Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-021-01885-1

BRIEF REPORT

Listeners are initially flexible in updating phonetic beliefs over time

David Saltzman[1] · Emily Myers[1]

Accepted: 13 January 2021

Abstract

Perceptual learning serves as a mechanism for listeners to adapt to novel phonetic information. Distributional tracking theories posit that this adaptation occurs as a result of listeners accumulating talker-specific distributional information about the phonetic category in question (Kleinschmidt & Jaeger, *Psychological Review*, *122*, 148–203, 2015). What is not known is how listeners build these talker-specific distributions—that is, if they aggregate all information received over a certain time period, or if they rely more heavily upon the most recent information received and down-weight older, consolidated information. In the present experiment, listeners were exposed to four interleaved blocks of a lexical decision task and a phonetic categorization task in which the lexical decision blocks were designed to bias perception in opposite directions of a "s"–"sh" contrast. Listeners returned several days later and completed the identical task again. In each individual session, listener's perception of a "s"–"sh" contrast was biased by the information in the immediately preceding lexical decision block (though only when participants heard the "sh"-biasing block first, which was likely driven by stimulus characteristics). There was evidence that listeners accrued information about the talker over time since the bias effect diminished in the second session. In general, results suggest that listeners initially maintain some flexibility with their talker-specific phonetic representations, but over the course of several exposures begin to consolidate these representations.

Keywords Speech perception · Perceptual learning

Perceptual learning is an inherent component of speech perception. Talkers vary significantly in their phonetic properties (e.g., Hillenbrand, Getty, Clark, & Wheeler, 1995), and consequently listeners must adjust their mapping between acoustics and phonetic categories for each new talker that they encounter. Luckily for the listener, this variability tends to have a statistical structure that is characteristic of the talker. For instance, a given talker may have a consistently longer mean voice onset time for voiceless stops (VOT; Allen, Miller, & DeSteno, 2003), or consistently low F2 value for vowels (Hillenbrand et al., 1995). Further, individual talkers also differ in their variability—that is, one talker may have wide variability in their productions, whereas another may produce a narrower range of acoustic values (Newman, Clouse, & Burnham, 2001).

An array of findings supports the view that listeners are sensitive to the phonetic characteristics of a given talker, and that they adjust their perceptual criteria to use this information to reach a stable phonetic percept (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Kleinschmidt & Jaeger, 2015; Kraljic & Samuel, 2007; Theodore & Miller, 2010). Many accounts of perceptual learning share the notion that talker adaptation involves tracking the statistics of a talker's speech over time, discovering the distributional acoustic patterns associated with each novel talker, and using this information to create probabilistic maps between acoustics and linguistic representations (Maye, Weiss, & Aslin, 2008; McMurray, Aslin, & Toscano, 2009; recently formalized using a Bayesian framework in Kleinschmidt & Jaeger, 2015). Under this view, distributional information is combined with contextual information (e.g., "who is the talker"; "what word is likely in this context") to generate a talker-specific, contextually bound probability that a given acoustic token will match a likely phonetic category. This class of theories predicts that changing the statistical distribution of tokens in the input will ultimately result in perceptual adaptation.

This article is a replication and replacement of Saltzman and Myers (2018), which was retracted after the authors discovered an error in stimulus presentation during the phonetic categorization task.

✉ Emily Myers
emily.myers@uconn.edu

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

Published online: 19 March 2021

Perceptual learning paradigms (e.g., Bertelson, Vroomen, & De Gelder, 2003; Norris, McQueen, & Cutler, 2003) demonstrate many situations in which listeners quickly adapt to phonetic characteristics of a novel talker. Listeners might hear an ambiguous speech token whose ambiguity is resolved either by lexical context (e.g., Norris et al., 2003) or audiovisual information (e.g., Bertelson et al., 2003), accompanied by a clear version of the contrasting phonetic category. The speech stream thus contains both top-down contextual information (i.e., "in the lexical context, 'epi_ode', 's' is the only probable interpretation of the ambiguous sound") as well as bottom-up distributional information (i.e., listeners are exposed to a bimodal distribution of tokens—one ambiguous, one clear—that is shifted for each of the exposure conditions). Using the distributional learning framework, the effect found in perceptual learning studies can be explained as the listener pairing top-down information about phoneme identity with distributional information about the statistics of the novel talker's input, which in turn allows for reshaping of their phonetic categories (Pajak, Fine, Kleinschmidt, & Jaeger, 2016).

The argument that listeners maintain distributional information for each unique talker is described in Kleinschmidt and Jaeger (2015), which follows earlier research that listeners do indeed maintain talker-specific information (Goldinger, 1996; Nygaard & Pisoni, 1998). For instance, in a study by Kraljic and Samuel (2005), listeners were exposed to a male and female talker whose fricatives were biased in opposite directions, then tested on these tame talkers. Listeners maintained separate talker-specific criteria for the two talkers (see also Eisner & McQueen, 2006; Luthra et al., under review). Moreover, talker-specific distributional representations help to explain how perceptual learning effects persist over time (Kraljic & Samuel, 2005). For instance, in a study from Eisner and McQueen (2006), participants maintained talker-specific information over a 12-hour delay, unaffected by exposure to different speakers in the intervening period between exposure and testing, suggesting that any new mapping between acoustics and phonology was specific to the test talker. In Kraljic and Samuel (2005, Experiment 3), participants engaged in a lexically guided perceptual learning (LGPL) task in which they were first exposed to phonetically biasing information—namely, ambiguous tokens embedded in an unambiguous lexical context, then exposed to unaltered exemplars of previously ambiguous sounds from the same talker, and then tested. This led to an extinction of the perceptual learning effect, which is congruent with the notion from distributional tracking theories that listeners would integrate the good exemplars into the talker-specific phonetic distribution, thus disrupting the shifted category representations that they had formed for this new talker. A following experiment confirmed that this disruption was due to unlearning and not simply an extinguishing of the original effect via selective adaptation (Experiment 4).

