

RESEARCH ARTICLE

Control of Movement

Vocal pitch enables differential motor learning of speech segments

 Robin Karlin,^{1,4}  Emily Tesch,^{2,4}  Ding-Lan Tang,^{3,4}  Yuyu Zeng,⁴  Caroline A. Niziolek,^{4,5*} and  Benjamin Parrell^{4,5*}

¹Department of Speech, Language and Hearing Sciences, University of Missouri, Columbia, Missouri, United States;

²Department of Speech and Hearing Science, Arizona State University, Tempe, Arizona, United States; ³Academic Unit of Human Communication, Learning, and Development (HCLD), The University of Hong Kong, Hong Kong SAR, People's Republic of China; ⁴Waisman Center, University of Wisconsin-Madison, Madison, Wisconsin, United States; and

⁵Department of Communication Sciences and Disorders, University of Wisconsin-Madison, Madison, Wisconsin, United States

Abstract

Sensory feedback is crucial for accurate motor control. One process of movement correction is sensorimotor adaptation, or motor learning in response to perceived sensory errors. Recent studies demonstrate that people can adapt to opposing errors on a single movement given a context that differentiates when each error occurs. In speech production, linguistic structure (e.g., the same vowel in different words) can provide context for differential adaptation, but it is unclear whether this is restricted to the same effectors (i.e., lips, tongue, jaw) or also includes movements of other speech effectors (i.e., the larynx). Reaching studies show that contextual movements need not be produced with the same effector as the learning target, but so far, they have only tested left-right pairs. We present three simultaneous adaptation experiments in speech that examine whether laryngeal movements for pitch can provide context for oral articulator movements for vowels. In each experiment, the resonances that correlate with vowel articulator position were perturbed in three directions that were predictable given a pitch context. First, Mandarin speakers differentially adapted given pitch contexts that signal differences in word meaning, suggesting that lexical pitch provides context for vowels. Second, English speakers differentially adapted given arbitrary pitch matching contexts on the word “head,” suggesting that nonmeaningful pitch movements provide context for vowels. Third, English speakers did not differentially adapt when listening to contextual pitch, indicating that mere auditory input of pitch is insufficient. Together, these results suggest that sensorimotor context for learning can be provided by effectors other than the learning target.

NEW & NOTEWORTHY Previous work shows that sensorimotor learning can be specific to different motor contexts, but to date, this research has only examined contexts provided by the same effector as the learning target or its contralateral pair. We show that laryngeal movements for pitch enable differentiated learning of oral articulator movements for vowels, even when the pitch is linguistically meaningless. This indicates that motor contexts that enable learning can be generated by effectors distinct from those that undergo learning.

altered auditory feedback; contextual learning; sensorimotor adaptation; speech production

INTRODUCTION

Extensive research has shown that sensory feedback plays a crucial role in maintaining accurate motor control, including speech motor control (1–4). Motor behavior produces sensory feedback (e.g., visual, somatosensory, auditory feedback) that the central nervous system uses for both online control

and to alter motor plans for future movement. External perturbations of this feedback have been used extensively to examine these processes, with a particular focus on mechanisms of feedback-driven updates to future movement plans, generally referred to as sensorimotor adaptation. For example, visual feedback of reaching movements can be altered such that participants see their hand further to the right or



*C. A. Niziolek and B. Parrell contributed equally to this work.

Correspondence: R. Karlin (rkarlin@health.missouri.edu).

Submitted 17 December 2024 / Revised 7 January 2025 / Accepted 13 March 2025



left than reality (5–7); typical studies in speech apply perturbations to the auditory feedback that speakers hear of their own voice, e.g., lowering the first resonant frequency (F1) of the vowel in *head* /hed/ to result in a token sounding more like *hid* /hɪd/ (8, 9). Historically, a single perturbation has been applied consistently to either a single movement, or to all movements in a study. In such studies, motor learning observed on the target movements could reflect either a generalized change to the entire repertoire for that effector or a change limited to the trained movements themselves. As such, the scope of sensorimotor adaptation is unclear. That is, it is not known the extent to which people are learning about movement in general versus more specific learning of the movements targeted by the experiment—and if it is specific, what defines a specific movement.

A growing literature has begun to examine what conditions allow for separate, context-dependent adaptation of one movement (10–13). In these simultaneous adaptation experiments, a single target movement is perturbed in two opposing directions during a single experimental phase. The direction of perturbation is consistently associated with a secondary contextual cue. For example, a forward reach may be perturbed to the right when associated with a particular color (e.g., green) or when followed by a rightward reach, and perturbed to the left when associated with a distinct color (e.g., white) or when followed by a leftward reach. These studies have shown that only cues that involve the motor system in some way (such as being paired with different reaches) enable differential learning, whereas purely sensory cues (such as different colored lights) do not (12–15). However, studies that test pairs of movements have been limited in that they have only tested context-target pairs that use either the same effector (e.g., two reaches with one hand) or contralateral effector pairs (e.g., a reach with the left hand providing context for a reach with the right hand). Thus, it is unclear whether movements from a different type of effector can provide sensorimotor context for a target movement.

There have recently been a handful of studies on context-dependent adaptation in speech, which established the use of opposing auditory perturbations as a tool to investigate the scope of learning in speech. The organization of language poses a particularly interesting arena for investigation, as there are many theoretical units that could be the target of sensorimotor adaptation, such as a single word, a phoneme, or even an entire class of speech sounds. However, similarly to manual reaching studies, examination of specific learning thus far has been limited to word contexts that recruit contextual movements of the same articulators, such as simultaneous perturbation of the same vowel in different words with different preceding sounds (e.g., ‘head’ vs. ‘Ted,’ 16), or simultaneous perturbation of the same syllable in different words (e.g., ‘pedigree’ vs. ‘pedicure,’ 17). Similar to the work in the reaching literature, both of these studies showed that speakers were able to implement different adaptations to different words at the same time, indicating that sensorimotor adaptation can apply to a more specific target than one movement (i.e., a word rather than a phoneme or syllable). However, in both of these studies, either lexical information or the movements of supralaryngeal articulators could provide the context.

