Transfer of Statistical Learning from Passive Speech Perception to Speech Production

Timothy K. Murphy[1,2], Nazbanou Nozari[3] , and Lori L. Holt[4]

1. Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213
2. Center for the Neural Basis of Cognition, Pittsburgh, PA 15213
3. Department of Psychological and Brain Sciences, Indiana University, IN 47405
4. Department of Psychology, University of Texas at Austin, Austin, TX, 78712, USA

Corresponding Author.

Timothy K. Murphy

Department of Psychology, Carnegie Mellon University, Baker Hall, Floor 3, Frew St, Pittsburgh, PA 15213

Email: tkmurphy@andrew.cmu.edu

Acknowledgments. This work was supported by funding from the National Science Foundation BCS-1941357 to LH and by BCS-2217415 to NN and LH. TM was supported by the Predoctoral Training Program in Behavioral Brain Research (T32GM081760, awarded institutionally to LH and Julie Fiez). Christi L. Gomez and Erin D. Smith provided critical support with study recruitment and data collection. Emril Radoncic and Reva Prabhune were essential in piloting and troubleshooting. David Plaut provided helpful feedback.

Abstract

Communicating with a speaker with a different accent can affect one's own speech. Despite the strength of evidence for perception-production transfer in speech, the nature of transfer has remained elusive, with variable results regarding the acoustic properties that transfer between speakers and the characteristics of the speakers who exhibit transfer. The current study investigates perception-production transfer through the lens of statistical learning across passive exposure to speech. Participants experienced a short sequence of acoustically variable minimal pair (beer/pier) utterances conveying either an accent or typical American English acoustics, categorized a perceptually ambiguous test stimulus, and then repeated the test stimulus aloud. In the Canonical condition, /b/-/p/ fundamental frequency (F0) and voice onset time (VOT) covaried according to typical English patterns. In the Reverse condition, the F0xVOT relationship reversed to create an 'accent' with speech input regularities atypical of American English. Replicating prior studies, F0 played less of a role in perceptual speech categorization in Reverse compared to Canonical statistical contexts. Critically, this down-weighting transferred to production, with systematic down-weighting of F0 in listeners' own speech productions in Reverse compared to Canonical contexts that was robust across male and female participants. Thus, the mapping of acoustics to speech categories is rapidly adjusted by short-term statistical learning across passive listening and these adjustments transfer to influence listeners' own speech productions.

Keywords: Statistical Learning**,** Speech Perception, Speech Production, Phonetic Cue Weighting, Phonetic Convergence, Auditory Word Repetition

The close interaction of speech perception and production is undeniable. Perception of *one's own* speech influences speech production (e.g., Guenther, 1994; Bohland, Bullock, & Guenther, 2010). For example, altering speech acoustics and feeding speech back to a talker with minimal delay results in rapid compensatory alterations to productions that are predictable, replicable, and well-accounted for by neurobiologically plausible models of speech production (e.g., Guenther, 2016; Houde & Jordan, 1998).

Similarly, perception of *another talker's* speech can influence production. Talkers imitate sublexical aspects of perceived speech in speech shadowing tasks (Fowler et al. 2003; Goldinger 1998; Shockley et al. 2004) and phonetically converge to become more similar to a conversation partner (Pardo et al. 2017). However, results are variable and hard to predict. Shadowers imitate lengthened voice onset times (VOT), but not shortened VOTs (Lindsay et al. 2021; Nielsen 2011; but see also Schertz & Pacquette-Smith, 2023). Phonetic convergence occurs only for some utterances or some acoustic dimensions, but not others (Pardo et al., 2013). Talkers may converge across some dimensions but diverge on others (Bourhis & Giles 1977; Earnshaw 2021; Heath 2015), making it difficult to predict which articulatory-phonetic dimensions will be influenced (Ostrand & Chodroff, 2021). Phonetic convergence is also variable across talkers' sex (Pardo et al., 2017), with some studies reporting greater convergence among female participants (Namy et al., 2002), others among males (Pardo, 2006; Pardo et al., 2010), or more complicated male-female patterns of convergence (Miller et al., 2010; Pardo et al., 2017). In sum, the direction and magnitude of changes in speech production driven by perceived speech are dependent on multiple contributors (Pardo 2006, Babel 2010) likely to include social and contextual factors (Bourhis & Giles, 1977; Giles, et al. 1991; Pardo 2006). This has made it challenging to characterize production-perception interactions fully.

Some have argued that a better understanding of the cognitive mechanisms linking speech perception and production will meet this challenge (Babel, 2012; Pardo, 2022). Here, we propose an approach that is novel in two ways: (1) *Statistical learning*. Instead of investigating phonetic convergence at the level of individual words, we manipulate the statistical relationship of two acoustic dimensions, fundamental frequency (F0) and voice onset time (VOT) and study the effect of perceptual statistical learning across these dimensions on listeners' own speech. (2) *Subtlety and implicitness*. Acoustic manipulation of the statistical regularities of speech input is barely perceptible and devoid of socially discriminating information, since it is carried on the same voice, therefore allowing us to investigate the basic perception-production transfer without influence of additional (important, but potentially complicating) sociolinguistic factors.

Our approach builds on the well-studied role of statistical learning in speech perception. *Dimension-based statistical learning* tracks how the effectiveness of acoustic speech dimensions in signaling phonetic categories varies as a function of short-term statistical regularities in speech input (Idemaru & Holt, 2011, 2014, 2020; Idemaru & Vaughn, 2020; Liu & Holt, 2015; Lehet & Holt, 2017; Schertz et al., 2015; Schertz & Claire, 2020; Zhang & Holt, 2018; Zhang, Wu, & Holt, 2021). This simple paradigm parametrically manipulates acoustic dimensions, for example voice onset time (VOT) and fundamental frequency (F0), across a two-dimensional acoustic space to create speech stimuli varying across a minimal pair (*beer-pier).* The paradigm selectively samples stimuli to manipulate short-term speech regularities, mimicking common communication challenges like encountering a talker with an accent that deviates from local norms. Across Exposure stimuli (Figure 1A, B, red) the short-term input statistics either match the typical F0xVOT correlation in English (Canonical condition, e.g., with higher F0s and longer VOTs for *pier*) or

introduce a subtle and barely detectable 'accent' with a short-term F0xVOT correlation opposite of that typically experienced in English (Reverse condition, e.g., lower F0s with longer VOTs for *pier*).