One issue that has received less attention is the processes by which listeners integrate new, recently encountered talker-specific information (here termed "recent statistics") with existing information that listeners have accumulated about the total talker-specific distribution of acoustic cues (here termed "global statistics"). Kleinschmidt and Jaeger (2015) state that "in situations like a recalibration experiment where listeners encounter odd-sounding, often synthesized speech in a laboratory setting, they may have little confidence, a priori, that any of their previous experiences are directly informative" (p. 13), and thus predict listeners will be maximally flexible during these experiments as the value of previous experiences with the category in question are not believed to be informative. The results of Kraljic and Samuel (2005) appear to confirm that recently encountered statistics are given a stronger weighting; that is, if listeners heavily weight new tokens, the most recent input should more strongly shape the phonetic category. Furthermore, in a series of experiments by Van Linden and Vroomen (2007, Experiments 1–4), listeners were exposed to both lip-reading and lexically biasing information for a "t"–"p" contrast in a blocked design, and the effect of the biasing information was sampled sporadically in each block. Their results demonstrate that (1) listeners can shift their category boundaries flexibly within an experiment and (2) use the most recent statistics when building a distribution. Contrastively, Kleinschmidt and Jaeger (2015) posit that talker-specific distributions cannot be created or maintained if a listener simply tracks the recent statistics from a talker (p. 26), and go on to demonstrate a model for how beliefs about a talker are updated over experiences (Fig. 17). Support for giving greater weight to earlier experience, in line with a "global statistics" account, was also demonstrated in Kraljic, Samuel, and Brennan (2008), in which participants only showed the expected perceptual learning effect when the ambiguous stimuli (which should shift the listener's category boundary) were presented before the clearly produced stimuli. When the order was reversed (a block of clearly produced stimuli presented before a block of ambiguous stimuli) there was no perceptual learning effect.

In the current study, we ask whether listeners are continuously flexible in their adjustment to new and conflicting phonetic information about a talker, and how this affects their ability to create a talker-specific cue distribution. One possibility is that participants aggregate all the input from a given talker into one unified distribution, assigning equal weight to each token in memory. In this case, a listener who hears ambiguous tokens in an "s"-biasing context, for instance, and is tested on this contrast should see the previously attested shift in phonetic category boundary. Subsequent exposure to an "sh"-biasing block, however, will simply add new tokens to the emerging "s" and "sh" distributions for the talker (see "global statistics" in Fig. 1c), leaving the category boundary somewhere in the middle of the distribution. Under this view,