Here, we investigate the extent to which vocal pitch, i.e., the fundamental frequency (f_0) of a spoken word, can serve

as the context for differential sensorimotor adaptation of speech segments (here, vowels). f_0 provides crucial insight on the question of specific sensorimotor adaptation because it creates different motor contexts in the larynx, without directly involving the supralaryngeal articulators targeted by the auditory perturbations. In *experiment 1*, we test the hypothesis that f_0 as lexical tone in Mandarin can serve as a context for differential adaptation of segments by perturbing the vowel *ei* in minimal pairs that consist of identical segments and differ only in tone. We show that Mandarin speakers can adapt F1 production to simultaneous opposing perturbations that are cued by lexical tone, indicating that differential adaptation is not reliant on differing kinematics of the target articulator. We also conduct two follow-up studies to pinpoint the source of motor context from lexical tone production. First, we examine whether these results are unique to lexical tone, or whether nonlexical f_0 can similarly facilitate simultaneous adaptation to opposing formant perturbations of a single target syllable by having English-speaking participants imitate distinct auditory tones (*experiment 2*). Second, we test a control condition that includes the same auditory pitch cue as in *experiment 2* but without imitation in production (*experiment 3*), which we do not expect to enable adaptation as it is a nonmotor cue. These follow-up studies show that pitch imitation, but not auditory pitch cues alone, facilitates adaptation to simultaneous opposing formant perturbations, indicating that speech motor learning can be specific to a vocal pitch context, even when the pitch confers no linguistic meaning.

METHODS

Participants

Sixty-one people with no reported history of hearing, speech, or neurological disorders participated in the study (*experiment 1*: $n = 20$, 16/3/1 women/men/nonbinary, age 18–31 yr, median = 23, SD = 3.8; *experiment 2*: $n = 20$, 16/4 women/men, age 18–30 yr, median = 20, SD = 3.9; *experiment 3*: $n = 21$, 14/7 women/men, age 18–43 yr, median = 23, SD = 6.3). Data from two participants was excluded due to an error in the procedure that led to pitch cues that were not calibrated to their habitual pitch (for more information, see *Participant-Specific Pitch Cues*). Participants in *experiment 1* were native speakers of Mandarin Chinese; participants in *experiment 2* and *experiment 3* were native speakers of American English that did not speak any tonal languages. All participants passed an automated Hughson-Westlake hearing screening (pure-tone thresholds ≤ 25 dB HL in both ears at 250, 500, 1,000, 2,000, and 4,000 Hz). No participant took part in multiple experiments. Participants were compensated for their participation either monetarily or through extra credit in a course in the University of Wisconsin-Madison Communication Sciences and Disorders Department. All participants provided written, informed consent. All procedures were approved by the Institutional Review Board at the University of Wisconsin-Madison.

A sample size of 20 participants per experiment provides 80% power to detect an effect size of $d = 0.58$, which is a much smaller effect size than those reported in previous simultaneous adaptation studies (16).

Apparatus

The study was conducted in a quiet room with participants seated in front of a computer screen. Participants were exposed to multiple real-time perturbations of F1 (first resonant frequency of the vowel). Participants spoke words as they appeared on the screen into a desk-mounted microphone (Sennheiser MKE 600) and received auditory feedback through over-ear headphones (Beyerdynamic DT 770 PRO) at ~80 dB SPL mixed with masking noise at ~60 dB SPL to limit potential bone- or air-conducted perception of unperturbed speech. Speech was recorded, processed, perturbed (on some trials), and played back to participants using a modified version of Audapter (18). Formant frequency alterations were applied in mels, a logarithmic transformation of frequency where equal differences in mels are judged by listeners to correspond to equal changes in pitch. The measured latency of this system was ~19 ms.

Procedures

In *experiment 1*, there were three target words differing only in lexical tone: 飞 fēi “to fly” (tone 1; high tone); 肥 fēi “fat” (tone 2; rising tone), and 费 fèi “cost” (tone 4; falling tone) (Fig. 1A, left). Words were presented to the participant on a computer monitor using simplified Chinese characters. In each trial, the target word was on the screen for 1.5 s, and there were 1.25 s between the end of one trial and the beginning of the next trial, with a random jitter of up to 250 ms in either direction (i.e., 1–1.5 s between stimuli). The target words were pseudorandomly ordered within each phase (see last paragraph in this section) such that no two sequential trials had the same target. The vowel *ei* in each word received one of three perturbations to F1, with a maximum perturbation of 125 mels: F1 up, F1 down, or no perturbation. For each participant, the F1 perturbation was consistent within a word; the perturbation received by each word was

counterbalanced across participants to the extent possible (each of the possible 6 permutations assigned to 3 or 4 participants).

In *experiment 2*, there was only one target word, “head” (Fig. 1A, middle). In each trial during the main experiment, participants heard a 300-ms pure tone (high, mid, or low; for more information on how the frequencies of the pure tones were determined, see *Participant-Specific Pitch Cues*), followed by a 50-ms gap, and then the orthographic stimulus “head” appeared on the screen; the timing of the rest of the trial was identical to *experiment 1*. Participants were instructed to match the pitch of their production of the word “head” to the pure tone cue. Participants were instructed to speak the word as normally as possible, rather than singing. Pure tones were pseudorandomly presented such that no two adjacent trials had the same target pitch, as in *experiment 1*. The vowel /e/ in each pitch-match of “head” received one of three perturbations, with a maximum perturbation of 125 mels: F1 up, F1 down, or no perturbation. The perturbation received by each pitch match was counterbalanced across participants to the extent possible (each of the possible 6 permutations was assigned to 3 or 4 participants). After the experiment was complete, participants completed an abbreviated version of the Edinburgh Lifetime Music Experience Questionnaire (ELMEQ, 19) to collect information on instrumental and voice training.