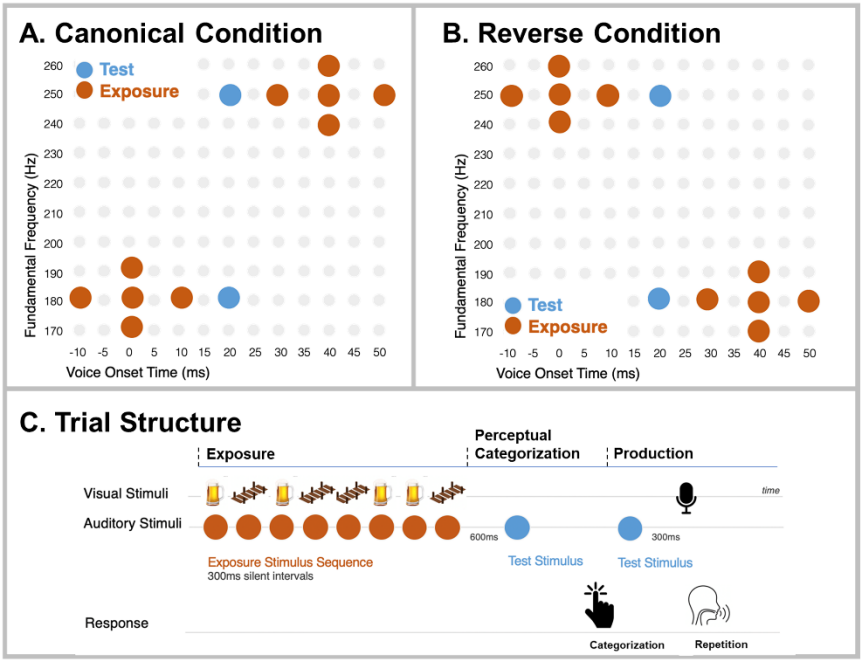

Figure 1. Stimulus and Trial Structure. A. Canonical Distribution. B. Reverse Distribution. The Test stimuli (blue) have ambiguous VOT and are identical across Canonical and Reverse conditions. C. Trial Structure. *Exposure phase:* Participants listened passively to 8 Exposure stimuli, each paired with a visual stimulus. *Perceptual Categorization phase*: After 600 ms they heard one of two Test stimuli with Low or High F0 and categorized it as *beer* or *pier*. *Repetition phase*: they heard the same Test stimulus again and repeated it aloud.

Test stimuli are constant across conditions (Figure 1A, B, blue). They have a neutral, perceptually ambiguous VOT thereby removing this dominant acoustic dimension from adjudicating category identity. But F0 varies across Test stimuli. Therefore, the proportion of Test stimuli categorized as *beer* vs. *pier* provides a metric of the extent to which F0 is perceptually weighted in categorization as a function of experienced short-term speech input regularities (Wu & Holt, 2022).

Although the manipulation of short-term input statistics is subtle and unbeknownst to the listeners, the Exposure regularity rapidly shifts the perceptual weight of F0 in *beer-pier* Test stimulus categorization (Idemaru & Holt, 2011). Listeners down-weight F0 reliance upon introduction of the accent. This effect is fast and robust against the well-known individual differences in perceptual weights and the variability with which individuals perceptually weight different acoustic dimensions (Kong & Edwards, 2011, 2016; Schertz et al., 2015, 2016). In all, this well-replicated finding (1) demonstrates reliable changes in the perceptual system as a function of brief exposure to subtle changes in the statistical properties of the acoustic input and (2) establishes a statistical learning paradigm as an ideal tool for examining the impact of these changes on speech production.

In the current study, we used dimension-based statistical learning to investigate whether adjustments to the perceptual space influence speech production in systematic ways. Following Hodson and colleagues (2023), participants passively experienced short sequences of *beer* and *pier* Exposure stimuli sampling Canonical or Reverse distributions followed by one of the two F0-differentiated Test stimuli. They categorized the Test stimulus as *beer* or *pier*, then heard it again and repeated it aloud (Figure 1C). If production is rapidly adjusted to the change in the perceptual space evoked by passive listening across statistically structured sequences of sound, we predict a down-weighting of *production* F0 in the Reverse (compared to the Canonical) condition. Secondarily, we examine both perception and production effects separately in male and female participants to assess whether the adjustment is influenced by participant sex.

Methods

Participants

Although previous studies which have used this experimental paradigm have found large effect sizes for dimension-based statistical learning in *perception*, we do not have a prior effect size for potential dimension-based statistical learning in *production*. Assuming an effect size of 0.45 with alpha of 0.05 and power of 0.8, in a within-subject design, we would need 41 participants. Because a secondary goal of this project is to assess the effect separately in male and female participants, we doubled this sample size. To allow for possible attrition, we set the target sample size of 45 male and 45 female participants.

Ninety participants (45 females) were recruited using Prolific (www.prolific.co), an online participant enrollment tool. Sex was determined by participants' responses to the question: "What sex were you assigned at birth, such as on an original birth certificate?" In answer to a question regarding gender, 45% of participants identified as cisgender female, 48% identified as cisgender male, and 7% identified as non-binary. Here, we used the biological variable sex.

The study was conducted under a protocol approved by the Institutional Review Board at Carnegie Mellon University. All participants were adult native-English speakers located within the United States, ages 18 to 40 years old (M_{age} = 28.6, SD = 6 years), and compensated at an hourly rate of $10. Following data collection, three (two female) participants were removed due to poor quality audio recordings.

Stimuli

Acoustic stimuli were based on natural utterances of *beer* and *pier* spoken by an adult female native English speaker digitally recorded in a sound attenuated booth, as described by Idemaru and Holt (2020). All stimuli were derived from two initial recordings, one *beer* and one *pier*, chosen for their similarity in duration (385 ms) and F0 contour. Following the approach of McMurray and Aslin (2005), we identified 15 splice points (~2-3 ms apart, at zero crossings) in both recordings. Then, we removed the interval between *beer* onset and the first splice point and inserted a corresponding interval from the *pier*, creating a new stimulus along the VOT series. Repeating this process resulted in a fine-grained series of syllables varying in VOT from *beer* to *pier* in approximately 2-3 ms steps. From this series, syllables with VOTs of 0, 10, 20, 30, 40, and 50 ms served as stimuli. An additional stimulus with -10 ms VOT was created by taking a splice of pre-voicing from *beer* and inserting it before the burst of the 0 ms VOT *beer.*

Next, we manipulated the fundamental frequency (F0) across the VOT series to create a 2-dimensional F0xVOT acoustic space, with adjustment of the F0 onset frequency (170-250 Hz in 10-Hz steps) at vowel onset manipulated manually using Praat 5.3 (Boersma & Weenink, 2017). The F0 contour decreased quadratically to 150 Hz at stimulus offset. Stimuli were normalized to the same root mean-squared amplitude.

We sampled three types of stimuli from the F0xVOT acoustic space. Exposure stimuli conveyed a specific F0xVOT short-term regularity (Canonical, Reverse) across passive listening (Figure 1C, Exposure). They possessed unambiguous VOTs diagnostic of /b/ (-10, 0, 10 ms) and /p/ (30, 40 ,50 ms) and F0 frequencies spanning 170, 180, 190, 240, 250, 260 Hz. The Canonical condition stimuli (Figure 1A, red) were sampled to exhibit the typical English F0xVOT relationship (Abramson & Lisker 1986) with *beer* associated with shorter VOT (-10, 0, 10 ms) and lower F0 (170, 180, 190 Hz) and *pier* associated with longer VOT (30, 40, 50 ms) and higher F0 (240, 250, 260 Hz). The Reverse condition stimuli (Figure 1B, red) reversed this F0xVOT correlation; shorter VOTs consistent with *beer* were paired with higher F0s and longer VOTs signaling *pier* were paired with lower F0s. We constructed each trial as a sequence of four *beer* (short VOT) and four *pier* (long VOT) stimuli randomly selected from either the Canonical or Reverse distributions, and randomly ordered with 300-ms inter-stimulus silent intervals (Figure 1C).

