Cognition 270 (2026) 106412

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Exposure to an accent transfers to speech production in a single shot

Timothy K. Murphy [a,b,*], Lori L. Holt [c], Nazbanou Nozari [d,e]

[a] Department of Otolaryngology, University of Wisconsin-Madison, Madison, WI, USA
[b] Waisman Center, University of Wisconsin-Madison, Madison, WI, USA
[c] Department of Psychology, The University of Texas at Austin, Austin, TX, USA
[d] Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA
[e] Cognitive Science Program, Indiana University, Bloomington, IN, USA

ARTICLE INFO

Keywords:
Psycholinguistics
Speech perception
Speech production
Production-perception links
Speech planning
Statistical learning

ABSTRACT

Listening to another speaker's voice can lead to predictable changes in a listener's own voice. This means that perception can alter production. A key question is whether overt production and its auditory consequences are critical for observing such changes. We answer this question in two experiments ($N = 269$) by passively exposing participants to speech that carries different acoustic patterns and investigating changes to production. Experiment 1 shows that decreasing the number of productions by an order of magnitude does not decrease the influence of perception on production. Experiment 2 takes this further by demonstrating that perceptual influence manifests on the very first overt production after exposure to new speech regularities. Collectively, these results show that perception can alter production without relying on feedback from overt production and its auditory consequences. This finding, in turn, strongly supports speech production models that include internal simulations.

1. Introduction

When speaking to a person, many of us find ourselves speaking more like them. However, the mechanisms underlying this phonetic convergence are not well understood. On the one hand, social factors, such as the perceived social status and likeability of the interlocutor, are known to influence convergence (e.g., Babel, 2012; Pardo, Gibbons, Suppes, & Krauss, 2012). On the other hand, the basic effect is found even when social influences are minimized (e.g., Murphy, Nozari, & Holt, 2024), pointing to a core cognitive mechanism underlying convergence. Recently, we have shown that exposing individuals to recordings of simple words, like *beer* and *pier*, that follow the statistical regularities of American English or deviate from them slightly to convey an accent shifts how listeners use acoustic dimensions in speech perception and transfers to similarly impact listeners' own speech productions (Murphy et al., 2024; Murphy, Nozari, & Holt, 2025). We have further shown the ecological validity and broad scope of this effect (Huffaker, Holt, & Nozari, 2025; Thorburn et al., in press). Key questions remain: How critical is the role of overt production in this transfer? Is convergence driven by repeated production and fine-tuning of the production system — perhaps through sensory feedback from one's own voice — or do shifts in speech perception more readily transfer to influence production? Here, we answer these questions. In doing so, we shed light on the mechanisms underlying phonetic convergence and address the bigger-picture question of how perception affects action.

It is well-known that perceptual consequences of one's own actions guide both learning and action monitoring, including speech (e.g., Miall & Wolpert, 1996). This is elegantly laid out in a neurobiologically plausible computational model of speech production called Directions into Velocities of Articulators (DIVA; e.g., Tourville & Guenther, 2011). Fig. 1 shows the regulation of production through auditory feedback in DIVA.[1] Production activates the Speech Sound Map (left ventrolateral premotor cortex) for a phoneme, syllable, or word. This triggers a forward model, which sends a motor command to the Articulator Map (motor cortex) to produce the utterance. In parallel with the forward model, the Speech Sound Map also activates a feedback control system. Projections to higher-order auditory cortical areas (posterior superior temporal gyrus/sulcus and planum temporale) generate an expectation for the auditory percept of the currently produced utterance in the Auditory Target Map. This auditory target is compared to the incoming auditory signal (heard speech) in the Auditory State Map. If the Auditory Target and Auditory State maps do not match, an error signal is

* Corresponding author at: Department of Otolaryngology, University of Wisconsin-Madison, Madison, WI, USA.
E-mail address: tmurphy37@wisc.edu (T.K. Murphy).

[1] Note that DIVA also contains a somatosensory feedback loop, but we focus on the auditory loop as it is relevant to the current study.

https://doi.org/10.1016/j.cognition.2025.106412
Received 24 March 2025; Received in revised form 8 December 2025; Accepted 15 December 2025
Available online 25 December 2025

generated in the Auditory Error Map (higher-order auditory cortical areas). The error signal is conveyed to the Feedback Control Loop Map (right ventral premotor cortex) which, in turn, issues corrections to the Articulator Map (motor cortex; see Nozari, 2022, for embedding in the broader production process).

In this manner, DIVA explains how sensory feedback from one's own speech is used to adjust articulation. Evidence for this mechanism (and its neural correlates) comes from compensatory articulation adjustments in response to perturbations of auditory feedback from one's own voice (e.g., Houde & Jordan, 1998; Tourville, Reilly, & Guenther, 2008). While DIVA focuses on *within-individual* adjustments, phonetic convergence to other people's speech raises the possibility that DIVA's framework could be extended to cover the interaction between production and perception *between a speaker and a listener*. Potentially, exposure to another's utterance may shift the Auditory Target Map in a listener's auditory cortex. If a subsequent auditory signal in the Auditory State Map generated by the listener's own utterance does not match the shifted target, the feedback control system would be triggered and subsequent utterances would shift, accordingly. This mechanism would provide an elegant and parsimonious explanation for phonetic convergence.

Yet, despite its elegance, DIVA has been criticized for its reliance on sensory feedback from overt production, as it slows down performance (Hickok, 2012; Nozari, 2025a, 2025b). In contrast, models that rely inherently on internal simulation do not rely nearly as strongly on the end-state of a motor command and its corresponding sensory state (e.g., Houde & Nagarajan, 2011). Evidence for the latter model in adjustments to perturbations to self-produced speech is mixed. While some claim the necessity of overt sensory feedback (Daliri, Chao, & Fitzgerald, 2020), a recent study has shown single-shot adaptation in self-produced speech (Hantzsch, Parrell, & Niziolek, 2022). We test the proposal of internal simulation in the context of adapting to others' speech. If this applies to phonetic convergence, the extent of overt production should not play a critical role in transfer (Experiment 1), and, perhaps even more radically, overt production may be altogether unnecessary for observing transfer (Experiment 2).