In *experiment 3*, a participant-specific high, mid, or low auditory pitch cue preceded the orthographic stimulus “head,” but participants were instructed to read the word aloud normally and to not match the pitch they heard (Fig. 1A, right). If participants started to anticipate the presentation of the word “head” and started speaking before the tone was finished, they were reminded to wait until after the tone was done. The experimenter also monitored the participants’ productions for influence from the preceding pitch cue, using a visual tracker that displayed values from MATLAB’s built-in pitch tracking function; no participants

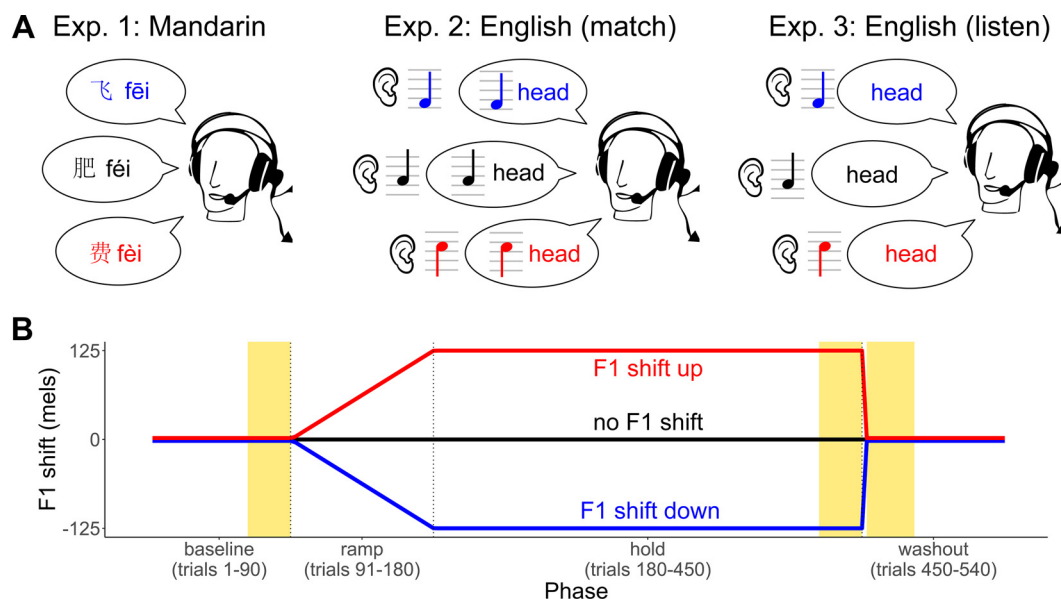


Figure 1. Schematic of all three experiments. *A*: an illustration of the tasks for each experiment. *B*: an illustration of the phases and perturbations used in this study; all experiments used this scheme. Yellow shaded areas indicate windows of analysis.

demonstrated a tendency to inadvertently match the pitch. The perturbation received by each word was counterbalanced across participants to the extent possible (each of the possible 6 permutations was assigned to 3 or 4 participants).

All experiments had four phases (Fig. 1B): a baseline phase with veridical feedback (30 trials each of 3 words; 90 total trials); a ramp phase where the perturbations were gradually introduced up to a maximum of 125 mels (30 trials each of 3 words; 90 total trials); a hold phase with constant perturbation of 125 mels (90 trials each of 3 words; 270 total trials); and a washout phase with veridical feedback (30 trials each of 3 words; 90 total trials).

Participant-Specific Pitch Cues

In *experiment 2* and *experiment 3*, the frequencies of the pure tone targets were determined at the beginning of the experiment, based on the participant's habitual pitch. In a pre-test phase, participants read the phrases "My lion is yellow," "Our llama ran away!" and "Does Mary owe you money?" three times each, in random order. These phrases are highly sonorous and use a wide intonational range, thus providing an approximation of their habitual pitch. f_0 tracks were automatically extracted using wave_viewer, a MATLAB-based GUI (20). To avoid undue influence from mistracked samples, f_0 values were excluded 1) first if they were beyond minimum and maximum acceptable boundaries (lower than 50 Hz or higher than 500 Hz), and 2) then if they were more than three standard deviations beyond the median pitch after the removal of values outside the acceptable boundaries.

The median f_0 value of the cleaned f_0 tracks was then used as the baseline for the low tone. The experimenter confirmed that the baseline f_0 was a likely candidate (not based on mistracked pitch) based on their impression of the participant's voice and gender-related differences, using a general guideline of ~100 Hz for male participants and ~200 Hz for female participants. This value was then matched to the closest canonical Western musical note to produce the low tone. Canonical musical notes were used in case there were participants with perfect pitch that might be bothered by f_0 values that were slightly off from canonical notes. The mid tone was three semitones higher than the low tone, and the high tone was three semitones higher than the mid tone, for an overall difference of six semitones between high and low, where semitones are the smallest pitch interval that divides the Western musical scale. This value was based on the range of the Mandarin falling tone (*tone 4*) in pilot data from *experiment 1*, such that speakers in both experiments used a similar pitch range. Pilot testing showed that this range was generally comfortable; three semitones is also well above pitch discrimination thresholds for both typical listeners and "tone-deaf" listeners (21).

Pure tones (sine waves) were then generated for each pitch, with a duration of 300 ms. This duration was based on typical durations of the vowel in "head" in previous experiments (22) and was chosen to promote a spoken production of the word, rather than a sung production. To ensure equal loudness percepts between the tones, the amplitude of each tone was set based on the 80-phon curve for each frequency (23, 24).

After the tone values were determined, participants completed a practice phase with nine trials (3 trials/pitch cue)

that were identical in procedure to the remainder of the experiment, with no perturbation to the vowel formants. For *experiment 2*, the experimenter assessed whether the participant was speaking the words (as opposed to 'singing' them), and if they were reliably producing pitch differences. If either of these criteria was not met, the practice phase was repeated. For *experiment 3*, the experimenter visually inspected automatically tracked f_0 to ensure that participants were not matching the pitch inadvertently. If this criterion was not met, the practice phase was repeated. Participants were able to repeat practice on request.

Data Processing

Formant tracking was performed with wave_viewer (20) using the Praat formant tracking algorithm (25). Vowel onset and offset were set automatically using a participant-specific amplitude threshold. Errors in vowel onset and offset were corrected by hand-marking the location of the vowel using the spectrogram and waveform of the speech sample. Vowel onset was identified by the presence of F1 and F2 on the spectrogram and periodicity on the waveform. Vowel offset was identified when F1 and F2 were no longer visible on the spectrogram. Within this marked time range, formant values were tracked with a specific linear predictive coding (LPC) order (the number of coefficients used to estimate the formants) and pre-emphasis value (a filter to boost higher frequencies and improve signal-to-noise ratio for formants) for each participant. These parameters were adjusted on a per-trial basis if there were formant tracking errors. Trials with unresolvable formant tracking errors or with production errors (such as saying the wrong word, yawning during production, etc.) were excluded (1.4%, 0%–3.5% across participants). To focus on changes in F1 due to sensorimotor adaptation and avoid effects of formant transitions and online compensation resulting from the sensory feedback control system, which is typically measurable ~100–150 ms after the onset of F1 perturbation (3, 26, 27), a single mean F1 value for each trial was calculated from a window 25–100 ms after vowel onset.