Test stimuli (Figure 1, blue) served as both the probe for perceptual categorization and elicitation of speech production in the auditory repetition task. Test stimuli possessed a constant, perceptually ambiguous VOT (20 ms, see Idemaru & Holt, 2020) and either a High F0 (250 Hz) or a Low F0 (180 Hz) (Figure 1A, B; blue). Two stimuli with unambiguous VOTs and High or Low F0s (*beer:* 0 ms VOT, 180 Hz F0; *pier*: 40 ms VOT, 250 Hz F0). Forty-eight trials with unambiguous test stimuli were included to ensure participants did not perceive only unusual sounding probes.

Procedure

Online participants recruited via Prolific were automatically directed to the experiment, hosted on the online experimental platform Gorilla (www.gorilla.sc, Anwyl-Irvine et al., 2018, 2021). Participants were required to use the Chrome browser and all speech was presented in lossless FLAC format. Participants first completed consent and a simple demographics survey and then underwent a brief psychophysical check for compliance in wearing headphones using the dichotic Huggins pitch approach (Milne et al., 2020). Participants who did not pass the headphone check did not proceed to the experiment. Subsequently, a microphone check confirmed that participants' browsers and microphones were recording speech utterances.

The experiment then commenced, expanding the perceptual protocol of Hodson et al. (2023) to examine transfer to production. Participants were instructed about the trial structure via written instructions. As illustrated in Figure 1C, each trial had three phases: *Exposure*, *Perceptual Categorization*, and *Repetition*. In the exposure phase participants passively listened to a sequence of 8 Exposure stimuli (4 short VOT <15 ms signaling *beer* and 4 long VOT >25 ms signaling *pier,* randomly ordered) separated by 300 ms of silence (5900 ms total duration). As stimuli played diotically over headphones corresponding clipart images (*beer* for <15 ms VOT, *pier* for >25 ms VOT) appeared, synchronized to sound onset. The next phase, Perceptual Categorization, began with 600 ms of silence. Participants then heard a Test stimulus with perceptually ambiguous VOT (20 ms) and either Low (180 Hz) or High (250 Hz) F0 and categorized it as *beer* or *pier* via a keyboard response guided by onscreen text indicating the key/response correspondence as well as a question mark to indicate the need to respond. The Repetition phase began immediately after response. Participants heard the same Test stimulus and, 300 ms later, saw an image of a microphone that signaled them to repeat the Test stimulus

aloud. Participants' utterances were recorded over their own computer microphone and stored digitally as .weba files.

The Perceptual Categorization and Repetition phases were identical across blocks. Blocks differed in the distinctive (Canonical, Reverse) short-term regularities conveyed by the Exposure phase. The first block was always Canonical, with subsequent blocks alternating between Reverse and Canonical blocks. This resulted in 248 test trials (124 Canonical, 124 Reverse; blocks of 40-42 trials) presented across six blocks. Two of the three Canonical blocks were composed of 41 trials while the third was composed of 42 trials. A small programming discrepancy led to two of the three Reverse blocks having 42 trials whereas the third had 40 trials.

Among the 248 test trials, 200 trials (100 Canonical, 100 Reverse) presented Ambiguous Test stimuli to assay dimension-based statistical learning in perception and its transfer to production. The remaining 48 trials (24 Canonical, 24 Reverse) presented Unambiguous stimuli so that participants did not perceive only unusual sounding probes. Ambiguous and Unambiguous stimuli were randomized within condition (Canonical, Reverse). Participants had 15-sec breaks after each 15 trials and between blocks.

Production F0 Measurements

We designed custom Praat and R scripts to extract F0 from the speech productions. In Praat (version 5.3), "To TextGrid (silences)..." identified and isolated word productions in the 2.5 second audio recordings. Then, "To Pitch (ac)" characterized the F0 frequency of first 40 ms of voicing, where F0 differences between onset obstruent consonants are typically most pronounced (Lea 1973; Hombert, Ohala, & Ewan, 1979; Hanson, 2009; Xu & Xu, 2021). After F0 values were log transformed, outliers +/- 3 standard deviations relative to a participant's mean F0 were removed from further analyses. Next, z-score normalization on a by-participant basis accounted for F0 variability across talkers that is impacted by multiple factors, including sex (Titze, 1989). Thus, a F0 value of 0 represented the mean F0 for a participant across all productions and values of +/- 1 corresponded to a standard deviation above and below the mean, respectively. Normalization provided a means of aligning F0 variability across participants prior to group analyses.

Analysis

Statistical analysis involved mixed effects models via the *lme4* package (Bates, Mochler, Bolker, and Walker, 2015) in R (version 4.1.3, R Core Development Team, 2022). In keeping with recommendations of Barr, Levy, Scheepers, and Tily (2013), we strove for including the maximal random effects in the models. Most models, however, did not tolerate the maximal random effect structure. For consistency, we report the models with random intercept of both subjects and items, which were tolerated by all models. The former captures variability among subjects; the latter among exposure sequences that changed from trial to trial. To assure that excluding random slopes did not radically alter any of the main conclusions, we also report the output of the models with the largest random effect structure tolerated by each model in Appendix B.

For perceptual categorization data, a logit mixed-effects logistic regression model included a binary response (*beer, pier*) as the dependent variable. The model included Condition (Canonical, Reverse), Test stimulus F0 (Low F0, High F0), and participant Sex (Male, Female) and their 2- and 3-way interactions as fixed effects, and by-subject and by-item random intercepts included. For speech production data, a continuous z-score normalized F0 dependent measure allowed for a standard (non-logit) linear mixed effects model. Here, too, fixed effects of Condition, Test stimulus, Sex and its interactions were modeled, with by-subject and by-items modeled as random

effects. Dependent categorical variables were center coded (1 vs -1). P-values were based on Satterthwaite approximates using the *LmerTest* package (version 3.1-3, Kuznetsova, Brockhoff, & Christensen, 2016). Analyses collapsed data from the three Canonical blocks and, separately, from the three Reverse blocks.

We conducted the production analyses in two steps: (1) Our first analysis used Test stimulus F0 to predict production F0. This analysis is parallel to the perceptual analysis and captures the whole process, which includes the change to perception as well as changes to production. (2) Our second analysis used perceptual responses as the main predictor of production F0. This analysis already partials out the contribution of perceptual changes as a function of exposure to the Canonical and Reverse distributions, which allows us to isolate the production component of transfer. The data, analysis code, and full tables of the results are available at https://osf.io/cwg4d/.

Results

Perceptual Categorization

Figure 2 plots categorization responses as a function of Canonical and Reverse short-term speech regularities. Table 1 presents the results of the analysis.