Specifically, we use a statistical learning paradigm to manipulate the correlation between two acoustic speech features, voice onset time (VOT) and fundamental frequency (F0). VOT denotes the time between the release of a stop consonant and the start of vocal fold vibrations of the following vowel and is the main acoustic dimension distinguishing consonants such as /b/ (Short VOT) and /p/ (Long VOT). F0 is the physical property akin to pitch. In standard American English, VOT and F0 are highly correlated (Lisker, 1986). We keep this as our Canonical minimal pair, *beer/pier*. We also create a slight accent in a Reverse condition by reversing the correlation between VOT and F0 in the same minimal pair uttered by the same voice. We passively expose participants to short sequences of stimuli drawn either from the Canonical or the Reverse distributions. Based on past work, we expect exposure to the Reverse distribution to rapidly shift the effectiveness of F0 in signaling *beer* versus *pier* (e.g., Idemaru & Holt, 2011; Zhang, Wu, & Holt, 2021; Hodson, DiNino, Shinn-Cunningham, & Holt, 2022). We also expect this effect to transfer to production and create phonetic convergence (Murphy et al., 2024, 2025).

To test the importance of overt production and subsequent auditory feedback, we compare transfer between two groups differing in the number of productions (36 vs. 360). To anticipate the results, reducing production by an order of magnitude does not negatively impact transfer. We then present a large-scale, single-shot experiment to investigate whether *any* production in the Reverse condition is required

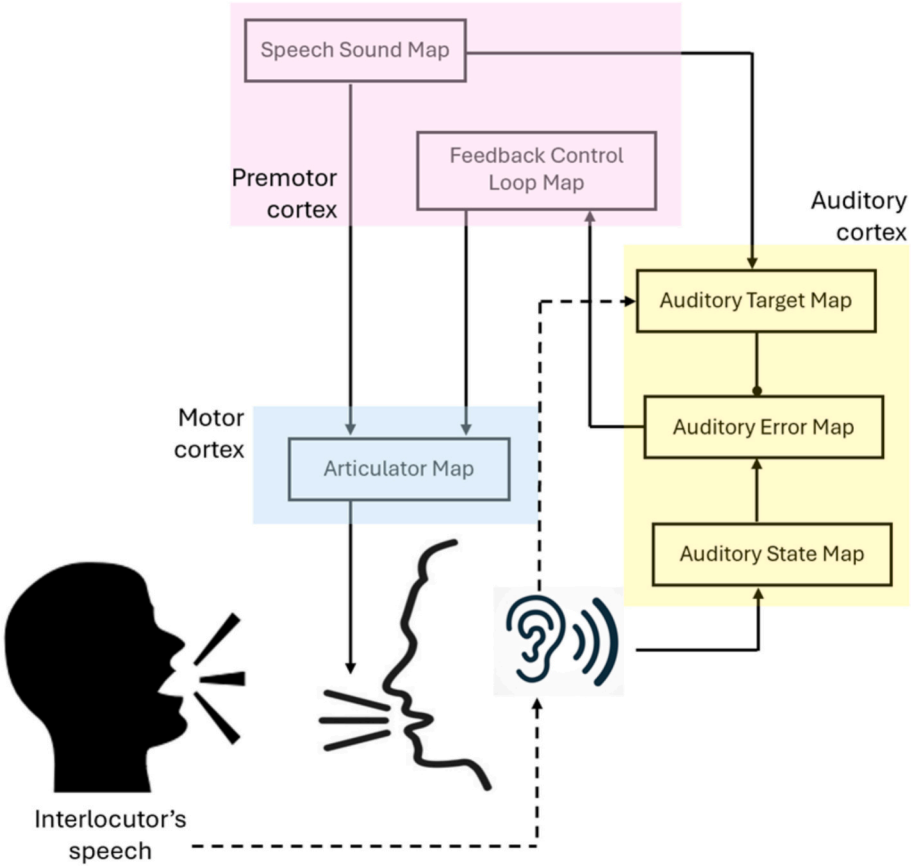

Fig. 1. Regulation of production through auditory feedback in DIVA for self-produced speech and its potential extension to other-produced speech. Adjustments to production are made through a comparison between the Auditory Target Map and Auditory State Map. The Auditory Target Map could be altered by hearing other-produced speech (the dashed-line path). Figure is loosely adapted from Kearney and Guenther (2019).

to observe transfer. If overt production and its corresponding auditory feedback are necessary for articulatory adjustments, there should be no transfer evident in the first Reverse condition production. If, on the other hand, articulation can be adjusted in the absence of overt production and sensory feedback, we would expect to see transfer and phonetic convergence on the very first Reverse condition production.

2. Experiment 1

2.1. Methods

2.1.1. Participants

We estimated the sample size using simulation-based power analyses (SIMR, Green & MacLeod, 2016; Kumle, Võ and Draschkow, 2021), with simulation model parameters informed by data reported in Murphy et al. (2024). A sample size of 48 was needed to detect a small effect size of d