Statistical Analysis

Statistical analyses were conducted on the change in F1, in mels, in each phase compared with baseline productions. The baseline F1 value was calculated as the mean F1 value of the last 10 productions in the baseline phase of each word. The statistical model includes the last 10 trials of each word from the baseline phase, the last 10 trials of each word from the hold phase (as a measure of maximum adaptation, taken when participants have the maximum exposure), and the first 10 trials of each word from the washout phase (as a measure of persistence of adaptation, taken before learning is washed out by the return to veridical feedback).

Linear mixed-effects models were performed in R using the lme4 package (28, 29). Models were built incrementally with maximum likelihood comparisons using the anova function in the lmerTest package (30) to determine which fixed effects remain in the model. Potential fixed effects included the formant shift applied (levels: F1 up, F1 down, no shift), phase of the experiment (levels: baseline, hold, washout), and the interaction between shift and phase.

Models also included participants as a random effect. Post hoc tests were performed with the emmeans package in R (31), using the Tukey adjustment for multiple comparisons. Reported means are estimated means, plus or minus standard error, in distance from baseline in mels. Effect sizes were calculated using the function eff_size in the emmeans package, based on the final model.

IRB STATEMENT

All procedures were approved by the Institutional Review Board at the University of Wisconsin-Madison.

RESULTS

Experiment 1

Mandarin Chinese (Mandarin) is a lexical tone language with four lexical tones. Tone has a high functional load in the language (32); minimal pairs where the tone is the sole difference are common. In this experiment, we test whether lexical tone can serve as motor context for sensorimotor learning in vowels. The results from *experiment 1* show that, as a group, Mandarin speakers can learn simultaneous,

opposing adaptations of the same segmental content using tone as context (Fig. 2A). During the hold phase, participants overall adapted their F1 up in opposition to a downward shift (26.8 ± 9.3 mels compared with baseline, $P < 0.0001$), and adapted their F1 down in opposition to an upward shift (-41.9 ± 9.2 mels, $P < 0.0001$), but did not change their production of the unshifted word (3.8 ± 9.2 mels, $P = 1.00$).

The adaptive responses remained in early washout: participants maintained higher F1 in the shift-down condition (27.5 ± 9.3 mels compared with baseline, $P < 0.0001$), lower F1 in the shift-up condition (-44.7 ± 9.0 mels, $P < 0.0001$), and showed no change in the no shift condition (-1.6 ± 9.2 mels, $P = 1.00$). There was no significant change between hold and early washout in any shift condition ($P > 0.98$ for all shifts).

Crucially, all shift conditions were significantly different from each other in the expected direction during both the hold and washout phases (all $P < 0.0005$); the difference between shift up and shift down was large in both the hold (Cohen's $d = 1.33$) and washout phases (Cohen's $d = 1.39$). These results indicate that Mandarin speakers learned three different adaptations on three different words that were differentiated by lexical tone alone, and support the hypothesis

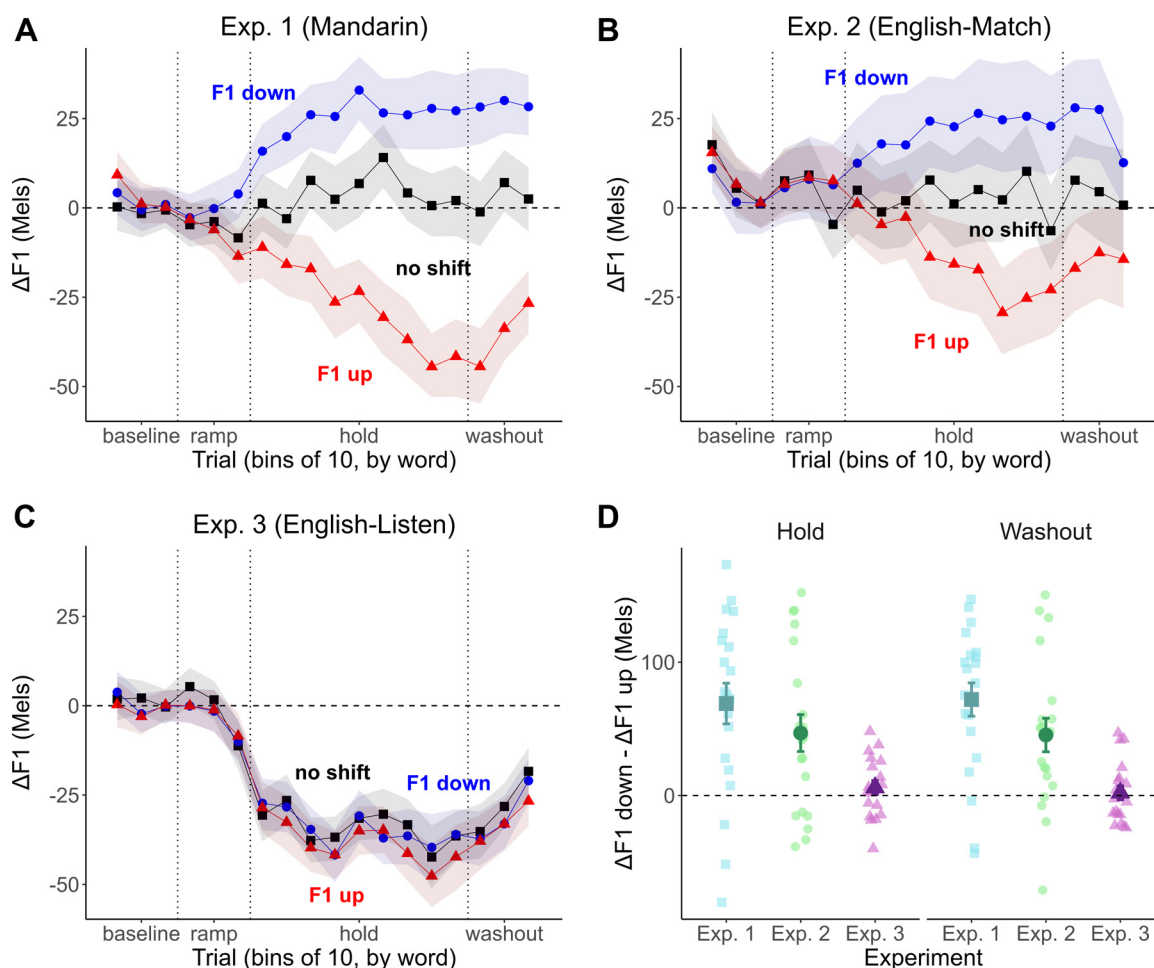


Figure 2. Main experimental results. **A:** change from baseline for *experiment 1* (Mandarin). **B:** change from baseline for *experiment 2* (English, pitch matching). **C:** change from baseline for *experiment 3* (English, listening only). **D:** difference in change from baseline between the shift down condition and the shift up condition ($\Delta F1 \text{ down} - \Delta F1 \text{ up}$) for each experiment. The left three columns show the hold phase and the right three columns show the washout phase.