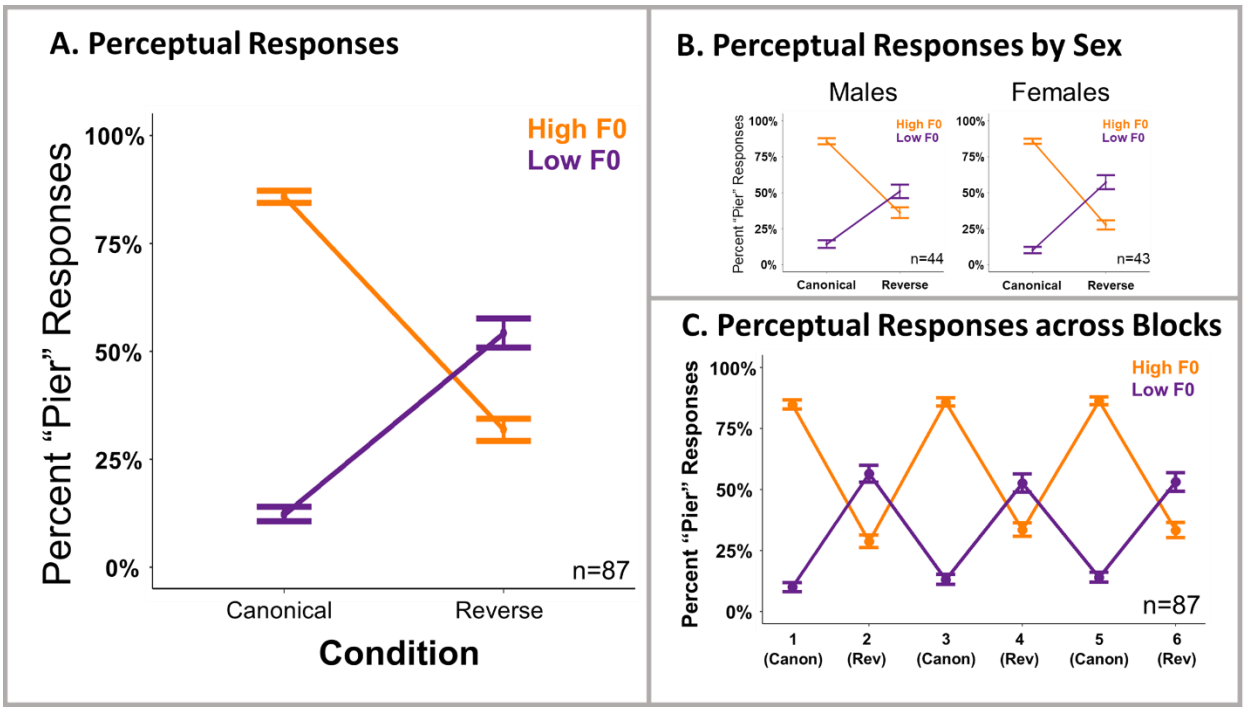

Figure 2. Results of Perceptual Categorization. Percentage of *pier* responses to High and Low F0 Test stimuli in Canonical and Reverse conditions are shown at the group level (A), broken down by Sex (B) and broken down by blocks (C). Averages reflect subject means ± SE).

As expected, there was a main effect of Test stimulus F0, such that the Test stimulus with the High F0 was more likely to be labeled as *pier* ($z = 9.94$, $p < .001$). Crucially, as in prior studies of dimension-based statistical learning, there was a significant interaction of Test stimulus F0 and Condition ($z = 16.09$, $p < .001$). Passive exposure to short-term speech input regularities impacted

the effectiveness of F0 in signaling *beer-pier* category identity. Neither the main effect of Sex, nor its interaction with Condition was significant[1]. There was, however, a significant three-way interaction between Sex, Condition and Test stimulus F0 ($z = 6.39$, $p < .001$). To better understand the nature of this interaction, we conducted separate tests on Male and Female participants. The results showed significant Condition by Test stimulus F0 interactions for both Male and Female participants, with a larger coefficient for Female participants ($\beta = 1.22$, $SE = .09$, $z = 14.19$, $p < .001$; $\beta = 1.49$, $SE = .08$, $z = 17.54$, $p < .001$, for Males and Females respectively).

Table 1 Regression Table – Perception

Predictor	*β*	*SE*	*z*	*p*
(Intercept)	-0.23	0.10	-2.26	.024
Condition	0.13	0.08	1.48	.139
Test cue F0	0.84	0.08	9.94	<.001
Sex	-0.05	0.06	-0.83	.406
Condition:Test cue F0	1.36	0.08	16.09	<.001
Condition:Sex	-0.04	0.02	-1.68	.094
Test cue F0:Sex	-0.01	0.02	-0.47	.635
Condition:Test cue F0:Sex	0.14	0.02	6.39	<.001

Note: Reference levels are condition (Reverse), Target stimulus F0 (Low F0), Sex (Male)

In summary, listeners relied on F0 to guide decisions about speech category identity when local speech input regularities conformed to English norms. When regularities shifted to create an 'accent', F0 was much less effective in signaling the speech categories. This replicates Hodson et al. (2023), who first demonstrated that passive exposure to speech elicits dimension-based statistical learning. Adding to that result, we also showed that the effect is robust in both male and female participants. Next, we examine the influence of this perceptual statistical learning on production.

Repetition (Speech Production)

Figure 3 plots z-score-normalized speech production F0s elicited in response to High and Low F0 Test stimuli in the context of Canonical and Reverse short-term speech regularities. As described under Analyses, two models were run on these data. The first model predicted changes to production F0 as a function of Test Stimulus F0. Table 2 presents this model's results. As in perceptual categorization, there was a significant effect of Test stimulus F0, such that the High F0 Test stimulus prompted a larger magnitude normalized F0 than the Low F0 test stimulus ($t = 15.36$, $p < .001$). There was also a significant effect of Condition, such that productions made in the Canonical Condition exhibited a higher F0 than the Reverse Condition ($t = 2.27$, $p = .026$). The interaction between Test Stimuli F0 and Condition was significant ($t = 19.02$, $p < .001$) in a manner consistent with transfer of perceptual statistical learning to production.

[1] The interaction between Sex and Condition was significant in a mixed effects model with a random slope of Test Cue F0 by Subject (see Table B1); however, post-hoc analyses run separately in males and females revealed no significant effect of Condition in either group ($z = 0.86$, $p = .391$, $z = 0.29$, $p = .769$, for Males and Females, respectively).