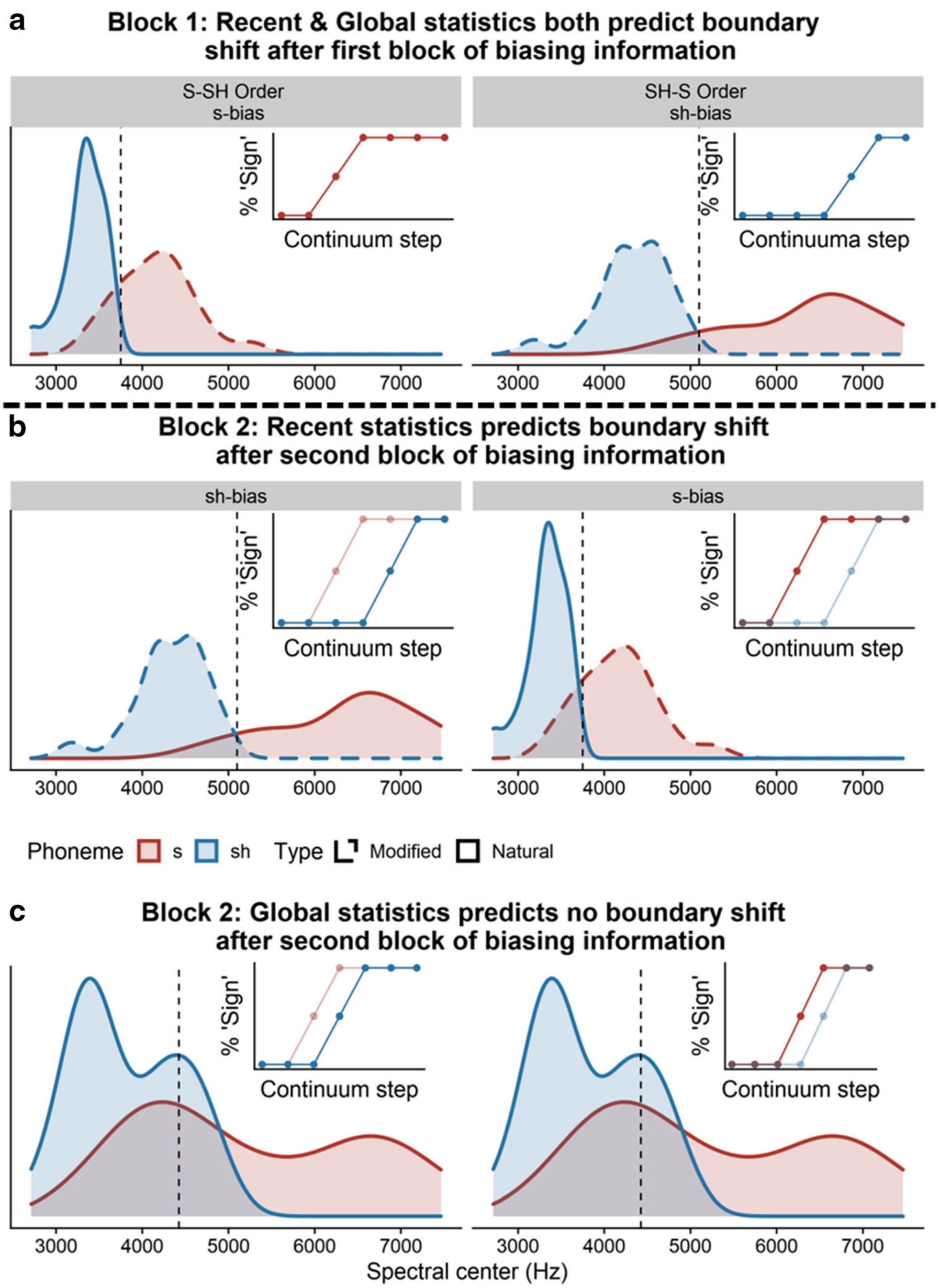

Fig. 1 Schematic showing the probability density function over the centroid frequencies of the "s" and "sh" tokens that listeners hear on each block of the LD task (see Drouin, Theodore, & Myers, 2016). Blue shows the "sh" tokens, red shows the "s" tokens. In SH-bias blocks, listeners hear naturally produced versions of the "s" tokens (red) and altered, ambiguous versions of the "sh" tokens, while the reverse is true of S-bias blocks. In each panel, a vertical dotted line indicates a hypothetical "ideal" boundary that minimizes mis-categorizations of the exposure set. Inset into each density function plot is the expected categorization function from participants in during the PC task following each LD block. Solid lines indicate the expected categorization function in response to the most recent LD block, while the transparent line represents the expected categorization function from the earlier LD block. The two orders of lexical decision blocks are represented, with the S-SH order in the left column, and the SH-S order in the right column. **a** For Block 1, both recent statistics and global statistics hypotheses predict the same boundary shift in response to the biasing information contained in the LD block. **b** For Block 2, the recent statistics hypothesis predicts that listeners will resolve on a boundary dictated by the immediately previous LD block, now shifting their categorization function in the opposite direction of Block 1. **c** For Block 2, the global statistics view predicts that listeners will generate distributions over the entire set of LD stimuli they have received thus far, and thus both groups of participants will show the same boundary value at Block 2

a participant who had heard the "sh"-bias first would show a shift to incorporate ambiguous tokens in the "sh" category, but her categorization function after being exposed to "s" tokens next would be equivalent to the participant who heard the bias blocks in the opposite order, since both participants would have heard the full complement of stimuli by the end of the experiment. Essentially, this leads to a prediction that the order of presentation of these blocks will matter, with categorization functions equalizing after listeners have heard both "s" and "sh"-biasing blocks. The biasing effect should be even more diminished if the participant were to return and complete the same task again, as the listener should be updating their beliefs about the talker's distribution with an aggregate of all of the information they received in their first exposure to the talker. In addition, the talker should now be more familiar, which should allow the listener to safely utilize this prior experience with the talker to inform their future experience with input from said talker. In essence, the more speech a listener hears from a talker, the "heavier" the distributional information for that talker, and the harder it should be to shift.

An alternative is that listeners are maximally flexible, easily disregard old information about a talker, looking only to the most recently encountered tokens when considering how to process incoming information (see "recent statistics" in Fig. 1b). This would predict that listeners will shift and reshift their phonetic criteria on the basis of recent information, and that the shift for the second-encountered bias will be just as large as that for the first set of biasing information a listener hears. Upon a second exposure to the same task, we should see listeners continue to shift and reshift their category boundaries as a result of the biasing information. Following this hypothesis, it is possible listeners will create a very flexible talker-specific distribution (or perhaps do not create one at all, which is discussed later) and simply move around in that distributional space.

To test these alternatives, in the current study we manipulated lexical bias within-participant, otherwise closely following methods of Kraljic and Samuel (2005). Listeners were exposed to four interleaved blocks of a lexical decision task and a phonetic categorization task (see Fig. 1) in which the lexical blocks were designed to bias perception in opposite directions. Listeners also returned several days later for a second session in which they completed the identical task from their first session.

If listeners behave per the global statistics hypothesis, performance on the final phonetic categorization task in Session 1 will be the same regardless of whether the lexical-decision block immediately preceding it was "s" or "sh"-biasing, and therefore we should see a main effect of order (i.e., a shift in the predicted direction after the first biasing block, then an equilibration of the effect after being exposed to the opposite-direction bias). Also, per this hypothesis, the main effect of bias could still be present during Session 2, but significantly reduced in size compared with Session 1 (or extinguished altogether in an extreme case), leading to a Bias × Session interaction. However, if listeners behave per the recent statistics hypothesis, we should see a large boundary shift in their categorization functions following each biasing block, regardless of the order of the blocks. This effect should also reproduce during the second session.

Method

Participants

A total of 114 participants (55 females, $M_{\text{age}} = 36.49$ years, $SD = 12.74$ years) were recruited from the online participant pool Prolific. All participants indicated that they were monolingual English speakers with normal hearing and no history of language impairment. Participants were geographically restricted to the United States. Informed consent was obtained from every participant in accordance with the guidelines of the University of Connecticut Institutional Review Board. Participants received monetary compensation for their participation in each session.

Stimuli

Stimuli for the lexical decision (LD) task were taken from Myers and Mesite (2014).[1] These items consisted of 200 total words, 100 filler nonwords, 60 filler real words, 20 critical "s" words, and 20 critical "sh" words. The critical words were real words containing either a "s" or "sh" in a word-medial position. Acoustically modified versions of these words were created by replacing the "s" or "sh" with an ambiguous, 50%/50% blend of the two sounds. Further details about stimuli can be found in Myers and Mesite (2014). In the "s"-biasing condition, listeners heard words containing the ambiguous blend ("?") in "s"-containing words and unaltered versions of the "sh"-words (e.g., "epi?ode", "flourishing"). In the "sh"-biasing condition, the ambiguous blend appeared in "sh"-words and listeners also heard unaltered versions of the "s"-words (e.g., "flouri?ing", "episode").

Items for the phonetic categorization (PC) task consisted of a seven-step continuum from "shine" to "sign," which were created in PRAAT (Boersma & Weenink, 2017) by blending (through waveform averaging) fricatives derived from the words "sign" and "shine" at different proportions, from 20% "s"–80% "sh" (Step 1) to 80% "s"–20% "sh" (Step 7). The blended fricatives were then inserted into the "sign" frame. The "shine"–"sign" continuum was pilot tested to ensure consistent perception of the endpoints of the continuum. The same talker was used for the LD and PC stimuli.

[1] See the Stimulus Selection and Stimulus Construction subsections in Methods and Materials of Myers and Mesite (2014).

Procedure

The experiment took place over two sessions (see Fig. 2). In Session 1, participants engaged in alternating blocks of a LD task and PC task. LD blocks contained lexical information that biased listeners to perceive an ambiguous phoneme as either “s” or ”sh,” and PC blocks tested the effects of having heard this biasing information. Participants were randomly assigned to either the S-SH group (in which the first LD block contained “s”-critical words and the second the ”sh”-critical words) or the SH-S group (the reverse order). The PC task was identical across groups. In Session 2, participants returned between 7 and 16 days later (*M* = 8.21 days, *SD* = 1.63 days) and completed the identical task as in Session 1 assigned to the same order of presentation as in their first session.

