resp ~ step*bias + (step * bias | Subject) + (step | Subject) + (bias | Subject)

Fixed Effects	df	chi square	p value	significance
Step	1	62.52	< 0.001	***
Bias	1	3.41	0.06	+
Step x Bias	1	0.06	0.80	

D. Experiment 2d: Mixed Exposure (2x), with Feedback

SH_resp ~ step*bias + (step | Subject) + (bias | Subject)

Fixed Effects	df	chi square	p value	significance
Step	1	93.64	< 0.001	***
Bias	1	7.38	0.007	**
Step x Bias	1	8.10	0.004	**

Model syntax on line below each experiment label. (A) Experiment 2a results, mixed talker presentation once through, no feedback for talker decision. (B) Experiment 2b results, mixed talker presentation once through, with feedback for talker decision. (C) Experiment 2c results, mixed talker presentation **twice** through, no feedback for talker decision. (D) Experiment 2d results, mixed talker presentation **twice** through, with feedback for talker decision. Significance is indicated by asterisk number adhering to classic convention ($p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), $p > 0.05$ (+).

General Discussion

Because of the considerable acoustic-phonetic variation between talkers, listeners must maintain different generative models (i.e., different sets of beliefs) for how different talkers (or different groups of talkers) produce their speech sounds (Kleinschmidt & Jaeger, 2015). A robust literature on lexically guided perceptual learning suggests that as listeners gain additional experience with a talker's voice, they continually update these generative models, doing so in a relatively talker-specific fashion (Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2007; Tzeng et al., in press). For instance, when a listener is exposed to a talker who produces the /s/ sound with a relatively low spectral center (so that it sounds closer to an /ʃ/), they will update their beliefs about how this talker (and other similar talkers) produce their /s/ sound. In ecological conditions, however, listeners rarely encounter one talker at a time—rather, listeners typically alternate between different talkers and therefore must simultaneously update multiple generative models, effectively assigning the correct acoustic distribution to the correct talker.

In the current lexically guided perceptual learning study, listeners were exposed to two talkers, one of whom produced their /s/ sound atypically and one of whom produced their /ʃ/ sound atypically. Thus, listeners were required to maintain separate generative models for these two talkers.[3] In Experiment 1, listeners were exposed to one talker's voice and tested on that talker prior to hearing the second one; that is, they only had to update one generative model at a time. Listeners were able to learn the idiosyncratic speaking styles of these two talkers with relatively little exposure (32 exposure trials per talker). In Experiment 2, listeners were exposed to the two talkers in an interleaved fashion, such that they were exposed to both talkers' voices before learning was assessed. We found that when exposure was interleaved, listeners required additional exposure to the talkers' voices (64 exposure trials per talker) before talker-specific learning was observed. In general, these results suggest that there is a cognitive cost associated with updating multiple generative models simultaneously, as opposed to sequentially.

Our results are consistent with domain-general theories of learning, which hold that trial-to-trial variability during learning induces contextual interference, making learning relatively challenging (Magill & Hall, 1990). In Experiment 1, listeners faced relatively little contextual interference, as they were exposed to and tested on one talker before being exposed to the second talker. By contrast, Experiment 2 required listeners to contend with a relatively high degree of

[3] The current data suggest that learning the phonetic contingencies of two interleaved talkers (Experiment 2) is more difficult than learning the idiolects of two blocked talkers (Experiment 1). However, the challenge of adapting to two interleaved talkers may have been exacerbated since the current study also required listeners to learn to interpret the *same* phonetic information (an ambiguous sound between /s/ and /ʃ/) differently depending on the talker. In principle, perceptual learning of interleaved talkers might be relatively easier if listeners were learning different contrasts (e.g., /s/-/ʃ/ for one talker and /b/-/v/ for the other), as there would be less conflict between talkers in how acoustics map onto phonetic categories. Future work will be necessary to clarify how the exposure schedule (blocked versus interleaved) may interact with between-talker "phonetic conflict" to affect perceptual learning.

contextual interference, as there was trial-to-trial variability in which generative model listeners needed to update. As such, additional experience was required before learning was observed.

Studies of learning suggest that while interleaved exposure may make learning more challenging, it may also lead to more robust learning, as a high degree of contextual interference may encourage learners to engage in deeper, more elaborate processing (Magill & Hall, 1990; Rohrer & Taylor, 2007). For instance, an interleaved exposure schedule may be advantageous for learning non-native speech sound categories by promoting the use of a reflexive learning system (Chandrasekaran, Koslov, et al., 2014). However, the benefits of interleaved training over blocked exposure may be mitigated when the contrast to be learned is particularly difficult (Fuhrmeister & Myers, 2020). The current study provides evidence that listeners need additional exposure to update talker-specific generative models for native-language phonetic categories when exposure to the two talkers is intermixed compared to when it is blocked—however, future studies would be needed to specifically investigate whether interleaved exposure leads to more robust learning of how different talkers produce their speech sounds.