that the lexical tone in Mandarin can provide context for sensorimotor learning of segments. (For additional analyses of the interaction between lexical tone and adaptation magnitude, see the Supplemental analysis and Supplemental Fig. S1A.)

Experiment 2

Experiment 1 provides support for the idea that lexical f_0 can serve as context to differentiate between segmentally identical motor plans. However, it is unclear whether the ability of f_0 to enable this learning is restricted to lexical pitch, or whether f_0 universally provides sensorimotor context for the segmental content of speech. In *experiment 2*, we test this by extending the simultaneous adaptation paradigm used in *experiment 1* to English speakers producing the word “head” with three different arbitrary (nonlexical) f_0 levels. Results from *experiment 2* show that, as a group, English speakers learn simultaneous, opposing adaptations with arbitrary produced f_0 as context (Fig. 2B). During the hold phase, participants overall adapted their F1 up in opposition to a downward shift (26.4 ± 11.6 mels compared with baseline, $P = 0.003$); however, the change in F1 in opposition to an upward shift, while numerically in the expected direction, did not reach statistical significance (-18.6 ± 11.6 mels, $P = 0.06$). Participants also did not change their production of the unshifted word (-4.3 ± 11.6 mels, $P = 1.00$).

During the washout phase, participants retained the adaptation to the downward shift (27.2 ± 11.6 mels compared with baseline, $P = 0.002$); for the upward shift condition, change in F1 was numerically in the expected direction but did not reach statistical significance, similar to the hold phase (-18.6 ± 11.6 mels, $P = 0.14$). Participants also did not change their production of the unshifted word (8.5 ± 11.6 mels, $P = 0.95$). There was no significant change between the hold and washout phases in any shift condition (all $P > 0.6$).

Crucially, upward shift and downward shift significantly differed from each other in the expected direction during both the hold and washout phases (both $P < 0.0001$). The difference between shift up and shift down was large in both the hold (Cohen’s $d = 1.01$) and washout phases (Cohen’s $d = 1.06$), but less so than for Mandarin speakers. These results suggest that speakers were able to learn simultaneous opposing adaptations when cued by matching different pitches, but not as robustly as Mandarin speakers given the lack of any significant change in the shift-up condition relative to baseline. (For additional analyses of the interaction between target pitch and adaptation magnitude, see the Supplemental analysis and Supplemental Fig. S1B.)

Post hoc analysis: ability to produce distinct pitch categories.

Although the pitch cues were three semitones apart, we observed that some participants had difficulty matching pitch. In particular, most participants had relatively clear high and low categories that aligned with the f_0 of the pure tone cue, but for several participants, the mid tone was not distinct, with many trials produced identically to either high or low tones (see Fig. 3A, bottom). This suggests that for these speakers, the problem lay in the perception of the mid tone as a distinct category, rather than difficulty in achieving high or low f_0 values. That is, these speakers may have been

perceiving the mid tone as “higher” or “lower” compared with the previous tone that they heard and producing a high or low pitch accordingly. As we hypothesized that distinct motor plans are crucial to learn simultaneous opposing adaptations, an inability to distinguish the mid tone as a distinct category might impair learning. To test this post hoc hypothesis, we conducted a further analysis comparing adaptation in participants who were able to consistently produce three distinct pitch categories and participants who were not.

To determine which participants were able to produce distinct pitches, we first extracted the median f_0 from vowel onset to vowel offset in each trial. We iterated a k -means clustering algorithm 1,000 times for each participant, with a target of three clusters ($k = 3$). On each run of the algorithm, we compared the assigned cluster for each trial (low, middle, or high f_0) to the target tone for the trial. We then extracted the mean percent correct classification overall, and the mean percent correct for the mid tone (mid tone trials classified as the middle cluster). There was a large separation between groups in percent mid tone correct (see Fig. 3B); nonmatchers ($n = 8$) were defined as speakers who had less than 60% of the mid tones classified correctly.

As musical training may have played a role in the ability to match pitch (and potentially a role in the likelihood of treating pitch and segmental plans as a cohesive unit), we compared musical background between matchers and nonmatchers using results from the postexperiment musical experience survey (ELMEQ). Pitch-matching ability was related to experience with vocal music: of the 12 matchers, eight had vocal training or experience singing in a choir (of which five also had instrumental experience), and four had experience with an instrument but no voice or choir experience. Of the eight nonmatchers, none had vocal training or experience singing in a choir; five had experience with an instrument.

To determine whether differential adaptation was contingent on pitch-matching ability, we compared matchers and nonmatchers in their magnitude of differential adaptation, calculated for each participant as the mean difference between the shift up and shift down conditions, for both the hold and washout phases (Fig. 3C). Given the small data set, one-tailed Welch’s t tests were used to assess differences between the groups in each phase, where the predicted direction is that matchers would have greater separation between shift conditions than nonmatchers. Although the mean difference between matchers and nonmatchers was numerically different (hold: 49.92 mels for matchers, 42.37 for nonmatchers; washout: 52.67 mels for matchers, 34.53 mels for nonmatchers), there was no statistically significant difference between matchers and nonmatchers in either hold [$t(14.544) = 0.26$, $P = 0.40$] or in washout [$t(12.647) = 0.66$, $P = 0.26$].

Post hoc analysis: comparing magnitude of adaptation to Mandarin speakers.