In the LD task, participants were asked to indicate whether the stimulus was a word or a nonword by pressing a corresponding key on the keyboard as quickly and accurately as possible. There were 200 trials in each LD block. In the PC task, participants were asked to indicate whether they perceived the stimulus as “sign” or “shine,” which they did by pressing a corresponding key on the keyboard as quickly as possible. Each categorization task consisted of eight repetitions of each of the seven tokens from the “sign” to “shine” continuum presented in random order, for a total of 56 trials per PC block. Response options were counterbalanced in both tasks.

To attempt to replicate in-laboratory quality standards in an online study, we implemented two headphone checks based upon Woods et al. (2017) that participants completed before starting the experiment. Participants were instructed to first adjust the volume to a comfortable listening level while wearing headphones, and then were presented with three pure tones and asked to indicate which sound was the softest. Unbeknownst to the participants, one of the sounds is presented 180^{o} out of phase across the stereo channels. Without using headphones, it is very difficult to complete this task accurately. There were six trials, and participants were considered to have passed if they responded correctly in at least four of the six trials. If participants failed the first headphone check, they were reminded to wear headphones if they were not already, and then advanced to a second round of the headphone check. Success or failure in passing the headphone check(s) was marked in each participants data file and used for later exclusionary purposes.

Results

Only participants who completed both sessions of the study (*n* = 98) were included for analysis. Further exclusions were made for not passing the headphone checks (*n* = 15 excluded) or for having below 80% accuracy at each endpoint for the PC tasks (*n* = 20

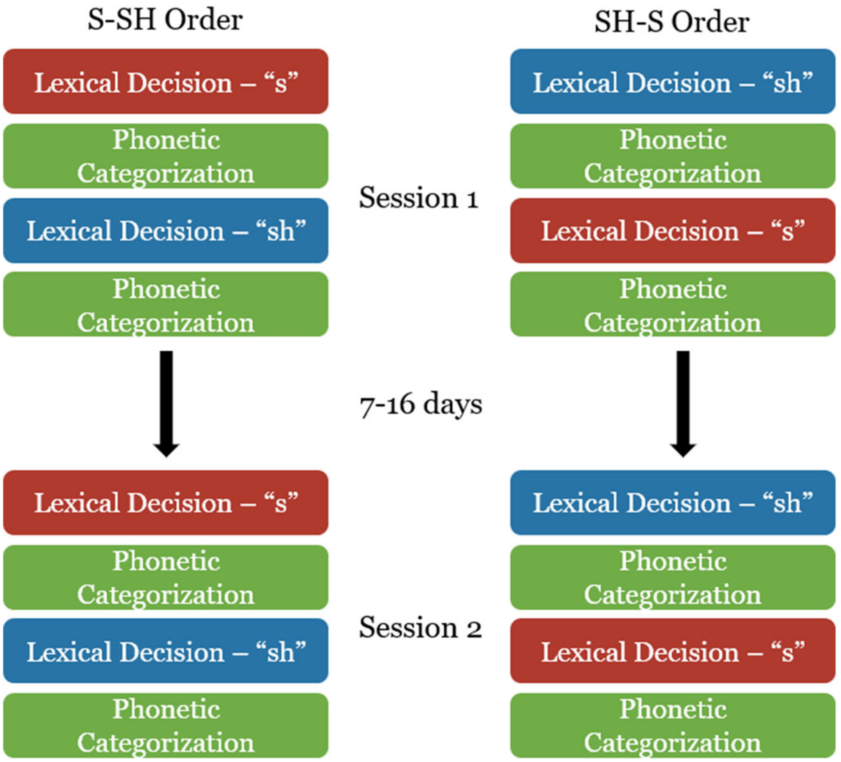

Fig. 2 Experimental schedule. Participants were assigned to either the S-SH group or SH-S group (see text for details). In each session, Lexical Decision (red: “S”-biasing block, blue: “SH”-biasing block) blocks alternated with Phonetic Categorization (green) blocks. Participants returned after 7–16 days to repeat the identical experimental procedure. (Color figure online)

excluded), leaving data from a total of 63 participants. Data from both sessions was combined for a single omnibus analysis.

A generalized linear mixed-effects model with a logit-link was performed in the R statistical computing language (R Development Core Team, 2014) using the “mixed” command from the *afex* package (Singmann, Bolker, Westfall, & Aust, 2020). The *p* values were estimated using a likelihood ratio test. A backward-stepping selection heuristic was used to achieve model selection. All factors were sum coded. The model selected contained fixed effects of continuum step (centered), biasing condition (S-bias vs. SH-bias), order of presentation (S-SH vs SH-S), and session (Session 1 vs. Session 2), and their interactions with by-subject random slopes and intercepts for continuum step, biasing condition, session, and their three-way interaction. The output from the *afex* ANOVA table is reported.

As expected, a significant main effect of continuum step was revealed, such that participants were more likely to indicate that they heard “sign” as the proportion of “s” in the fricative blend increased (χ^2 = 170.18, p < .001; see Fig. 3). In addition, a significant main effect of biasing condition was found, reflecting increased “sign” responses immediately