Results of Experiment 2 indicated that feedback did not influence whether talker-specific learning was observed. We had hypothesized that drawing attention to the two talkers might make it easier for listeners to learn each talker's phonetic idiosyncrasies, as previous work has suggested that attention to talker identity at encoding may modulate the strength of talker-specific effects (Goldinger, 1996; Luthra et al., 2018; Theodore et al., 2015). However, the talker identification task used in this study was orthogonal to the phonetic manipulation (i.e., the information that allowed listeners to disentangle whether the ambiguous sound was /s/ or /ʃ/ came from the lexical signal, not from talker identity), which may explain why no effects of feedback were observed.

In summary, the present study demonstrates that listeners can track the distinct, idiosyncratic ways that different talkers produce their speech sounds, providing further evidence that perceptual learning may constitute a mechanism by which listeners learn to accommodate phonetic variability across talkers. Critically, our results indicate that listeners are able to update multiple generative models at a time (as when they encounter two distinct talkers in an interleaved fashion). However, listeners may require additional exposure when they are required to update two generative models simultaneously, as compared to the situation where they only need to update one model at a time.

Open Practices Statement

All stimuli, data, and analysis scripts are publicly available at https://osf.io/bs7ja/. The experiments reported in this manuscript were not pre-registered.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Boersma, P., & Weenik, D. (2017). Praat: Doing phonetics by computer.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

Chandrasekaran, B., Koslov, S. R., & Maddox, W. T. (2014). Toward a dual-learning systems model of speech category learning. *Frontiers in Psychology*, *5*(July), 1–17.

Chandrasekaran, B., Yi, H. G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin and Review*, *21*(2), 488–495.

Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception and Psychophysics*, *70*(4), 604–618.

Davis, M. H., & Sohoglu, E. (2020). Three functions of prediction error for Bayesian inference in speech perception. In D. Poeppel, G. R. Mangun, & M. S. Gazzaniga (Eds.), *The Cognitive Neurosciences* (6th ed., pp. 177–189). Cambridge, MA: The MIT Press.

Drouin, J. R., & Theodore, R. M. (2018). Lexically guided perceptual learning is robust to task-based changes in listening strategy. *Journal of the Acoustical Society of America*, *144*(2), 1089–1099.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, Supplement*, *14*(1), 4–58.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, *119*(4), 1950–1953.

Fuhrmeister, P., & Myers, E. B. (2020). Desirable and undesirable difficulties: Influences of variability, training schedule, and aptitude on nonnative phonetic learning. *Attention, Perception, and Psychophysics*, 1–17.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, *22*(5), 1166–1183.

Johnson, K. A. (2008). Speaker normalization in speech perception. *The Handbook of Speech Perception*, 363–389.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3933–3936).

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience*, *34*(1), 43–68.

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*(2), 148–203.

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. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Dartmouth Publishing Group.

Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement : When adults learn new words. *Cognitive Psychology*, *55*, 306–353.

Luthra, S., Fox, N. P., & Blumstein, S. E. (2018). Speaker information affects false recognition of unstudied lexical-semantic associates. *Attention, Perception, and Psychophysics*, *80*(4), 894–912.

Luthra, S., Magnuson, J. S., & Myers, E. B. (in press). Boosting lexical support does not enhance lexically guided perceptual learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Magill, R. A., & Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. *Human Movement Science*, *9*(3–5), 241–289.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543–562.

McQueen, J. M., Norris, D., & Cutler, A. (2006). The dynamic nature of speech perception. *Language and Speech*, *49*(1), 101–112.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481–498.

Saltzman, D. I., & Myers, E. B. (in press). Listeners are initially flexible in updating phonetic beliefs over time. *Psychonomic Bulletin & Review*.

Shea, J. B., & Morgan, R. L. (1979). Effects of contextual interference and age on acquisition, retention, and transfer of motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(2), 179–187.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). afex: Analysis of Factorial Experiments. R package version 0.21-2. https://CRAN.R-project.org/package=afex.

Theodore, R. M., Blumstein, S. E., & Luthra, S. (2015). Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis. *Attention, Perception, and Psychophysics*, *77*(5), 1674–1684.

Tzeng, C. Y., Nygaard, L. C., & Theodore, R. M. (in press). A second chance for a first impression: Sensitivity to cumulative input statistics for lexically guided perceptual learning. *Psychonomic Bulletin & Review*.

White, K. S., & Aslin, R. N. (2011). Adaptation to novel accents by toddlers. *Developmental Science*, *14*(2), 372–384.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, and Psychophysics*, *79*(7), 2064–2072.