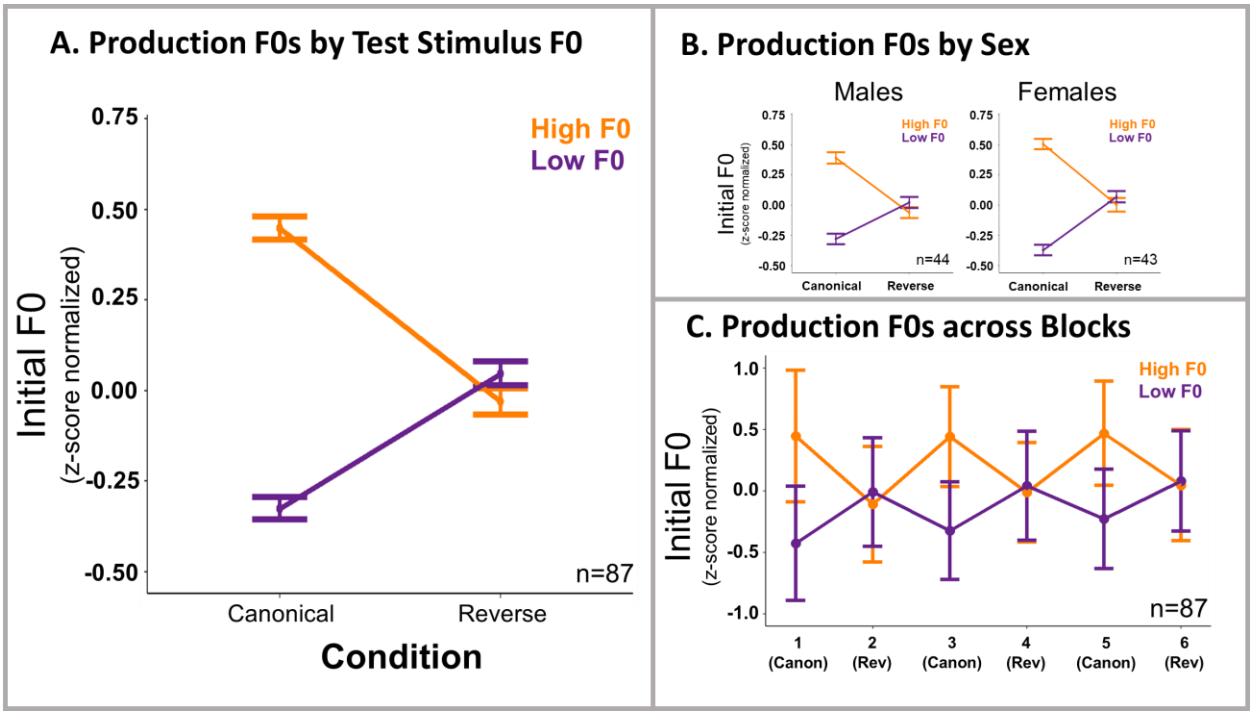

Figure 3. Results of Repetition. F0 values in speech production by Test stimulus F0 are shown at the group level (A), broken down by Sex (B) and broken down by blocks (C). Averages reflect subject means ± SE).

There was no significant main effect of Sex, though there was a significant 3-way interaction among Sex, Condition and Test stimulus F0 (t = 3.50, p = < .001). Separate post-hoc tests on Male and Female data revealed a significant interaction between Condition and Test stimulus F0 for both groups, with a larger coefficient for Female participants (β = 0.19, SE = 0.02, t = 12.87, p < .001; β = 0.25, SE = 0.01, t = 20.07, p < .001, for Males and Females, respectively).

Table 2 Regression Table – Production (by Test stimulus F0)

Predictor	*β*	*SE*	*t*	*p*
(Intercept)	0.01	0.01	0.50	.617
Condition	0.03	0.01	2.27	.026
Test stimulus F0	0.18	0.01	15.36	< .001
Sex	0.002	0.01	0.31	.760
Condition:Test stimulus F0	0.22	0.01	19.02	< .001
Condition:Sex	-0.01	0.01	-1.28	.199
Test stimulus F0:Sex	0.04	0.01	4.84	< .001
Condition:Test stimulus F0:Sex	0.03	0.01	3.50	< .001

Note: Reference levels are condition (Reverse), Target stimulus F0 (Low F0), Sex (Male)

These results suggest that production is affected by the manipulation of short-term regularities in speech perceived passively. However, it is possible that the results are driven by changes to perception and not production. Our second analysis addresses this issue by modeling changes to production F0 as a function of participants' perceptual choices, thus removing the variance due

to the influence of Test Cue F0 on perception. Figure 4 shows production F0 changes based on perceptual responses and Table 3 summarized the results of this analysis.

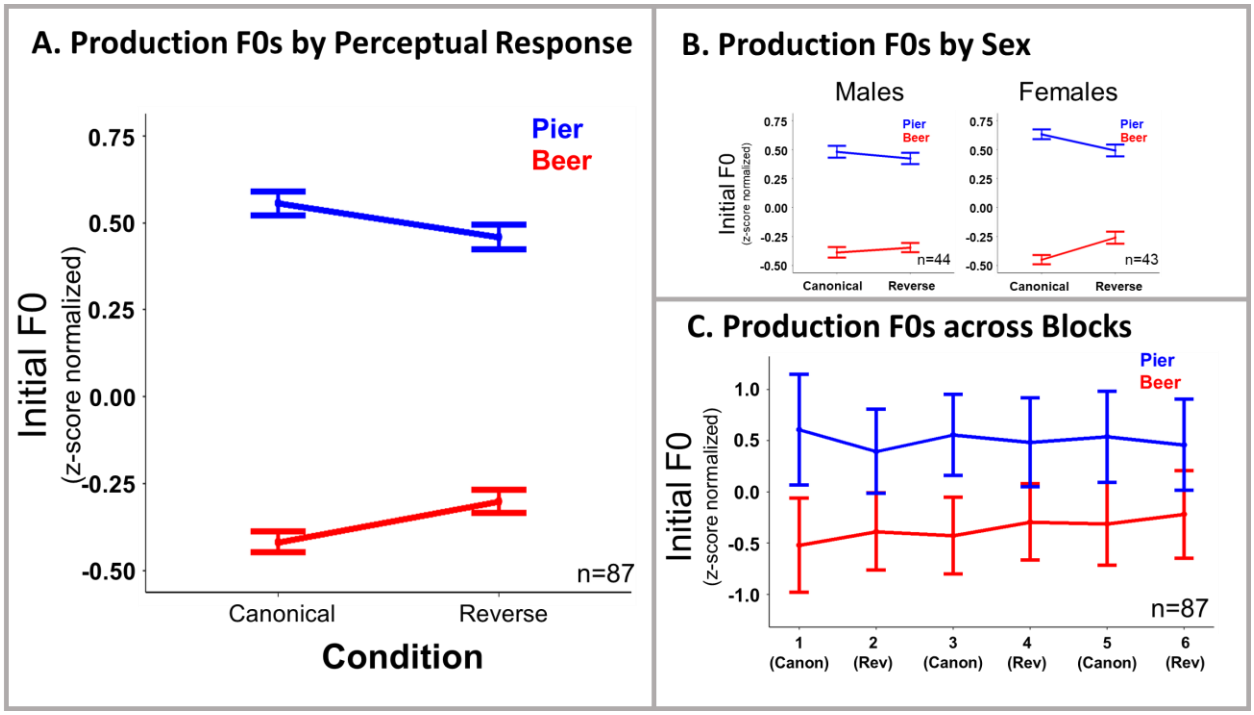

Figure 4. Results of Repetition. F0 values in speech production by perceptual responses are shown at the group level (A), broken down by Sex (B) and broken down by blocks (C). Averages reflect subject means ± SE).