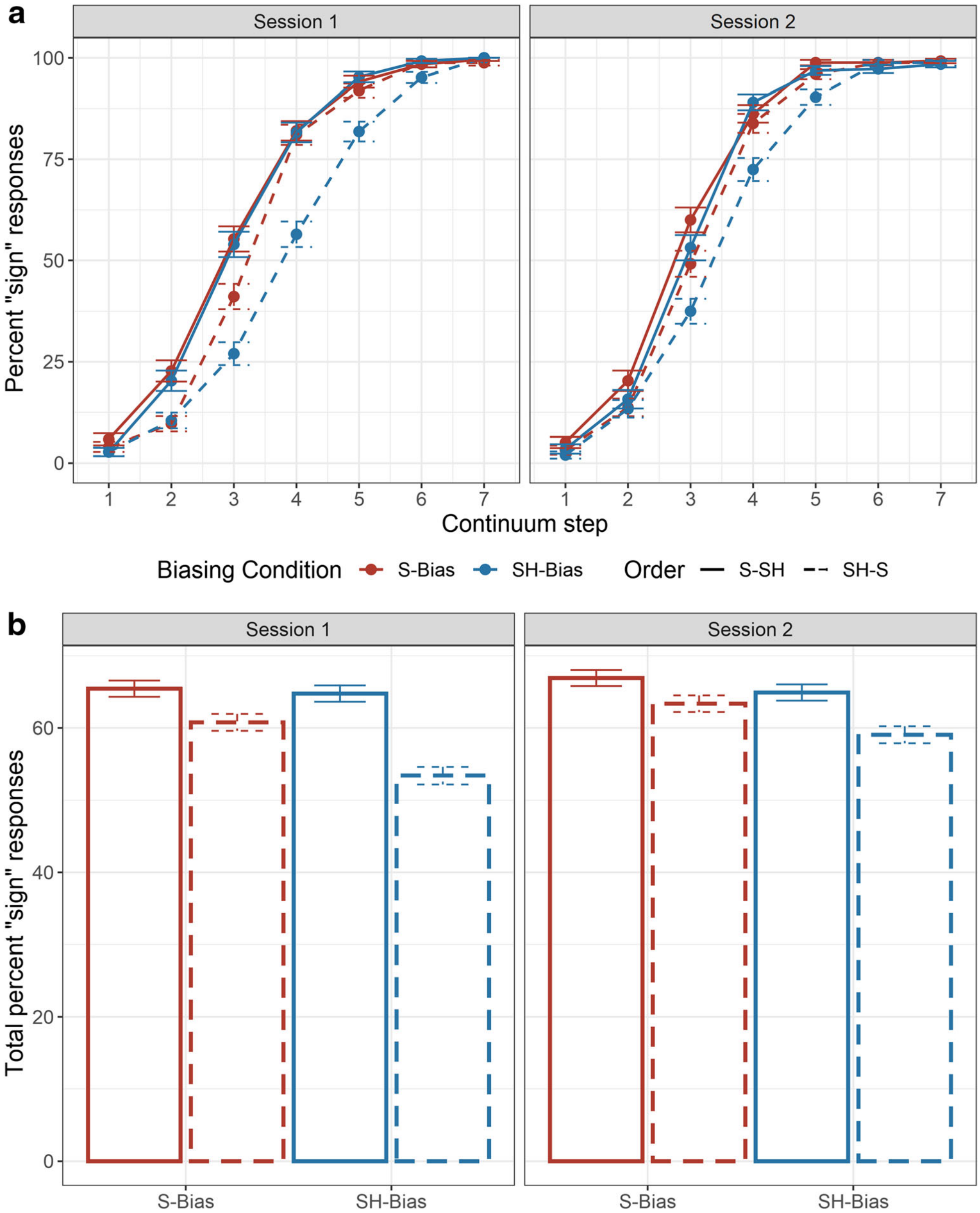

Fig. 3. **a** Data from the phonetic categorization task. Order (S-SH, SH-S) was a between-subjects factor. “Biasing condition” indicates the type (“sh”-biasing or “s”-biasing) of the immediately preceding LD block. Error bars reflect standard error of the mean. **b** The phonetic biasing effect collapsed across continuum step. Lower total percentage “sign” responses indicate a bias towards the “shine” side of the continuum, while higher total percentage “sign” responses indicate a bias towards the “sign” end of the continuum

following a "s" LD block and decreased "sign" responses immediately following a "sh" LD block (χ^2 = 13.17, p < .001). There was also a main effect of order of presentation, with greater overall "sign" responses for the S-SH order (χ^2 = 5.23, p = .02), and a main effect of session, with greater overall "sign" responses for Session 2 (χ^2 = 6.39, p = .01). In addition, there was a significant Biasing Condition × Order interaction (χ^2 = 9.27, p = .002), and a significant Biasing Condition × Order × Session interaction (χ^2 = 5.98, p = .01), which we unpack below.

The two-way interaction between biasing condition and order was explored using post hoc estimated marginal means comparisons from the R package *emmeans* (Lenth, Singmann, & Love, 2020). This interaction was driven by a finding that bias effects were limited to the SH-S order only—that is, participants showed a significant reshift of phonetic category boundary only when they heard the SH-biased condition first (log odds ratio = 0.981, p < .0001), but participants who heard the S-biased condition first showed no subsequent shift in boundary when confronted with the SH-biased condition (log odds ratio = 0.105, p = .60).

This asymmetry between the SH-S and S-SH orders also emerged in the three-way interaction between biasing condition, order of presentation, and session. The interaction was driven by the magnitude of the biasing effect in the SH-S order decreasing in Session 2 compared to Session 1 (log odds ratio = 0.60, p = .04, Tukey adjusted), indicating that participants appeared to be aggregating global statistics about the talker over time. There was no significant difference in the magnitude of the biasing effect in the S-SH order from Session 1 to Session 2 (log odds ratio = −0.23, p = .79, Tukey adjusted).

Intraindividual stability of perceptual learning effect

In order to examine the stability of the perceptual learning effect within individuals, the magnitude of the biasing effect for individual participants was estimated by fitting psychometric functions (using the R package *quickpsy*) to their categorization responses in the PC blocks and extracting the crossover point. The difference between the crossover point for the "s"-bias condition and "sh"-bias condition in each session was then calculated, which serves as an indicator of the size of the biasing effect. Next, we calculated the Pearson correlation between the size of this shift during Session 1 compared with Session 2. A small positive correlation (r = 0.25, p = .05) was observed, indicating that on average, participants showed a similar size shift in the first session as they did in the second session. The correlation in shift size did not change when participants were separated by order of presentation (r = .22 for both groups, though neither reach significance).

Lexical decision

Accuracy in the lexical decision blocks was high across the course of the experiment (M = 93.6%, SD = 24.4%; see Table 1), as expected from previous studies that have used this same word list (e.g., Kraljic & Samuel, 2005, 2006). To confirm these findings, responses in the lexical decision tasks were submitted to a generalized linear mixed-effects models with a logit-link, with fixed effects of word type (ambiguous words vs. clear words, both types of filler items were excluded from analysis), biasing condition (S-bias vs. SH-bias), order of presentation (S-SH vs SH-S), and session (Session 1 vs. Session 2) and their interactions. The model selected contained by-subject random slopes and intercepts for the three-way interaction of word type, biasing condition, and session, as well as random slopes and intercepts for word type and biasing condition.