The less consistent results for adaptation seen in English speakers in *experiment 2* compared with Mandarin speakers in *experiment 1* suggest that lexical tone and arbitrary pitch may differ in the extent to which they enable differential sensorimotor adaptation to vowel formant perturbations. To

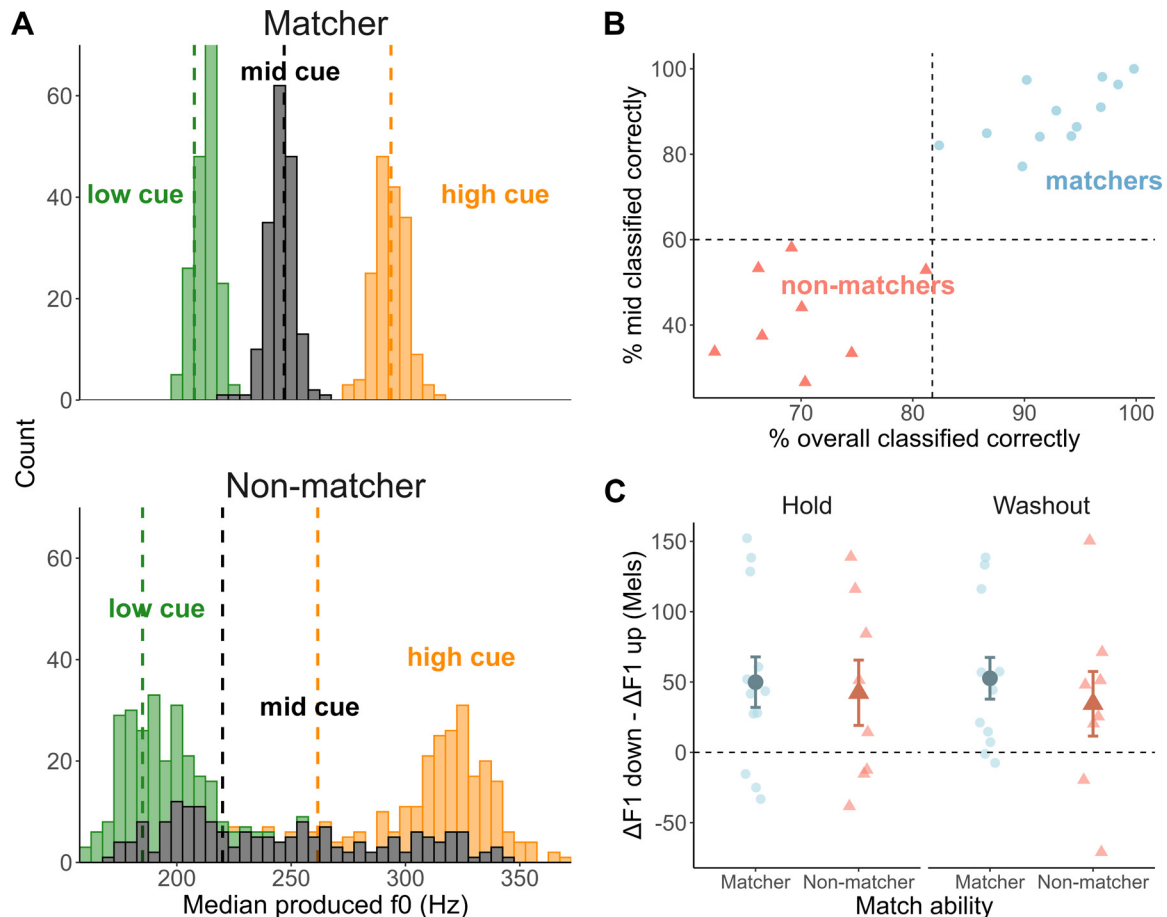


Figure 3. Participants who could match pitch vs. participants who could not match pitch in *experiment 2*. **A**, *top*: histogram of the actual f_0 produced by a participant who matched pitch successfully; received pitch cue is denoted by color; *bottom*, a histogram of the actual f_0 produced by a participant who did not match pitch successfully. Vertical dashed lines denote the pitch of the target tones. **B**: the results of the cluster analysis, which divides speakers into matchers and nonmatchers based on percent overall trials classified correctly (x-axis) and percent mid trials classified correctly (y-axis). **C**: difference in change from baseline between the shift down condition and the shift up condition ($\Delta F1$ down - $\Delta F1$ up) for matchers and nonmatchers.

directly compare the magnitude of simultaneous opposing adaptation in these two contexts, we also conducted a one-tailed t test comparing the two experiments on differential adaptation, defined for each speaker as the mean difference between the shift up and shift down conditions, for both the hold and washout phases. Here, the predicted direction is that Mandarin speakers in *experiment 1* would adapt more than English speakers in *experiment 2* given that pitch has lexical value in the former task and is entirely arbitrary in the latter.

Mean differential adaptation (Fig. 2D) was numerically larger in Mandarin than in English in both the hold (Mandarin: $M = 69.02$ mels, $SD = 68.3$ mels; English: $M = 46.90$ mels, $SD = 61.98$ mels) and washout phases (Mandarin: $M = 72.01$ mels, $SD = 55.78$ mels; English: $M = 45.41$ mels, $SD = 56.29$ mels). However, this difference was not statistically significant in either phase [hold: $t(37.65) = -1.07$, $P = 0.15$; washout: $t(37.997) = -1.50$, $P = 0.07$].

Experiment 3

The results from *experiment 2* indicate that even arbitrary f_0 can provide context for sensorimotor adaptation of segments,

suggesting that lexical relevance is not a necessary condition for context-dependent adaptation in speech. However, unlike in *experiment 1*, the speakers in *experiment 2* also heard a pitch cue before the trial, in addition to producing a different pitch. In this experiment, we test whether merely hearing a distinct pitch cue, without subsequent planning and production, provides sufficient context to anchor simultaneous opposing adaptation in English speakers. In studies of upper limb control, there is a large body of evidence showing that arbitrary cues unrelated to motor planning are insufficient for motor learning (11, 12, 14, 15). Thus, we predict that participants will not be able to adapt their speech according to external pitch cues.

The results from *experiment 3* show that English speakers cannot learn simultaneous, opposing adaptations of the same segmental content when they only receive an auditory cue (Fig. 2C). During the hold phase, participants overall produced a lower F1 than in baseline for all shifts (no shift: -34.3 ± 7.8 mels; shift down: -37.7 ± 7.8 mels; shift up: -39.6 ± 7.8 mels; all shifts significantly different from baseline, $P < 0.0001$). There were no significant differences between shifts (all $P > 0.86$).