As seen in Figure 4, when the contribution of perception is removed, the effect size clearly diminishes. The question is: Is there a significant production effect beyond those captured by perception? The results of the analysis suggest that there is. In addition to the main effect of Perceptual Response, there was a significant interaction between Perceptual Response and Condition ($t = -8.75$, $p < .001$), indicating within-word changes to F0 in productions as a function of Condition. A significant interaction between Perceptual Response and Sex ($\beta = -0.02$, $SE = .01$, $t = -3.36$, $p = .001$) was evident, indicating within-word changes to F0 as a function of Sex[2]. Moreover, there was a significant 3-way interaction among Sex, Condition, and Perceptual Response ($t= -4.62$, $p<.001$) with post-hoc tests revealing significant effects in both sexes, with a greater magnitude in Females ($\beta = -0.04$, $SE = .01$, $t = -3.61$, $p < .001$; $\beta = -0.09$, $SE = 0.01$, $t = -8.57$, $p < .001$, for Males and Females, respectively). This provides evidence of true transfer of perceptual statistical learning to production. This analysis shows that evidence of transfer of statistical learning to speech production is present even when the perceptual heterogeneity expected of F0-differentiated stimuli in the Reverse condition is factored out. For readers interested in changes to VOT, we have reported a series of analyses including that variable in Appendix A.

[2] The interaction between Perceptual Response and Sex was not significant in a mixed effects that included random slopes for both Condition and Perceptual Response by Subject (Table B3).

Table 3 Regression Table – Production (by Perceptual Response)

Predictor	*β*	*SE*	*t*	*p*
(Intercept)	0.04	0.01	4.32	<.001
Condition	0.01	0.01	0.77	.443
Perceptual Response	-0.44	0.01	-58.32	<.001
Sex	0.01	0.01	0.97	.334
Condition:Perceptual Response	-0.07	0.01	-8.75	<.001
Condition:Sex	-0.005	0.01	-0.65	.513
Perceptual Response:Sex	-0.02	0.01	-3.36	.001
Condition:Perceptual Response:Sex	-0.03	0.01	-4.62	<.001

Note: Reference levels are Condition (Reverse), Perceptual Response (Pier), Sex (Male)

Discussion

The findings of this study show that subtle acoustic regularities experienced in listening to a voice impact the details of our own speech. The influence of perceptual statistical learning on speech production is rapid, can result from passive listening, and impacts sublexical aspects of speech production in both male and female participants. The transfer we observe cannot be accounted for by mimicry of speech acoustics. Mimicry would predict consistent F0 patterns across conditions, since the speech tokens that elicited speech productions were constant across the experiment. Putting mimicry aside, the transfer of F0 down-weighting in an auditory repetition task can come from two sources: changes to the perception of the stimulus and/or changes to production. Comparison between the first and subsequent analyses allows us to segregate the contribution of each source. Hypothetically the down-weighting of F0 differences in productions in the Reverse condition might have arisen solely from perception, without transfer to speech production. If participants were to utter *beer* and *pier* with English-consistent F0 each time they heard a High-F0 or Low-F0 target then the overall F0 difference in the Reverse condition might be diminished relative to the Canonical condition simply because *perceptual* down-weighting leads to greater inhomogeneity in the proportion of *beer* versus *pier* percepts in the Reverse, compared to the Canonical, condition. This inhomogeneity would mean that High- and Low-F0 targets elicit a mix of high and low F0 productions entirely due to perception, without any transfer of learning to production.

Our first analysis, conditioned on Test stimulus F0, shows the combined perception plus production effect of transfer to be of a large effect size. A second analysis conditioning production F0 according to *beer* versus *pier* categorization instead of test stimulus acoustics removes the contribution of perception and shows smaller, albeit persistent, F0 down-weighting in Reverse condition productions. Together, the analyses suggest that although there is a sizable perceptual contribution, there is also a unique contribution of transfer of the effects of statistical learning to production. The fact that this influence differs in magnitude across analyses conditioned on perceptual categorization versus input acoustics also makes an important point: the long-term norms of speech production are not overwritten by the more subtle influences of rapid statistical learning evident across short-term input. This is consistent with the demonstration that auditory repetition of familiar words is largely lexical (Dell et al., 2013; Nozari & Dell, 2013; Nozari et al., 2010).

These findings align with positive reports of phonetic convergence on F0 in shadowing tasks (Garnier et al., 2013; Mantell & Pfordresher, 2013, Postma-Nilsenová & Postma, 2013; Sato et al., 2013; Wisniewski et al., 2013). At the same time, they also illustrate how our statistical learning approach can provide a solution to the challenges of capturing and characterizing phonetic convergence. One advantage is dimension selection. *A priori* predictions about the dimensions expected to exhibit phonetic convergence have proven challenging in the phonetic convergence literature, as beautifully demonstrated by an exhaustive search across more than 300 acoustic-phonetic features (Ostrand & Chodroff, 2021). Our statistical learning approach provides *a priori* predictions of the dimension impacted by convergence (Wu & Holt, 2022), eliminating the need to selectively – or exhaustively – sample dimensions across which to examine the nature of transfer.

A second advantage is the ability to make directional predictions. Dimension-based statistical learning elicits predictable, directional effects on perception. When short-term speech input provides robust information (here, VOT) to indicate category identity, secondary dimensions that depart from long-term norms of these categories (as, here, for F0 in Reverse condition) are down-weighted in their influence on perceptual categorization (Wu & Holt, 2022). This has proven to be the case across consonants (Idemaru & Holt, 2011), vowels (Liu & Holt, 2015), and also prosodic emphasis (Jasmin et al., 2022) categories. This is important in that it emphasizes that the transfer to production is not simply convergence in the sense of imitation. Rather, directional sublexical adjustments in the perceptual system are carried over to the production system. As a result, we would not expect all changes to the acoustics of speech to transfer to production (see, e.g., our VOT analysis). This, in turn, may help to explain why phonetic convergence studies often yield inconsistent reports.

A third advantage is the ability to set aside sociolinguistic factors. Our manipulation of acoustic F0 was barely perceptible, and devoid of socially discriminating information because the voice was constant across conditions. With this approach, we observed transfer in both male and female participants. The consistency of our findings across sex may have been supported by our approach, which allowed us to eliminate sociolinguistic factors that may contribute to the variability of findings reported in the phonetic convergence literature (Pardo et al., 2017). A sizeable literature now exists detailing social and contextual factors eliciting convergence, such as talker attractiveness (Babel, 2012), conversational topic (Walker, 2014), and even cultural primes (Hurring et al., 2022; Walker et al., 2019). Further understanding of how these factors influence convergence will benefit from an understanding of the cognitive mechanisms of transfer (Pardo, 2022). Here, we put forward one such an account, in the framework of statistical learning wherein several computational approaches to the perceptual effects have been proffered (Harmon et al., 2019; Kleinschmidt & Jaeger, 2015; Liu & Holt, 2015; Wu, 2020).

At the broadest level, the results demonstrate that subtle statistical regularities experienced in passive listening to *another talker's speech* can transfer to influence one's own speech production. Statistical learning involving short-term regularities in perceived speech impacts sublexical aspects of speech production in a predictable manner, even when the speech targets that elicit production are held constant to prevent mimicry. In sum, by yielding specific *a priori* predictions of the sublexical aspects of speech expected to be impacted by transfer of statistical learning, dimension-based statistical learning across passive exposure to speech provides a valuable new framework for understanding perception-production transfer.

Open Practices Statement: Data, R scripts used for statistical analyses, and results are available at https://osf.io/cwg4d/.