There was a main effect of word type (χ^2 = 24.54, p < .001), with mean accuracy higher for clear words (98.3%) than for ambiguous words (91.1%). There was also a main effect of session (χ^2 = 5.90, p = .02), with overall slightly higher mean accuracy in Session 1 (95.2%) compared with Session 2 (94.2%). There were also three significant two-way interactions: Biasing Condition × Order of Presentation (χ^2 = 8.27, p = .004), Word Type × Biasing Condition (χ^2 = 3.95, p = .047), and Biasing Condition × Session (χ^2 = 9.09, p < .003). None of the three-way interactions or the four-way interaction reached significance.

Discussion

Distributional learning accounts of speech typically do not specify precisely how individual episodes are aggregated over time in order to inform perceptual learning (Kleinschmidt & Jaeger, 2015; Maye et al., 2008; McMurray et al., 2009). As a

Table 1 Participants' performance on LD blocks the experiment, as a function of biasing condition, order of presentation, and session

Biasing condition	Order	Session	Mean accuracy	*SD*
S-bias	S-SH	1	93.0%	25.5%
S-bias	S-SH	2	94.2%	23.4%
S-bias	SH-S	1	93.6%	24.5%
S-bias	SH-S	2	94.5%	22.8%
SH-bias	S-SH	1	92.8%	25.8%
SH-bias	S-SH	2	92.7%	26.1%
SH-bias	SH-S	1	94.0%	23.7%
SH-bias	SH-S	2	94.2%	23.3%

Note. For brevity, accuracy is displayed here collapsing across the four word types.

foray in to answering this question, we conceived of two possible integration schemes—one in which listeners use only the most recent statistical information they have received to adapt to a novel talker (recent statistics), or an alternative account where they continue to integrate new information about the talker with older information already learned (global statistics). The latter account is more consistent with the mechanism for adaptation put forth in the ideal adapter model (Kleinschmidt & Jaeger, 2015) and is similar to what was observed in the present study, though with some critical differences. In the first session, it appears that listeners remained quite flexible; those assigned to the SH-S order of presentation appeared to use biasing lexical information to shift their category boundary first in one direction (e.g., toward the "shine" end of the continuum), and then, when confronted with the opposite bias (e.g., toward the "sign" end of the continuum), back in the other direction. This result suggests that listeners use a relatively short temporal window of integration when they are considering how to interpret the speech of a talker, strongly weighting recent biasing information instead of building a session-long distributional scheme for the talker.

However, consistent with the global statistics hypothesis, by the second testing session, participants in this SH-S condition were less flexible in shifting their category boundaries. In Session 2, the lack of a significant difference between the categorization functions in response to the second exposure block (i.e., "s"-bias for the SH-S order and "sh"-bias for the S-SH order) conforms with the global statistics prediction, which is that listeners will combine the information from the first exposure block with that of the second block, thus leading to an equalization in categorization in response to the second exposure block. In effect, if listeners simply aggregate over all phonetic information about the talker, and the more "reversals" they hear from this talker, the more the curve will regress to the midpoint between the biasing conditions. This follows closely with the findings from Vroomen, van Linden, De Gelder, and Bertelson (2007), who observed that the biasing effect began to decrease for their participants who received at least 64 exposures to the ambiguous stimuli, and those of Theodore and Monto (2019) and Tzeng, Nygaard, and Theodore (2020), which found evidence for a global statistics integration scheme in perceptual learning. Relatedly, a peculiarity introduced by manipulating phonetic bias within-subjects is that participants are exposed to both ambiguous and clear productions of the same word in different lexical decision blocks, a fact that participants may become aware of after the completion of the first session. A parallel to Kraljic et al. (2008) may be drawn, wherein presenting a block of clear items followed by a block of ambiguous items led to an extinguishing of the perceptual learning effect (however, they did not reuse words from one block to the next as in the present study). Therefore, the reduction in the biasing effect in Session 2 could reflect a similar level of unlearning from hearing the formerly ambiguous words produced clearly.

An important issue to address is the asymmetric perceptual learning effect seen for our two orders of presentation (learning for SH-S order, no learning for S-SH order). This can likely be attributed to the particular stimulus set used (though other studies have found similar asymmetries in lexically guided perceptual learning, see Drozdova, Hout, & Scharenborg, 2016; Giovannone & Theodore, 2021; Kraljic et al., 2008), which is explored in detail in Drouin et al. (2016). As can be seen in Fig. 4, the acoustic distribution of the modified "s" tokens differ much more profoundly from their unmodified counterparts than do the "sh" tokens (compare the red curves in the SH-biased—unmodified /s/ to S-biased—modified /s/ panels). This results in the S-biasing condition providing a much stronger basis on which to shift the category boundary than the SH-biasing condition. We speculate that when this condition comes first, listeners become entrenched in this bias scheme. In contrast, the SH-biased condition provides a weaker basis to shift the category

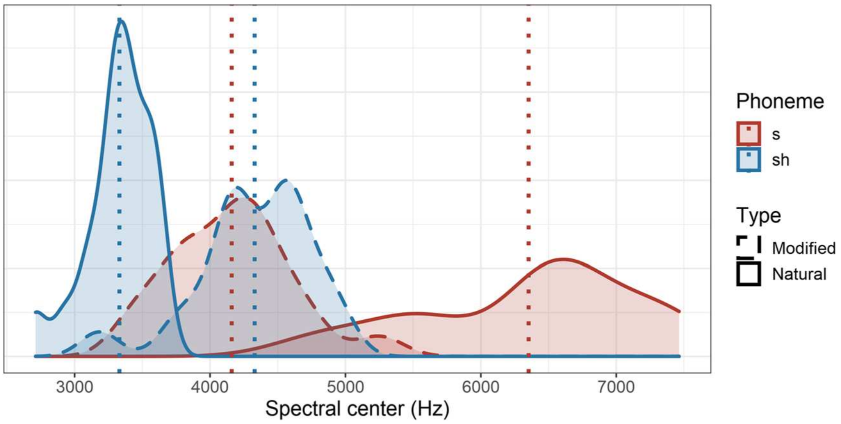

Fig. 4 Probability density function over the centroid frequencies of the "s" and "sh" tokens, both natural and modified, that participants hear in the LD blocks over the course of the experiment. Dotted lines represent mean centroid frequency. Note much larger change in mean for modified "s" token from natural "s" tokens compared with change for modified "sh" tokens from natural "sh" tokens

boundary, leaving listeners susceptible to the stronger effect of the S-biasing condition when they encounter it afterwards. This analysis points to the importance of careful control of the acoustic distributional information in modified natural stimuli. It also suggests that the degree of malleability of listeners perceptual schemes depends in part on the strength of the initial bias, and that particularly unambiguous cues might be needed to shift category boundaries when listeners have strong assumptions about the acoustic qualities of the talker.