In the washout phase, F1 remained lower than in the baseline phase for all shifts (no shift: -38.8 ± 7.8 mels; shift down: -37.3 ± 7.8 mels; shift up: -43.3 ± 7.8 mels; all shifts significantly different from baseline, $P < 0.0001$). As during the hold phase, there were no significant differences between shifts (all $P > 0.92$). Thus, although participants overall changed their produced F1, participants did not change to oppose the perturbations tied to the different pitch cues that they heard.

DISCUSSION

In this study, we tested the efficacy of f_0 as a contextual cue for simultaneous adaptation to opposing vowel formant perturbations. In *experiment 1*, we showed that Mandarin speakers can learn simultaneous, opposing adaptations to the same segmental content when the direction of the perturbation is consistently cued by a lexical tone. In *experiment 2*, we showed that English speakers who heard and then matched an arbitrary pitch that carried no linguistic meaning were similarly able to adapt to the opposing perturbations tied to the separate pitches. Together, these experiments indicate that f_0 in general provides sufficient sensorimotor context for simultaneous, opposing adaptation of vowels, regardless of the presence or absence of lexical information. Finally, in *experiment 3*, we showed that English speakers who simply listened to a pitch cue before producing a word could not learn simultaneous opposing adaptations, even though this cue perfectly predicted the upcoming perturbation. This result reinforces findings from reaching that cues that are unrelated to the movement cannot serve as a context for simultaneous adaptation to opposing sensory perturbation (12–14), expanding this finding to speech articulation.

Due to the arbitrary nature of the pitch matching, *experiment 2* is particularly informative because it isolates f_0 as the context for adaptation. In *experiment 1*, there were multiple possible paths to learning simultaneous opposing adaptations. First, the three words used in *experiment 1* are three distinct, yet segmentally identical lexical entries in Mandarin: it is possible that different lexical entries simply have different motor plans, as opposed to one shared plan for segmental articulation (33). Second, although the segments are phonologically the same, there are consistent phonetic differences between the words (e.g., shorter duration in *tone 4* words; 34) that could also suggest the segmental parts of these words have distinct motor plans. In either of these scenarios, this experiment would not involve simultaneous adaptation of the “same” movement, since the *ei* in each word would be a distinct unit. However, in *experiment 2*, participants were saying the same word (*head*) on every trial, with no differences in lexical status or intrinsic duration, isolating f_0 planning as the motor context for sensorimotor adaptation. Furthermore, the success of arbitrary f_0 in enabling contextual learning suggests that all speech-related f_0 can provide context for segments, specifically including intonation. However, it is unclear whether the same pattern would hold for naturally planned, nonlexical intonation contours, which often stretch over many segments or even many words. Additional studies that examine intonation as a motor context would provide valuable insight on this issue.

It is possible that the aforementioned additional distinctive characteristics of Mandarin words contributed to the numerically larger adaptation response in Mandarin speakers. Although the difference in adaptation magnitude between Mandarin speakers in *experiment 1* and English speakers in *experiment 2* did not reach statistical significance, there was a fairly substantial numerical difference between *experiment 1* (69 mels during hold, 72 mels during washout) and *experiment 2* (47 mels during hold, 45 mels during washout), and a difference in effect size (Mandarin Cohen’s $d = 1.33$, English Cohen’s $d = 1.01$, both during hold). One possible reason for the lack of statistical difference is that our sample sizes were too small to adequately power an analysis of adaptation magnitude between the two groups; this seems like a plausible path given the difference in effect sizes. It may also be the case that differential adaptation in Mandarin was somewhat hindered by overlap between the contextual movements: the three Mandarin tones used in this study are traditionally analyzed as H (*tone 1*), LH (*tone 2*), and HL (*tone 4*). Although the overall contour of each tone is quite distinct from the others, they all contain some movement toward high tone. This contrasts with English pitch matching, which was essentially three level tones at H, M, and L, and thus did not have phonetic or phonological overlap. Further studies with larger sample sizes or that examine the effects of contextual overlap may shed light on this issue.

Another possible explanation is that similar learning did occur for English and Mandarin speakers, but context-dependent learning in English was slower than in Mandarin. The model in the study by Avraham et al. (35) proposes that context-dependent sensorimotor adaptation is a motor instantiation of associative learning, linking together two well-studied functions of the cerebellum. Although the motor control system is able to learn in highly specific contexts, the separate contexts (i.e., associations between context as conditioned stimulus and target movement as conditioned response) have to be established first. In this, Mandarin likely has an advantage in that the three target sequences are different words, and are thus more likely to already have three separate forward models. The English speakers, on the other hand, may have had to establish the association between arbitrary f_0 differences and the oral articulators first. However, the English speakers in the match group did reach a plateau in their adaptation, with little change over the last 30 trials, suggesting that it is unlikely that the need to form novel associative contexts alone drove the (numerical) difference in adaptation magnitude between English and Mandarin. Longer time to plateau has been observed in context-dependent adaptation in reaching and in speech (12, 17), though Howard et al. (12) note that participants ultimately achieved similar magnitudes of learning. Studies that isolate the availability of previously established contexts or that continue the training phase for longer may provide further insight.

Finally, although these experiments have demonstrated that neither linguistic information nor movement of the same anatomical structure is required for contextual differentiation of speech sounds, it is still unclear exactly what permits differential adaptation. Previous studies in reaching have shown that simultaneous adaptation is also possible

when the contextual movement is merely planned alongside the target movement—ultimately even without the actual execution of the contextual movement (11–13). This indicates that planning alone can generate distinct sensorimotor neural states. As a result, Sheahan et al. (13) proposed that differential adaptation is licensed not by contextual movement(s) per se, but by the generation of separate sensorimotor neural states at the moment of executing the target movement. It has been hypothesized that different sensorimotor states could engage separate predictive models, possibly in the cerebellum, a critical structure for motor learning (12, 35–38). Sensorimotor neural states do not need to be generated precisely simultaneously with the target movement to license contextual learning, but the impact of contextual cues on the sensorimotor state appears to decay after ~500 ms, possibly driven by the cellular architecture of the cerebellum (12, 39). Given the timescales of speech, a separation of greater than 500 ms between segmental and tone planning is unlikely. Thus, under this framing, one could conclude that f_0 is planned either before or simultaneously with segments, such that distinct sensorimotor states are available during segmental planning. This runs contrary to models that posit that tone planning occurs after segmental planning (40–42), and lends some support to models that posit that segments and f_0 are planned in parallel streams (43–47).