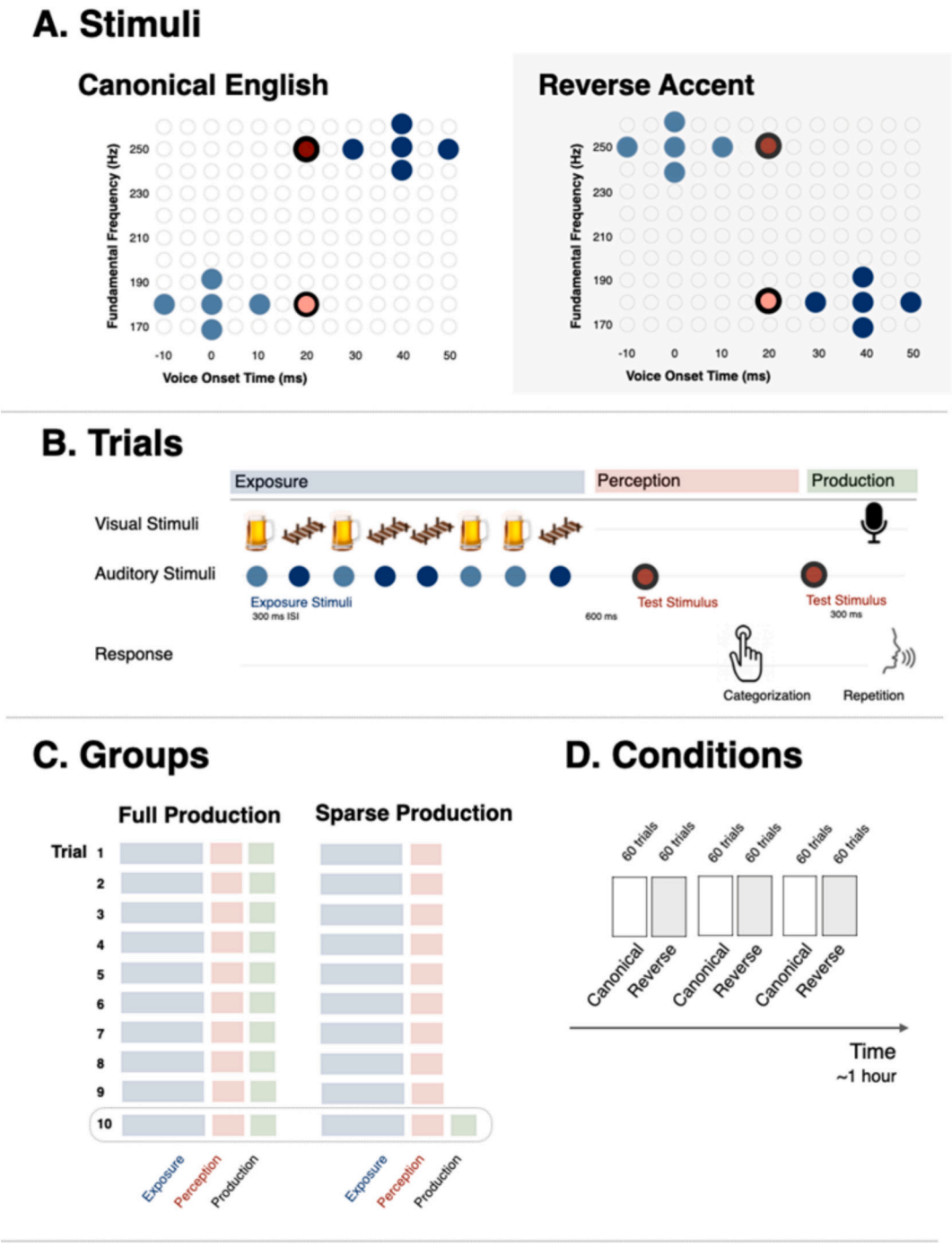

Fig. 2. Study Design. (A) Stimuli. Exposure stimuli are shown in blue for Canonical (left, white background) and Reverse (right, gray background) regularities. Test stimuli are shown in red. (B) Trial Structure. Exposure, Perception, and Production phases of a typical trial. (C) The Full Production group was prompted to produce speech on every trial. Production prompts occurred every tenth trial among the Sparse group. (D) Across ~1 h, participants in both Full and Sparse conditions experienced interleaved Canonical (60 trials) and Reverse (60 trials) speech regularities. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

= 0.2 with power = 0.85 at alpha = 0.05 alpha. To account for possible data loss, we collected data from 56 participants.

Participants were native speakers of American English recruited via Sona Systems at Carnegie Mellon University and Prolific (www.prolific.co), receiving credit or cash, respectively. The study protocol was approved by the Institutional Review Board of Carnegie Mellon University. Four participants were excluded due to poor audio recording of speech productions, leaving a final sample of 52 (39 female, M_{age} = 28.7, SD_{age} = 6.9).

2.1.2. Stimuli

We created a two-dimensional grid of acoustic speech stimuli from utterances digitally recorded in a sound-attenuated booth by an adult female native-American English speaker. One *beer* and one *pier*, chosen for their similarity in duration (385 ms) and F0 contour served as the base stimuli (see Idemaru & Holt, 2020). For each, we identified 15 splice points (~2–3 ms apart, at zero crossings). We removed the interval between *beer* onset and the first splice point and inserted a corresponding interval from *pier*, creating a new stimulus. We repeated this process to arrive at a fine-grained sampling from *beer* to *pier* across VOT (McMurray & Aslin, 2005). From this larger set (light dots, Fig. 2a) Experiment 1 used 0, 10, 20, 30, 40, and 50 ms VOT as stimuli. We created an additional, −10 ms VOT, stimulus by inserting a 10-ms splice of pre-voicing from *beer* before the burst of the 0 ms VOT stimulus.

We next manipulated the fundamental frequency (F0) of this series to create a 2-dimensional acoustic grid, using Praat 6.1 (Boersma & Weenink, 2021) to adjust vowel onset F0 to 170 to 250 Hz in 10-Hz steps. From these initial frequencies, F0 decreased quadratically to 150 Hz at stimulus offset for all stimuli. Finally, we normalized root-mean-squared amplitude across stimuli. In all, this created a densely sampled grid of stimuli varying acoustically in VOT and F0 vowel onset frequency and varying perceptually from *beer* to *pier,* as illustrated in Fig. 2a. We saved stimuli digitally in a lossless format.

2.2. Procedure

We sampled regions of this 2-d acoustic stimulus space in a manner consistent with English speech regularities (Canonical), or with an accent (Reverse). For the Canonical condition, there were 5 Exposure stimuli (Fig. 2a; blue dots) with shorter VOT (−10, 0, 10 ms) and lower F0 (170, 180, 190 Hz) and 5 Exposure stimuli with longer VOT (30, 40, 50 ms) and higher F0 (240, 250, 260 Hz). The 10 Exposure stimuli of the Reverse condition reversed this F0xVOT correlation (Fig. 2a, gray shaded panel). For both Canonical and Reverse conditions, we created sequences of 8 Exposure stimuli by randomly sampling 4 shorter VOT (consistent with *beer*; light blue) and four longer VOT (consistent with *pier;* dark blue) stimuli and ordered them randomly (300 ms silent intervals, 5900 ms total duration; Fig. 2b). As participants passively experienced this sequence of Exposure stimuli, clipart images corresponding to each word appeared synchronized to sound onset. We assessed the influence of Canonical and Reverse speech regularities with two Test Stimuli that each possessed a constant, perceptually ambiguous VOT (20 ms, see Idemaru & Holt, 2020) and either a High (250 Hz) or a Low (180 Hz) F0, as shown by red dots in Fig. 2a. With VOT neutralized, the categorization of these Test stimuli reveals listeners' reliance on F0 in *beer-pier* categorization. As depicted by Fig. 2b, on each trial one of the two Test stimuli followed the sequence of 8 Exposure stimuli after a 600-ms silent interval (with equal probability). Participants categorized the Test stimulus with a keypress to indicate *beer* or *pier* at their own pace. On some trials, after a categorization response was registered, participants heard the same Test stimulus again, and 300 ms later a microphone icon prompted them to say it aloud. They had 2500 ms to orally respond. Utterances were recorded digitally. As illustrated in Fig. 2c, participants in the Full group received production prompts on each trial. Those in the Sparse group received a prompt to produce every tenth trial. This created a 10:1 disparity in speech productions across groups (360 trials Full group, 36 trials Sparse group).