Qualitative changes to learned phonetic information may emerge over time, especially after sleep (Earle & Myers, 2015; Fuhrmeister & Myers, 2017; Fuhrmeister et al., 2020). Sleep-mediated consolidation appears to stabilize learned phonetic information, and protect this information from interference (Earle & Myers, 2015; Fenn, Nusbaum, & Margoliash, 2003). If these same principles operate in this paradigm, it follows that distributional information that listeners heard during Session 1 would become stabilized overnight, yielding a lessened ability to respond to distributional learning in Session 2. This was apparent in the present study, as the biasing effect diminished in Session 2 compared with Session 1, which could reflect sleep mediated consolidation stabilizing the learning from Session 1.[2] Though it should be noted that these patterns are also consistent with the notion that participants are instead continuing to learn more about the idiolect of this talker and aggregating information across all exposures, weighting them all equally, which would diminish the biasing effect as well. Future work will need to address this question, as the design of the present study cannot elucidate definitive support for one hypothesis over the other.

One caveat with the design of the present experiment should be noted: because the phonetic information that was provided to listeners in the first session was essentially inconsistent or erratic (and likely unrealistic), it is possible that listeners adapt a conservative strategy in interpreting the speech of the talker, and they do not settle into any particular phonetic boundary for that talker for at least their first exposure. This would explain the significant biasing effect seen in Session 1 and could come about from bottom-up mechanisms (the distribution is too broad and shallow for the system to settle) or from top-down mechanisms (the talker is viewed somehow as unreliable; see, for instance, Kraljic et al., 2008). Kleinschmidt and Jaeger (2015) also discuss the possibility that listeners may not always form talker-specific beliefs, especially in situations where there are no expectations that they would be useful again in the future (such as in an experiment). While this hypothesis could explain the findings in Session 1, participants were instructed during the consent process that the two sessions would be identical. Given that these instructions came before an explanation of the experiment's procedure, and were not wholly explicit, it is possible the participants assumed there was no reason to form talker-specific beliefs for a transient situation. Nevertheless, multiple studies of perceptual learning have found that talker-specific phonetic distributions persist over time, implying that participants may form them regardless of the situation (Kraljic & Samuel, 2005; Eisner & McQueen, 2006). Future research should explore the effect of top-down instructions on listener's willingness to create talker-specific distributions.

An auxiliary question that this study allows us to answer is whether individuals are consistent in the size of the boundary shift that they display across sessions. A secondary analysis showed that there was a weak, marginally significant positive relationship between the size of the biasing effect across sessions. Intraindividual consistency of perceptual learning has been largely unexplored, though Zheng and Samuel (2020) found no relationship between an individual's performance on a lexically guided perceptual learning task and their performance on an accent accommodation task (in which listeners identify words spoken with an accent), though this is not necessarily the same as the present study, in which participants were retested on the same task twice. Individual differences in language learning are becoming of increasing interest to explain the gulf in outcomes between learners; for instance, incidental language learning paradigms have found that factors like declarative learning abilities, procedural memory, some learning styles, personality factors, and sequence learning can have an effect on learning performance (Granena, 2013; Grey, Williams, & Rebuschat, 2015; Hamrick, 2015). More relevant to distributional tracking theories is the idea that statistical learning may be a skill unto itself, with accompanying individual differences; Siegelman and Frost (2015) found that their participants' performance on a series of statistical learning tasks was stable at an individual level across time. Though the present study does support this notion, it should be noted that the metric of stability we used is inherently confounded with learning the talker-specific phonetic representations. Therefore, it is possible that those who do show a "stable" perceptual learning effect (i.e., the same magnitude boundary shift in both sessions) are simply poorer or slower phonetic adapters. Nevertheless, it is apparent that there are individual differences in the integration of phonetic information over time; some participants showed no biasing effect when tested in the second session, which appears to indicate the information they obtained about the speaker in these two sessions was sufficient for them to settle upon phonetic representations, while others still remained flexible. Future research will need

[2] Eisner and McQueen (2006) found no difference in perceptual learning between participants who were tested 12 hours after exposure and those tested the next day after sleep. However, because there was only a single exposure/test, it is impossible to conclude whether there were differences in the strength of consolidation (i.e., whether the group that was tested after sleeping would then show less learning in a second exposure).

to be directed to better understand the origins of this difference.

Conclusion

For the novel talkers in the current experiment, listeners appear to be initially flexible to the most recent biasing information they are provided with, but later go on to aggregate the information they have previously learned about the talker, becoming less flexible. Therefore, listeners seem to use a global integration scheme for new phonetic information, but even within this scheme there are still open questions about the dynamics of this process; future work will need to address how new, unexpected phonetic information from a learned talker is accommodated by the perceptual system, which rapidly adapts despite the wealth of aggregated phonetic information from the talker. A global statistics account in which listeners weighted all tokens from a talker equally would generate the prediction that it would be extremely difficult to adapt to the speech of very well-known talkers if a new perturbation or disruption was introduced. It would mean, for instance, that listeners who had only seen Meryl Streep playing American roles would struggle when confronted with her Polish-accented English in *Sophie's Choice*, or that one might fail to understand the distorted speech of one's parent after dental surgery. Therefore, room needs to be made in this theory for the intervention of top-down expectations to rapidly upweight new phonetic information. This delicate interplay is likely to be an enduring question in the understanding of spoken word recognition.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13423-021-01885-1.