Appendix A

VOT analysis

We measured production VOT on each trial using the Deep and Robust VOT annotator (Dr.VOT; Shrem, Goldrick, and Keshet, 2019), with post-processing manual inspection. Next, we z-scored the VOTs following the by-participant approach described for F0.

We used VOT to verify that perceptual categorization (*beer, pier*) responses were followed by speech productions that corresponded to the perceptual response (e.g., longer VOTs following *pier* vs. *beer* responses). **Figure A1** shows the raw (1A.A) and z-scored (1A.B) distribution of VOTs as a function of *beer-pier* perceptual responses. Perceptual responses were strongly associated with the VOT of subsequent productions ($t = -141.94$, $p < .001$) and production VOT distributions were not influenced by (Canonical, Reverse) condition ($t = -1.51$, $p = .13$, Table A1). Point biserial correlation between VOT and perceptual categorization reveals a strong relationship that is comparable across Canonical and Reverse conditions ($r = 0.74$, $p < .001$ for the Canonical condition and $r=0.74$, $p < .001$, for the Reverse condition). In sum, participants' *beer-pier* perceptual responses to Test Stimuli were followed by speech productions possessing VOTs that align with these perceptual categories.

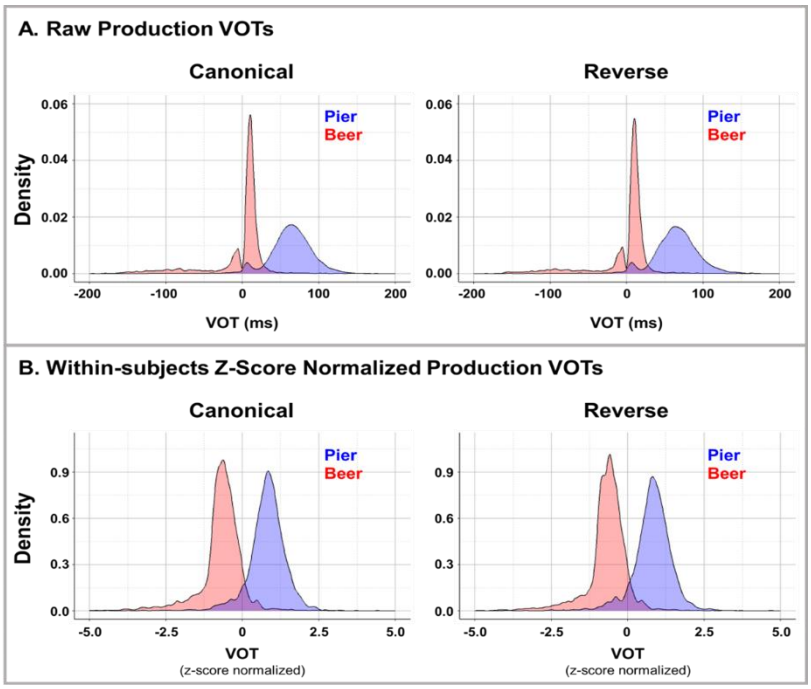

Figure A1. Distribution of raw Production VOTs (A) and z-scored VOTs (B) for beer and pier in Canonical and Reverse Conditions.

Table A1 Regression Table – Production VOTs (by Perceptual Response)

Predictor	*β*	*SE*	*t*	*p*
(Intercept)	0.05	0.01	4.16	<.001
Condition	-0.01	0.01	-1.51	.130
Perceptual Response	-0.75	0.01	-141.94	<.001
Condition:Perceptual Response	-0.002	0.01	-0.33	.743

Note: Reference levels are Condition (Reverse), Perceptual Response (Pier)

Given that VOT is well-aligned with the perceptual response we next examined its utility as a continuous measure of participants' intended utterance in testing transfer of statistical learning to the weighting of F0 in speech production. We fit a model predicting normalized speech production F0 as a function of utterance VOT (a continuous-measure proxy for intended production) and condition (Canonical, Reverse). Table A2 shows the predicted interaction (β=-0.02, SE=.01, t=-2.02, p= .043). The significant interaction replicates the transfer of statistical learning to speech production that persists when the F0 of the stimulus eliciting the utterance is factored out, and results are examined as a function of a participant's *intended* speech production (here assessed with VOT, assessed via perceptual response in Figure 4).

Table A2 Regression Table – Production F0s (by Production VOTs)

Predictor	*β*	*SE*	*t*	*p*
(Intercept)	0.01	0.02	0.56	.580
Condition	0.01	0.02	0.64	.527
VOT	0.32	0.01	40.33	<.001
Condition:VOT	-0.02	0.01	-2.02	.043

Note: Reference levels are Condition (Reverse). All VOT measures z-score normalized within participants

In sum, participants are largely consistent in producing words that correspond to their preceding perceptual choices. Transfer of statistical learning to speech production is observed in analyses utilizing VOT as a continuous measure of participants' intended productions.

Appendix B

Mixed effects models with the largest random effect structure tolerated by each model

Mixed effect models presented below share the same fixed effect structure as corresponding models in the main manuscript but also include the largest random effect structure tolerated by each model.

Table B1 Perceptual Categorization

*Perceptual Response ~ Condition * Test cue F0 * Sex + (1 + Test cue F0 | Subject) + (1 | Item)*

Predictor	*β*	*SE*	*z*	*p*
(Intercept)	-0.38	0.13	-2.97	.003
Condition	-0.004	0.10	-0.04	.966
Test cue F0	1.12	0.17	6.45	<.001
Sex	-0.08	0.08	-0.99	.322
Condition:Test cue F0	1.70	0.10	16.40	<.001
Condition:Sex	-0.07	0.03	-2.74	.006
Test cue F0:Sex	-0.04	0.14	-0.29	.773
Condition:Test cue F0:Sex	0.16	0.03	6.13	<.001

Note: Reference levels are Condition (Reverse), Test cue F0 (LowF0), Sex (Male)

Table B2 Repetition (Speech Production by Target Cue F0)

Random effect structure in presented in Table 2 is already the largest random effect structure tolerated by this model.

Table B3 Repetition (Speech Production by Perceptual Response)

*Z-scored Production F0 ~ Condition * Perceptual Response * Sex + (1 + Perceptual Response + Condition | Subject) + (1 | Item)*

Predictor	*β*	*SE*	*t*	*p*
(Intercept)	0.04	0.01	3.51	.001
Condition	0.004	0.01	0.35	.723
Perceptual Response	-0.42	0.03	-16.21	<.001
Sex	0.01	0.01	0.66	.512
Condition:Perceptual Response	-0.06	0.01	-7.33	<.001
Condition:Sex	-0.001	0.01	-0.13	.898
Perceptual Response:Sex	-0.02	0.03	-0.95	.346
Condition:Perceptual Response:Sex	-0.03	0.01	-4.64	<.001

Note: Reference levels are Condition (Reverse), Perceptual Response (LowF0), Sex (Male)

References

Abramson, A. S., & Lisker, L. (1985). Relative power of cues: F0 shift versus voice timing. *Phonetic linguistics: Essays in honor of Peter Ladefoged*, 25-33.

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

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.