Each participant completed six blocks of 60 trials each, with short breaks after each 20 trials. The blocks were identical, except for exposure, which was sampled randomly from the Canonical stimuli in blocks 1, 3, and 5 and Reverse stimuli in blocks 2, 4 and 6 (Fig. 2d). In each block, 48 trials involved the Test stimuli. The remaining 12 trials used unambiguous test stimuli (*beer:* 0 ms VOT, 180 Hz F0; *pier:* 40 ms VOT, 250 Hz F0) to avoid sole exposure to ambiguous stimuli during test, but these trials were discarded from the analyses.

The study was hosted on Gorilla (www.gorilla.sc, Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2018; Anwyl-Irvine, Dalmaijer, Hodges, & Evershed, 2021). Participants who did not pass a headphone check (Milne et al., 2021) or a microphone check did not proceed to the experiment. Those who proceeded heard diotic presentation of the speech stimuli over headphones and completed the experiment on their own computer.

2.3. Analytical approach

We extracted F0 from recordings and z-score normalized these values on a by-participant basis, as described by Murphy et al. (2024; see also Appendix A).

Statistical analysis involved mixed effects models via the *lme4* package (Bates, 2014 in R (version 4.1.3, R Core Development Team, 2022). We used mixed-effect logistic regression models with a binary response (*beer, pier*) as the dependent variable for perceptual categorization. Fixed effects included Statistical Regularity (Canonical, Reverse), Test Stimulus F0 (Low F0, High F0), and Group (Full, Sparse) alongside 2- and 3-way interactions. All fixed effects were center-coded (−1 or 1). *P*-values were based on Satterthwaite approximates using the *LmerTest* package (version 3.1–3, Kuznetsova, Brockhoff, & Christensen, 2017). The model examining production had the same general structure except that continuous normalized F0 was the dependent variable. Additionally, we replaced the fixed effect of Test Stimulus F0 (High, Low) with Perceptual Response (beer, pier). To better understand this decision, recall that the sequence of events is as follows: stimulus statistics ➔ perceptual change ➔ production change. To examine the change to perception, we use stimulus statistics as the independent variable. If we continue to do that to measure changes to production, we will have the intermediary perceptual change, which means that what we are observing in production could simply be a change to perception. Imagine that, after being exposed to Reverse stimuli, a participant hears a VOT-ambiguous High-F0 stimulus as "beer". She then goes on to produce a low-F0 production. Can we claim that the stimulus statistics have really altered the representational space in production? No, because the perceptual decision tells us that the participant actually intended to say "beer" and correspondingly produced a lower F0 to meet that goal.

We can avoid this challenge by making production F0 contingent not on stimulus statistics, but on the perceptual decision. The scenario is now like this: we have all the trials in Canonical and Reverse conditions, when the participant intended to say "beer". If there is no effect of the stimulus statistics on production, production F0 should be the same regardless of the condition. If, on the other hand, there has been a real change to production, we would expect production F0 for "beer" in the Reverse condition to be higher than that for "beer" in the Canonical condition, even though the intention has been exactly the same in both. In statistical terms, by making production F0 contingent on the perceptual judgment, we are partialing out any effects of stimulus statistics on perception, which allows us to remove a potential confound for the production analysis.

We included the largest random-effect structure tolerated by the model. For each model, analyses collapsed data over the three Canonical blocks and, separately, the three Reverse blocks.

2.4. Results

Code and analyses can be found at https://osf.io/cgp7u/. Full tables appear in Appendix B.

2.4.1. Perceptual categorization

As is clear in the top row of Fig. 3, both Full and Sparse groups exhibit shifts in perceptual categorization as speech input regularities change. The statistical model included a random intercept of subjects, as well as the random slopes of Statistical Regularity, Test Stimulus and their interaction over subjects.

As expected by English norms, High F0 test stimuli were categorized significantly more often as *pier* than *beer* ($z = 12.42$, $p < .001$). This effect interacted with Statistical Regularity ($z = 11.03$, $p < .001$), showing perceptual down-weighting of F0 in categorization responses in the Reverse condition. There was a main effect of Statistical Regularity ($z = -2.15$, $p = .031$) but no significant main effect of Group ($z = -0.67$, $p = .505$). Importantly, none of the interactions with Group was significant, including the critical 3-way interaction (see Table B1). Post-hoc tests showed that the F0 down-weighting was present for both groups (Full: $z = 6.45$, $p < .001$, Sparse: $z = 9.19$, $p < .001$).

2.4.2. Transfer to speech production

We next considered whether these perceptual effects transferred to speech production and, if so, whether transfer was influenced by 10:1 difference in production opportunities across Full and Sparse groups. Recall that this analysis is contingent on participants' perceptual responses (see Analytical Approach for a detailed explanation). The bottom row of Fig. 2 shows the results. The statistical model tolerated only a random intercept over subjects.

Consistent with English norms, productions elicited by responses labeled as *pier* had higher F0 than *beer* (Response: $t = -16.93$, $p < .001$); no other main effects were significant. Notably, the pattern in Response interacted with Statistical Regularity ($t = -5.93$, $p < .001$), showing transfer of perceptual down-weighting of F0 to production. Critically, there was no significant three-way interaction ($t = -1.58$, $p = .115$) despite the 10:1 difference in the number of productions (Table B2). Post-hoc tests revealed significant transfer for each group. There was a significant main effect of Response (Full: $t = -9.51$, $p < .001$; Sparse: $t = -14.77$, $p < .001$) and a significant interaction of Response and Statistical Regularity (Full: $t = -2.87$, $p = .004$; Sparse: $t = -5.70$, $p < .001$) for each group.