Acknowledgments This work was supported by NIH NIDCD Grant R01 DC013064 to E.B.M. The views expressed here reflect those of the authors and not the NIH or the NIDCD. We would like to thank Rachel Theodore for very helpful comments on an earlier version of this manuscript, and Julia Drouin for her contributions to the acoustic analysis.

The data and materials for all experiments are available at https://github.com/disaltzman/SM-PBR-18-Replication.

References

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*, 544–552. https://doi.org/10.1121/1.1528172

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*, 592–597. https://doi.org/10.1046/j.0956-7976.2003.psci_1470.x

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer (Version 6.0.26) [Computer software]. http://www.praat.org/. Accessed Oct 2020

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. https://doi.org/10.1016/j.cognition.2008.04.004

Drouin, J. R., Theodore, R. M., & Myers, E. B. (2016). Lexically guided perceptual tuning of internal phonetic category structure. *The Journal of the Acoustical Society of America*, *140*(4), EL307–EL313. https://doi.org/10.1121/1.4964468

Drozdova, P., Hout, R. V., & Scharenborg, O. (2016). Lexically-guided perceptual learning in non-native listening*. *Bilingualism: Language and Cognition*, *19*(5), 914–920. https://doi.org/10.1017/S136672891600002X

Earle, F. S., & Myers, E. B. (2015). Sleep and native language interference affect non-native speech sound learning. *Journal of Experimental Psychology: Human Perception and Performance*, *41*, 1680–1695. https://doi.org/10.1037/xhp0000113

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, *119*, 1950–1953. https://doi.org/10.1121/1.2178721

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature*, *425*(6958), 614–616. https://doi.org/10.1038/nature01951

Fuhrmeister, P., & Myers, E. B. (2017). Non-native phonetic learning is destabilized by exposure to phonological variability before and after training. *The Journal of the Acoustical Society of America, 142*(5), EL448–EL454. https://doi.org/10.1121/1.5009688

Fuhrmeister, P., Smith, G., & Myers, E. B. (2020). Overlearning of non-native speech sounds does not result in superior consolidation after a period of sleep. *The Journal of the Acoustical Society of America*, *147*(3), EL289–EL294. https://doi.org/10.1121/10.0000943

Giovannone, N., & Theodore, R. M. (2021). Individual differences in lexical contributions to speech perception. *Journal of Speech, Language, and Hearing Research*.

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. https://doi.org/10.1037/0278-7393.22.5.1166

Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning*, *63*, 665–703. https://doi.org/10.1111/lang.12018

Grey, S., Williams, J. N., & Rebuschat, P. (2015). Individual differences in incidental language learning: Phonological working memory, learning styles, and personality. *Learning and Individual Differences, 38*, 44–53. https://doi.org/10.1016/j.lindif.2015.01.019

Hamrick, P. (2015). Declarative and procedural memory abilities as individual differences in incidental language learning. *Learning and Individual Differences, 44*, 9–15. https://doi.org/10.1016/j.lindif.2015.10.003

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*, 3099–3111. https://doi.org/10.1121/1.411872

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. https://doi.org/10.1037/a0038695

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*, 141–178. https://doi.org/10.1016/j.cogpsych.2005.05.001

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13*(2), 262–268. https://doi.org/10.3758/BF03193841

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*, 1–15. https://doi.org/10.1016/j.jml.2006.07.010

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. https://doi.org/10.1111/j.1467-9280.2008.02090.x

Lenth, R., Singmann, H., & Love, J. (2020). Emmeans: Estimated marginal means, aka least-squares means. (R package version 1.4.5). https://cran.r-project.org/web/packages/emmeans/index.html. Accessed Oct 2020

Luthra, S., Mechtenberg, H., & Myers, E. B (In press). Perceptual learning of multiple talkers requires additional exposure. *Attention, Perception, & Psychophysics*.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*, 122–134. https://doi.org/10.1111/j.1467-7687.2007.00653.x

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, *12*, 369–378. https://doi.org/10.1111/j.1467-7687.2009.00822.x

Myers, E. B., & Mesite, L. M. (2014). Neural systems underlying perceptual adjustment to non-standard speech tokens. *Journal of Memory and Language*, *76*, 80–93. https://doi.org/10.1016/j.jml.2014.06.007

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*, 1181–1196. https://doi.org/10.1121/1.1348009

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Attention, Perception, & Psychophysics*, *60*, 355–376. https://doi.org/10.3758/BF03206860

Pajak, B., Fine, A. B., Kleinschmidt, D. F., & Jaeger, T. F. (2016). Learning additional languages as hierarchical probabilistic inference: Insights from L1 processing. *Language Learning*, *10*, 900–944. https://doi.org/10.1111/lang.12168

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language, 81*, 105–120. https://doi.org/10.1016/j.jml.2015.02.001

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2020). afex: Analysis of factorial experiments (R package version 0.27-2). https://CRAN.R-project.org/package=afex. Accessed Oct 2020

Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *The Journal of the Acoustical Society of America*, *128*, 2090–2099. https://doi.org/10.1121/1.3467771

Theodore, R. M., & Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker's phonetic distributions. *Psychonomic Bulletin & Review, 26*(3), 985–992. https://doi.org/10.3758/s13423-018-1551-5

Tzeng, C., Nygaard, L., & Theodore, R. M. (2020). A second chance for a first impression: Sensitivity to cumulative input statistics for lexically guided perceptual learning. *Psychonomic Bulletin & Review*, 1–29. https://doi.org/10.3758/s13423-020-01840-6. Advance online publication.

van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *J ournal of Experimental Psychology: Human Perception and Performance, 33*(6), 1483–1494. https://doi.org/10.1037/0096-1523.33.6.1483

Vroomen, J., van Linden, S., De Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory–visual speech perception: Contrasting build-up courses. *Neuropsychologia*, *45*, 572–577. https://doi.org/10.1016/j.neuropsychologia.2006.01.031

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. Attention, Perception, & Psychophysics, 79(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Zheng, Y., & Samuel, A. G. (2020). The relationship between phonemic category boundary changes and perceptual adjustments to natural accents. Journal of Experimental Psychology: Learning, Memory, and Cognition, 46(7), 1270–1292. https://doi.org/10.1037/xlm0000788

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