Babel, Molly. 2010. Dialect convergence and divergence in New Zealand English. *Language in Society 39*. 437–456.

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823

Boersma, P. & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 5.3, retrieved from http://www.praat. org/

Bohland, J. W., Bullock, D., & Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. *Journal of cognitive neuroscience*, *22*(7), 1504-1529.

Bourhis, R. Y., & Giles, H. (1977). The language of intergroup distinctiveness. *Language, ethnicity and intergroup relations*, *13*, 119.

Dell, G. S., Schwartz, M. F., Nozari, N., Faseyitan, O., & Coslett, H. B. (2013). Voxel-based lesion-parameter mapping: Identifying the neural correlates of a computational model of word production. *Cognition*, *128*(3), 380-396.

Earnshaw, K. (2021). Examining the implications of speech accommodation for forensic speaker comparison casework: A case study of the West Yorkshire FACE vowel. Journal of Phonetics, 87, 101062.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of memory and language*, *49*(3), 396-413.

Garnier, M., Lamalle, L., & Sato, M. (2013). Neural correlates of phonetic convergence and speech imitation. *Frontiers in psychology*, *4*, 600.

Giles, H., Coupland, N., & Coupland, J. (1991). 1. Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, *1*.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. Psychological Review, 105(2), 251–279.

Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological cybernetics*, *72*(1), 43-53.

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.

Harmon, Z., Idemaru, K., & Kapatsinski, V. (2019). Learning mechanisms in cue reweighting. *Cognition*, *189*, 76-88.

Heath, J. (2015, April). Convergence through divergence: compensatory changes in phonetic accommodation. In *LSA Annual Meeting Extended Abstracts* (Vol. 6, pp. 7-1).

Hodson, A. J., Shinn-Cunningham, B., & Holt, L. L. (2023). Statistical learning across passive listening adjusts perceptual weights of speech input dimensions. PsyArXiv. https://doi.org/10.31234/osf.io/4kxz3

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

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

Hurring, G., Hay, J., Drager, K., Podlubny, R., Manhire, L., & Ellis, A. (2022). Social priming in speech perception: Revisiting kangaroo/kiwi priming in New Zealand English. *Brain Sciences*, *12*(6), 684.

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

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1009.

Idemaru, K., & Holt, L. L. (2020). Generalization of dimension-based statistical learning. *Attention, Perception, & Psychophysics*, *82*(4), 1744-1762.

Idemaru, K., & Vaughn, C. (2020). Perceptual tracking of distinct distributional regularities within a single voice. *The Journal of the Acoustical Society of America*, *148*(6), EL427-EL432.

Kong, E., & Edwards, J. (2011). Individual Differences in Speech Perception: Evidence from Visual Analogue Scaling and Eye-Tracking. In *Proc. Int. Conf. Phonetic Sci* (pp. 1126-1129).

Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. Journal of Phonetics, 59, 40-57.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, *82*, 1-26.

Kwon, H. (2019). The role of native phonology in spontaneous imitation: Evidence from Seoul Korean.

Lea, W. A. (1973). Segmental and suprasegmental influences on fundamental frequency contours. *Consonant types and tone*, *1*, 15-70.

Lehet, M., & Holt, L. L. (2017). Dimension-based statistical learning affects both speech perception and production. *Cognitive science*, *41*, 885-912.

Lehet, M. & Holt, L. L. (2020). Nevertheless, it persists: Perceptual recalibration and normalization of speech impact different levels of representation. Cognition, 202,104328.

Lindsay, S., Clayards, M., Gennari, S., & Gaskell, M. G. (2022). Plasticity of categories in speech perception and production. *Language, Cognition and Neuroscience*, *37*(6), 707-731.

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1783.

Mantell, J. T., & Pfordresher, P. Q. (2013). Vocal imitation of song and speech. *Cognition*, *127*(2), 177-202.

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

Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2010). Alignment to visual speech information. *Attention, Perception, & Psychophysics*, *72*(6), 1614-1625.

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.

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, *21*(4), 422-432.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132-142.

Nozari, N., Kittredge, A. K., Dell, G. S., & Schwartz, M. F. (2010). Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of memory and language*, *63*(4), 541-559.

Nozari, N., & Dell, G. S. (2013). How damaged brains repeat words: A computational approach. *Brain and language*, *126*(3), 327-337.

Ostrand, R., & Chodroff, E. (2021). It's alignment all the way down, but not all the way up: Speakers align on some features but not others within a dialogue. *Journal of phonetics*, *88*, 101074.

Postma-Nilsenová, M., & Postma, E. (2013). Auditory perception bias in speech imitation. *Frontiers in Psychology*, *4*, 826.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. The Journal of the Acoustical Society of America, 119(4), 2382-2393.

Pardo, J. S., Jay, I. C., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics*, *72*(8), 2254-2264.

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*(3), 183-195.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, *79*(2), 637-659.

Pardo, J. S., Pellegrino, E., Dellwo, V., & Möbius, B. (2022). Vocal accommodation in speech communication. *Journal of Phonetics*, *95*, 101196.

Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J. L., & Nguyen, N. (2013). Converging toward a common speech code: imitative and perceptuo-motor recalibration processes in speech production. *Frontiers in psychology*, *4*, 422.

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of phonetics*, *52*, 183-204.

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2016). Individual differences in perceptual adaptability of foreign sound categories. *Attention, Perception, & Psychophysics*, *78*(1), 355-367.

Schertz, J., & Clare, E. J. (2020). Phonetic cue weighting in perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(2), e1521.

Schertz, J., & Paquette-Smith, M. (2023). Convergence to shortened and lengthened voice onset time in an imitation task. *JASA Express Letters*, *3*(2), 025201.

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

Shrem, Y., Goldrick, M., & Keshet, J. (2019). Dr. VOT: Measuring positive and negative voice onset time in the wild. *arXiv preprint arXiv:1910.13255*.

Tilsen, S. (2009). Subphonemic and cross-phonemic priming in vowel shadowing: Evidence for the involvement of exemplars in production. *Journal of Phonetics*, *37*(3), 276-296.

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.

Walker, Abby. 2014. Crossing oceans with voices and ears: Second dialect acquisition and topic-based shifting in production and perception. Columbus, OH: Ohio State University dissertation.

Walker, M., Szakay, A., & Cox, F. (2019). Can kiwis and koalas as cultural primes induce perceptual bias in Australian English speaking listeners?. *Laboratory Phonology*, *10*(1).

Wisniewski, M. G., Mantell, J. T., & Pfordresher, P. Q. (2013). Transfer effects in the vocal imitation of speech and song. *Psychomusicology: Music, Mind, and Brain*, *23*(2), 82.

Wu, Y. C. (2020). Behavioral, computational, and electrophysiological investigations of adaptive plasticity mechanisms in speech perception. [Doctoral Dissertation, Carnegie Mellon University]

Wu, Y. C. & Holt, L. L. (2022). Phonetic category activation drives adaptive plasticity in dimension-based statistical learning in speech perception. Journal of Experimental Psychology: Human Perception & Performance, 48, 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., & Holt, L. L. (2018). Simultaneous tracking of coevolving distributional regularities in speech. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(11), 1760.

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), e12947.