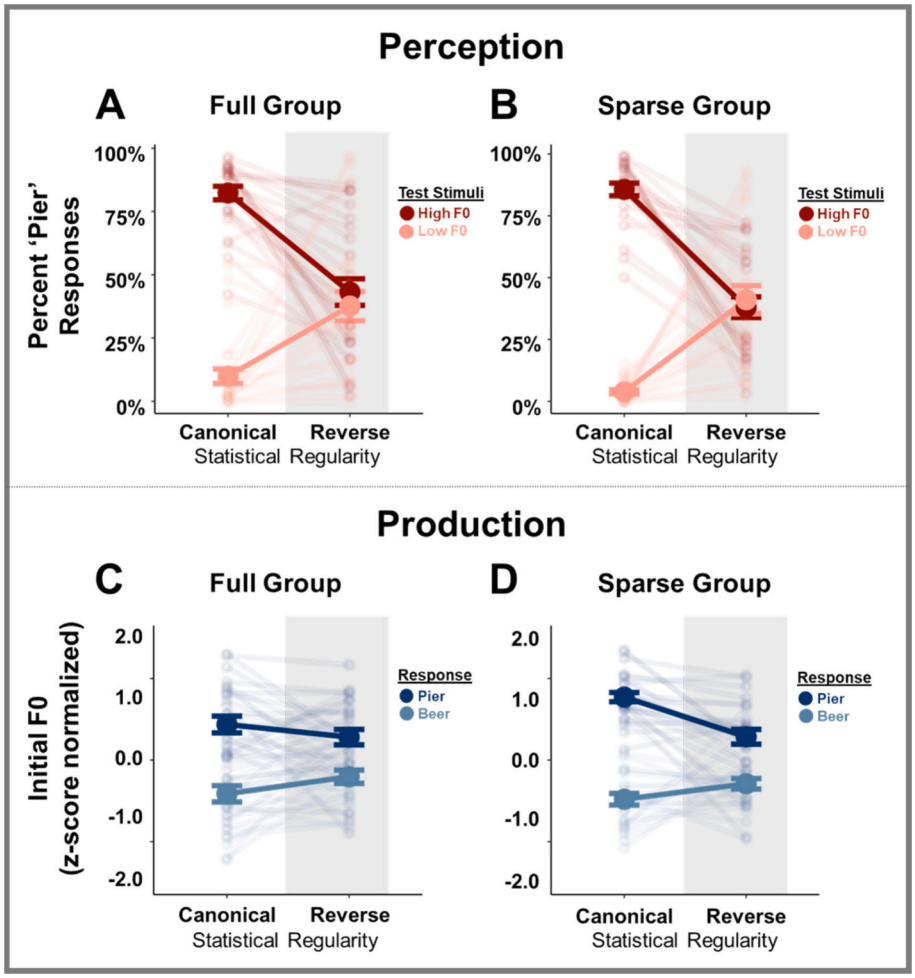

Fig. 3. Results of Experiment 1. (A) Percentage of 'pier' responses to High and Low F0 Test stimuli in the context of Canonical and Reverse statistical regularities for the Full group and (B) Sparse group. (C) F0s measured from participants' productions as a function of participants' perceptual responses, indicated by light blue ("beer") and dark blue ("pier") markers, for the Full group. The opposite slopes show that exposure to stimuli with a reversed American English F0 distribution drives their production F0s toward zero, making them less distinct. (D) F0s measured from participants' productions as a function of participants' perceptual responses for the Sparse group. Averages reflect subject means ± SE. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.5. Discussion

Experiment 1 replicated the past findings of F0 down-weighting in perception in the Reverse condition (e.g., Idemaru & Holt, 2011; Hodson et al., 2022) and its transfer to production (Murphy et al., 2024, 2025). The novel finding was that the production of 360 vs. 36 utterances did not have a significant effect on either perception or its transfer to production. This questions the role of overt production, and the sensory feedback it creates, in transfer of shifts in speech perception to production. The next experiment focuses on this specific question: is *any* overt production required for observing the transfer of statistical learning from perception to production?

3. Experiment 2

3.1. Methods

3.1.1. Participants

We used the same simulation-based approach to power calculation as in Experiment 1, with parameters adjusted to account for the single-shot nature of the Experiment 2 design. This produced a sample estimate of 214 participants to detect a small effect size of d = 0.2 with power = 0.85 at the 0.05 alpha significance level. We collected data from 250 participants on Prolific, anticipating some data loss. Thirty-three participants were excluded due to technical problems, leaving a sample of 217 (135 female, M_{age} = 30.1, SD_{age} = 5.5) participants.

3.1.2. Stimuli

Stimuli were identical to those of Experiment 1.

3.2. Procedures

The Experiment 2 procedure is illustrated in Fig. 4a. The general approach mirrored Experiment 1 except for changes made to accommodate testing whether *the very first production in the Reverse block* would be impacted by exposure to the regularities of the accent, without access to sensory feedback from production. To test this, a single Critical Canonical block (60 trials) preceded a Critical Reverse block (10 trials). Participants categorized the Test stimulus on each trial, with speech productions prompted every tenth trial. This resulted in 6 Canonical condition productions and a single Reverse condition production. To accommodate the single-shot utterance in the Reverse condition, the Test stimulus in this block was manipulated as a between-participant variable, with participants assigned randomly to either the High F0 Test stimulus or the Low F0 Test stimulus groups.

Immediately after the Reverse block, participants completed another block of 60 Normalization Canonical trials with speech production prompted on each trial. This provided 60 additional productions in support of z-score normalizing production F0 on a by-participant basis, as described for Experiment 1. In other ways, online testing proceeded as in Experiment 1.

3.3. Analytical approach

Productions collected across all blocks contributed to F0 normalization as described in Experiment 1. Only trials with production in the first two blocks were included in analyses (trials marked 1–7 in Fig. 4a). Modeling decisions were the same as Experiment 1. Perceptual judgments were modeled as a function of Test Stimulus F0, Statistical Regularity, and their interaction in a mixed-effect logistic regression model. Production F0 was modeled as a function of Perceptual Responses, Statistical Regularity and their interaction in a non-logistic version of the model.

3.4. Results

3.4.1. Perceptual categorization

Fig. 4b shows the results. The model tolerated the random intercept

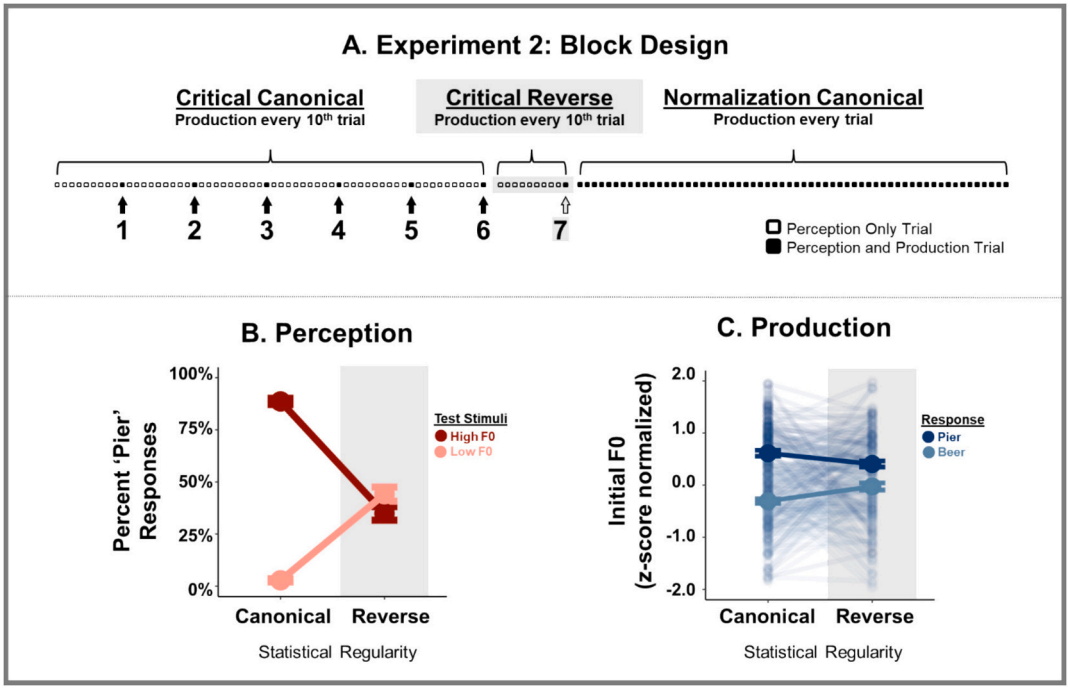

Fig. 4. Experiment 2 Design and Results. (A) Productions were elicited on every tenth trial in a Critical Canonical block (60 trials) and a Critical Reverse block (10 trials). Trials marked 1–7 were included in the analyses. A final Normalization Canonical block (60 trials) elicited productions on every trial for purposes of by-participant F0 normalization, as described in Experiment 1. (B) Perception Results. Percentage of pier responses to High and Low F0 Test stimuli in the context of Canonical and Reverse Statistical Regularities. (C) Production Results. F0s were measured from participants' productions as a function of participants' perceptual responses (Canonical, Trials 1–6; Reverse, Trial 7).

of subjects and the random slope of Statistical Regularity over subjects. As in Experiment 1, there was a main effect of Test Stimulus F0 in the direction expected of native English listeners ($z = 10.81$, $p < .001$). This effect interacted with Statistical Regularity ($z = 12.10$, $p < .001$), showing perceptual down-weighting of F0 in the Reverse condition (Table B3).

3.4.2. Transfer to speech production

The perceptual analyses showed that the 10 trials of the Critical Reverse block were sufficient to shift listeners' reliance on F0 in speech categorization. We next asked whether this shift transferred to influence the very first production (elicited on the final, tenth trial of that Critical Reverse block). Fig. 4c shows the results. The model tolerated a random intercept and random slope for Response over subjects. As in Experiment 1, *pier* F0 was significantly higher than *beer* F0 ($t = -7.96$, $p < .001$). Critically, this effect interacted with Statistical Regularity ($t = -2.98$, $p = .003$), showing transfer (Table B4).

3.5. Discussion

The perceptual shift from encountering the reversed regularity of the accent transferred to influence the very first overt production. The auditory feedback associated with overt production was not necessary for adjustments to speech production to be driven by perceptual shifts in speech perception.

4. General discussion

The relationship between perception and production has been of interest to language researchers for years, but it is now understood to be far from simple (see Baese-Berk, Kapnoula, & Samuel, 2024, for a review). The current study set out to test the role of overt production and its subsequent auditory feedback in phonetic convergence. We replicated prior findings by showing that the statistical regularities of incoming speech shift speech categorization and that this change transfers to affect production (Murphy et al., 2024, 2025). The novel finding was the independence of this transfer from overt production and its perceptual consequences. Experiment 1 showed that decreasing production attempts by an order of magnitude did not adversely affect transfer. Experiment 2 went further to show that phonetic convergence was evident on the very first overt production after exposure to a slight accent. This is important for several reasons.

First, the most well-accepted model of speech motor control, i.e., DIVA (e.g., Guenther, 2016; Meier & Guenther, 2023), relies on overt production and the sensory feedback it provides. Our findings show that such reliance is unnecessary. Instead, the current findings are better aligned with models that posit dynamic simulation of internal states as a means of adjusting the production command. An example of such a model is the state feedback control (SFC) models (e.g., Houde & Nagarajan, 2011). SFC shares its basic structure with DIVA. In both models, issuing a motor command is accompanied by predicting its sensory outcomes. But there are key differences between the two models: DIVA relies on the actual motor command and the end-state sensory consequences of its execution. This end-state sensory consequence is compared with the estimated consequence, and the correction is applied to the next motor command. In contrast, SFC, rather than relying on end-states, continuously models the trajectory of a motor command and its corresponding sensory states. In addition, an internal simulation of that motor command and its sensory consequences is generated in parallel. It is the real-time comparison between the sensory states estimated by the actual and the simulated motor commands that provides corrective feedback to the motor command. Since this process is continuous and dynamic, correction need not await the completion of the motor command and can thus be applied to the same trial.

The second reason for the importance of the current findings is its ties to a broader literature on the importance of the engagement of the production system in the transfer of statistical learning from perception to production beyond the acoustic-phonetic domain. For example, Kittredge and Dell (2016) reported that simple auditory exposure to new artificial phonotactic constraints (e.g., /s/ can only be an onset) was not sufficient for a speaker to demonstrate the same constraints in their own speech, even though such constraints are quickly learned when speakers produce them (e.g., Warker & Dell, 2006; see also Atilgan & Nozari, 2025, for generalization to other language modalities). Interestingly, the authors reported learning in an intermediate condition, when participants were given a task that required them to predict the upcoming auditory syllables. This finding was interpreted as the engagement of the production system through the act of prediction. Our claims agree with Kittredge and Dell's (2016) in showing that overt production is not always necessary for the transfer of learning between perception and production systems, and that prediction in the production system is involved in driving the transfer to production. However, unlike Kittredge and Dell, there is no active task requiring individuals to make predictions. We can thus show that rapid transfer of statistical learning between perception and production is possible even without the intentional engagement of the production system.

Third, the findings of the study speak more generally to the concept of "alignment" in language production, i.e., the notion that listeners align their own production to that of their interlocutors, at all levels of production (phonetic, phonological, lexical, syntactic, and semantic; Pickering and Garrod, 2004; Pickering & Garrod, 2013). Interestingly, there are some discrepancies in literature, especially at the phonetic level, where some studies find convergence, while others do not. One may blame methodological differences, but in truth, the variability in results can be observed across a wide range of methodologies, including free-form or semi-structured conversations (e.g. Gregory et al., 2001; Levitan & Hirschberg, 2011; Natale, 1975; Pardo et al., 2012), auditory repetition in shadowing tasks (e.g. Babel, 2012; Honorof, Weihing, & Fowler, 2011; Pardo, Jordan, Mallari, Scanlon, & Lewandowski, 2013; Shockley, Sabadini, & Fowler, 2004), and more controlled experimental tasks (e.g. Dias & Rosenblum, 2011; Kim, Horton, & Bradlow, 2011; Pardo, Urmanche, Wilman, & Wiener, 2017). The current results shed some light on this discrepancy. On the one hand, they provide strong support for one of the key claims of the alignment account, namely, the closely interwoven nature of perception and production. On the other hand, alignment accounts often emphasize (covert) "imitation" as a critical underlying mechanism. The current demonstration reframes the observed alignment as changes to information processing rather than covert imitation. In fact, individuals' speech did *not* converge, in the sense of a simple imitation; participants did not produce utterances with a reversed VOTxF0 correlation to match what they heard. Rather, the unusual but systematic reversal of VOTxF0 correlation in input caused the perceptual system to process information differently, i.e., to down-weight F0 (see Wu & Holt, 2022). This change in information processing updated how error signals are computed, which in turn affected production, without necessarily involving any acts of imitation.

Finally, the reader may wonder if the malleability shown here in the production system is compatible with a stable production system. Recall that phonemes have some degree of variability, which is why speech motor models, like DIVA, model them as regions as opposed to points. Production is stable as long as it does not easily swerve into a phoneme category different from the intended target. By making our production analyses contingent on perceptual responses, we first determine the target phoneme and then demonstrate the variability in F0 production within that target zone. In that sense, our results show that the system is quite stable. For a similar reason, we caution the reader against the temptation of extrapolating these results to learning novel accents, as the current demonstration does not extend to creating new phoneme categories or novel motor plans to execute them.

To summarize, perception can affect production without the overt or intentional engagement of the production system, calling for new ways of looking at one of the oldest and most fundamental questions in

cognitive science: how perception affects action.

CRediT authorship contribution statement

Timothy K. Murphy: Writing – review & editing, Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lori L. Holt:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Funding acquisition, Conceptualization. **Nazbanou Nozari:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors have no relevant financial or non-financial interests to disclose.

Acknowledgments

This work was funded by the NSF grant BCS-2346989 to N.N. and L. L.H. T.M. was supported by the 2022 Raymond H. Stetson Scholarship in Phonetics and Speech Science awarded by the Acoustical Society of America and is currently supported by the University of Wisconsin Voice Research Training Porgram (T32DC009401, awarded institutionally to Susan Thibeault).

Appendix A

F0 measurement and normalization

We extracted F0 from recordings of the productions using a custom Praat (version 6.1; Boersma and Weenink, 2021) and R (version 4.1.3, R Core Development Team, 2022) processing pipeline developed by Murphy, Nozari & Holt (2024). In Praat, "To TextGrid (silences)..." identified and isolated word productions in the 2.5 s audio recordings. Then, "To Pitch (ac)" characterized the F0 frequency of the first 40 ms of voicing, where F0 differences between onset obstruent consonants are typically most pronounced (Hanson, 2009; Hombert, Ohala, & Ewan, 1979; Lea, 1973; Xu & Xu, 2021). Next, we log-transformed F0 frequency and removed outliers +/− 3 standard deviations relative to a participant's mean F0 from further analyses. Finally, we accounted for the F0 variability across talkers impacted by multiple factors, including sex (Titze, 1989), by z-score normalizing F0 frequency on a by-participant basis. This yielded a measure for which 0 indicates the mean F0 for a participant across all productions. Values of ±1 indicate a standard deviation above or below the mean. These normalized measurements entered group analyses.

Appendix B

Full results of the main analyses in Experiments 1 and 2

Table 1
Experiment 1: Perceptual categorization.

Predictor	*β*	*SE*	*z*	*p*
(Intercept)	−0.64	0.11	−5.85	<0.001
Statistical Regularity	−0.21	0.10	−2.15	0.031
Test Stimulus F0	2.72	0.22	12.42	<0.001
Group	−0.14	0.21	−0.67	0.505
Statistical Regularity:Test Stimulus F0	5.33	0.48	11.03	<0.001
Statistical Regularity:Group	−0.26	0.17	−1.56	0.119
Test Stimulus F0:Group	0.41	0.43	0.97	0.334
Statistical Regularity:Test Stimulus F0:Group	1.65	0.93	1.77	0.077

Note: Reference levels are Condition (Reverse), Test Stimulus F0 (Low F0), Group (Full).

Table 2
Experiment 1: Transfer to speech production.

Predictor	*β*	*SE*	*t*	*p*
(Intercept)	0.04	0.02	1.79	0.074
Statistical Regularity	0.08	0.04	1.72	0.086
Response	−0.76	0.04	−16.93	<0.001
Group	0.04	0.04	0.90	0.367
Statistical Regularity:Response	−0.53	0.09	−5.93	<0.001
Statistical Regularity:Group	0.17	0.09	1.94	0.052
Response:Group	−0.23	0.09	−2.51	0.012
Statistical Regularity:Response:Group	−0.28	0.18	−1.58	0.115

Note: Reference levels are Condition (Reverse), Response (Pier), Group (Full).

Table 3
Experiment 2: Perceptual categorization.

Predictor	***β***	***SE***	***t***	***p***
(Intercept)	−0.61	0.11	−5.54	< 0.001
Statistical Regularity	−0.35	0.22	−1.60	0.11
Test Stimulus F0	3.25	0.30	10.81	< 0.001
Statistical Regularity:Test Stimulus F0	7.26	0.60	12.10	<0.001

Note: Reference levels are Statistical Regularity (Reverse), Test Stimulus F0 (Low F0).

Table 4
Experiment 2: Transfer to production.

Predictor	*β*	*SE*	*t*	*p*
(Intercept)	0.18	0.04	4.32	< 0.001
Statistical Regularity	−0.03	0.08	−0.45	0.656
Response	−0.68	0.09	−7.96	< 0.001
Statistical Regularity:Response	−0.45	0.15	−2.98	0.003

Note: Reference levels are Statistical Regularity (Reverse), Response (Pier)

Data availability

Data are available on OSF https://osf.io/cgp7u/

References

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods, 53*(4), 1407–1425.

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2018). Gorillas in our midst: Gorilla. sc. *Behavior Research Methods, 52*(2020), 388–407.

Atilgan, N., & Nozari, N. (2025). Statistical learning of orthotactic constraints: Evidence from typing. *Journal of Experimental Psychology: Learning, Memory, and Cognition. Advance online publication.*. https://doi.org/10.1037/xlm0001502

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics, 40*(1), 177–189.

Baese-Berk, M. M., Kapnoula, E. C., & Samuel, A. G. (2024). The relationship of speech perception and speech production: It's complicated. *Psychonomic Bulletin & Review*, 1–17.

Bates, D. (2014). Fitting linear mixed-effects models using lme4. In *arXiv preprint arXiv: 1406.5823*.

Boersma, P., & Weenink, D. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1. retrieved from http://www.praat.org/.

Daliri, A., Chao, S. C., & Fitzgerald, L. C. (2020). Compensatory responses to formant perturbations proportionally decrease as perturbations increase. *Journal of Speech, Language, and Hearing Research, 63*(10), 3392–3407.

Dias, J. W., & Rosenblum, L. D. (2011). Visual influences on interactive speech alignment. *Perception, 40*, 1457–1466.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498.

Gregory, S. W., Jr., Green, B. E., Carrothers, R. M., Dagan, K. A., & Webster, S. W. (2001). Verifying the primacy of voice fundamental frequency in social status accommodation. *Language & Communication, 21*(1), 37–60.

Guenther, F. H. (2016). *Neural control of speech*. MIT Press.

Hanson, H. M. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. *The Journal of the Acoustical Society of America, 125*(1), 425–441.

Hantzsch, L., Parrell, B., & Niziolek, C. A. (2022). A single exposure to altered auditory feedback causes observable sensorimotor adaptation in speech. *Elife, 11*, Article e73694.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*, 135–145.

Hodson, A., DiNino, M., Shinn-Cunningham, B., & Holt, L. L. (2022). Dimension-based statistical learning in older adults. In *Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 44, No. 44)*.

Hombert, J. M., Ohala, J. J., & Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language*, 37–58.

Honorof, D. N., Weihing, J., & Fowler, C. A. (2011). Articulatory events are imitated under rapid shadowing. *Journal of Phonetics, 39*(1), 18–38.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279*(5354), 1213–1216.

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience, 5*, 82.

Huffaker, K., Holt, L. L., & Nozari, N. (2025). *Transfer of Statistical Learning from Speech Perception to Production Generalizes to Reading*. https://doi.org/10.31234/osf.io/ng6cu_v1

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance, 37*(6), 1939–1956.

Idemaru, K., & Holt, L. L. (2020). Generalization of dimension-based statistical learning. *Attention, Perception, & Psychophysics, 82*, 1744–1762.

Kearney, E., & Guenther, F. H. (2019). Articulating: The neural mechanisms of speech production. *Language, Cognition and Neuroscience, 34*(9), 1214–1229.

Kim, M., Horton, W. S., & Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology, 2*(1), 125–156.

Kittredge, A. K., & Dell, G. S. (2016). Learning to speak by listening: Transfer of phonotactics from perception to production. *Journal of Memory and Language, 89*, 8–22.

Kumle, L., Võ, M. L. H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods, 53*(6), 2528–2543.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1–26.

Lea, W. A. (1973). Segmental and suprasegmental influences on fundamental frequency contours. *Consonant Types and Tone, 1*, 15–70.

Levitan, R., & Hirschberg, J. B. (2011). *Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions*.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling/b/ versus/p/in trochees. *Language and Speech, 29*(1), 3–11.

McMurray, B., & Aslin, R. N. (2005). Infants are sensitive to within- category variation in speech perception. *Cognition, 95*(2), B15–B26.

Meier, A. M., & Guenther, F. H. (2023). Neurocomputational modeling of speech motor development. *Journal of Child Language, 50*(6), 1318–1335.

Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Networks, 9*(8), 1265–1279.

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods, 53*(4), 1551–1562.

Murphy, T. K., Nozari, N., & Holt, L. L. (2024). Transfer of statistical learning from passive speech perception to speech production. *Psychonomic Bulletin & Review, 31*(3), 1193–1205.

Murphy, T. K., Nozari, N., & Holt, L. L. (2025). Bears don't always mess with beers: Limits on generalization of statistical learning in speech. *Psychonomic Bulletin & Review*, 1–12.

Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology, 32*(5), 790.

Nozari, N. (2022). Neural basis of word production. In L. R. Gleitman, A. Papafragou, & J. C. Trueswell (Eds.), *The Oxford handbook of the mental lexicon* (pp. 552–574). Oxford: Oxford University Press.

Nozari, N. (2025a). Monitoring, control and repair in word production. *Nature Reviews Psychology, 4*(3), 222–238.

Nozari, N. (2025b). The relationship between monitoring, control, conscious awareness and attention in language production. *Journal of Neurolinguistics, 74*, Article 101247.

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics, 40*(1), 190–197.

Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language, 69*, 183–195.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics, 79*, 637–659.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2), 169–190.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(4), 329–347.

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics, 66*(3), 422–429.

Thorburn C., Zhou L., Dick F., Nozari N., and Holt L.L. (in press). Speech Motor Control is Not Sequestered from General Auditory Processes. Journal of Experimental Psychology: General.

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America, 85*(4), 1699–1707.

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language & Cognitive Processes, 26*(7), 952–981.

Tourville, J. A., Reilly, K. J., & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage, 39*(3), 1429–1443.

Warker, J. A., & Dell, G. S. (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(2), 387–398.

Wu, Y. C., & Holt, L. L. (2022). Phonetic category activation predicts the direction and magnitude of perceptual adaptation to accented speech. *Journal of Experimental Psychology: Human Perception and Performance, 48*(9), 913–925.

Xu, Y., & Xu, A. (2021). Consonantal F0 perturbation in American English involves multiple mechanisms. *The Journal of the Acoustical Society of America, 149*(4), 2877–2895.

Zhang, X., Wu, Y. C., & Holt, L. L. (2021). The learning signal in perceptual tuning of speech: Bottom-up versus top-down information. *Cognitive Science, 45*(3), Article e12947.