However, it could be the case that mere simultaneous execution, with no need for simultaneous planning whatsoever, also generates different sensorimotor states. That is, although the differential adaptation of segments with distinct f_0 could reflect the relative time course of segmental versus f_0 planning, it could also simply reflect the fact that f_0 and segments are executed at the same time. A key question, then, is how task-relevant the contextual cue has to be to affect the relevant sensorimotor state. More concretely: is context-specific adaptation only possible when the contextual cue is sufficiently relevant to the target movement, or is the sensorimotor state of the entire body taken into consideration? The current study expands on previous work that has shown that the contextual movement does not need to use the same effector as the target movement: Reaching studies have found that the movement of one hand can provide context for the movement of the other hand, both when the hands are moving at the same time (11) and when the contextual hand precedes the target hand (10). However, humans frequently use both hands in concert for a single task, which may promote any manual task as relevant to another manual task. Similarly, although arbitrary f_0 is not relevant to language in the same way that lexical tone is, speakers constantly plan and control f_0 alongside segments when they are speaking. Thus, f_0 control in general may be more likely to be regarded as relevant for articulatory control, and thus be included in the sensorimotor context for speech segments. A recent study found that adaptation in upper limb control can be modulated by different speech contexts (48), which may suggest that a similar relationship exists between speech and manual control due to the frequency of co-speech gesture. Studies that test pairings of contextual and target movements that vary on a spectrum of relevance, e.g., pairing speech with other speech movements, hand movements, and foot tapping, could shed light on this question.

Conclusions

In sum, this series of three studies provides evidence that f_0 movements can be used as motor context for the movements for segments, even though they are not produced by the same set of articulators. This is the case for both lexical tone and for arbitrary pitch matching, suggesting that linguistic content is not necessary for f_0 to inform segmental control. Future work examining other uses of f_0 , such as intonation, would provide additional insight on the generality of f_0 as a motor context.

DATA AVAILABILITY

Data, experimental scripts, and analysis scripts are available on OSF (<https://doi.org/10.17605/OSF.IO/V7ZAF>). Some functions rely on additional code available at <https://github.com/carrien/free-speech>.

SUPPLEMENTAL MATERIAL

Supplemental analyses and Supplemental Fig. S1: <https://doi.org/10.17605/OSF.IO/V7ZAF>.

ACKNOWLEDGMENTS

Preprint is available at <https://osf.io/preprints/psyarxiv/mpuvr>.

GRANTS

This work was supported by grants to Caroline A. Niziolek and Benjamin Parrell from the National Science Foundation (Grant BCS 2120506) and National Institute on Deafness and Other Communication Disorders (Grant R01 DC019134), a grant to Robin Karlin from the National Institute on Deafness and Other Communication Disorders (Grant F32 DC019535), and a core grant to the Waisman Center from the National Institute of Child Health and Human Development (Grant P50 HD105353).

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

R.K., E.T., D.-L.T., C.A.N., and B.P. conceived and designed research; E.T. performed experiments; R.K. analyzed data; R.K. interpreted results of experiments; R.K. prepared figures; R.K. drafted manuscript; R.K., D.-L.T., Y.Z., C.A.N., and B.P. edited and revised manuscript; R.K., E.T., D.-L.T., Y.Z., C.A.N., and B.P. approved final version of manuscript.

REFERENCES

1. Houde JF, Jordan MI. Sensorimotor adaptation of speech I: compensation and adaptation. *J Speech Lang Hear Res* 45: 295–310, 2002. doi:10.1044/1092-4388(2002)023.
2. Houde JF, Jordan MI. Sensorimotor adaptation in speech production. *Science* 279: 1213–1216, 1998. doi:10.1126/science.279.5354.1213.
3. Jones JA, Munhall KG. Perceptual calibration of F0 production: Evidence from feedback perturbation. *J Acoust Soc Am* 108: 1246–1251, 2000. doi:10.1121/1.1288414.
4. Tourville JA, Reilly KJ, Guenther FH. Neural mechanisms underlying auditory feedback control of speech. *NeuroImage* 39: 1429–1443, 2008. doi:10.1016/j.neuroimage.2007.09.054.

5. Ghahramani Z, Wolpert DM, Jordan MI. Generalization to local remappings of the visuomotor coordinate transformation. *J Neurosci* 16: 7085–7096, 1996. doi:10.1523/JNEUROSCI.16-21-07085.1996.
6. Krakauer JW, Ghilardi M-F, Ghez C. Independent learning of internal models for kinematic and dynamic control of reaching. *Nat Neurosci* 2: 1026–1031, 1999. doi:10.1038/14826.
7. Simani MC, McGuire LM, Sabes PN. Visual-shift adaptation is composed of separable sensory and task-dependent effects. *J Neurophysiol* 98: 2827–2841, 2007 [Erratum in *J Neurophysiol* 99: 1579, 2008]. doi:10.1152/jn.00290.2007.
8. Munhall KG, MacDonald EN, Byrne SK, Johnsrude I. Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *J Acoust Soc Am* 125: 384–390, 2009. doi:10.1121/1.3035829.
9. Villacorta VM, Perkell JS, Guenther FH. Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *J Acoust Soc Am* 122: 2306–2319, 2007. doi:10.1121/1.2773966.
10. Gippert M, Leupold S, Heed T, Howard IS, Villringer A, Nikulin VV, Sehm B. Prior movement of one arm facilitates motor adaptation in the other. *J Neurosci* 43: 4341–4351, 2023. doi:10.1523/JNEUROSCI.2166-22.2023.
11. Howard IS, Ingram JN, Wolpert DM. Context-dependent partitioning of motor learning in bimanual movements. *J Neurophysiol* 104: 2082–2091, 2010. doi:10.1152/jn.00299.2010.
12. Howard IS, Ingram JN, Franklin DW, Wolpert DM. Gone in 0.6 seconds: the encoding of motor memories depends on recent sensorimotor states. *J Neurosci* 32: 12756–12768, 2012. doi:10.1523/JNEUROSCI.5909-11.2012.
13. Sheahan HR, Franklin DW, Wolpert DM. Motor planning, not execution, separates motor memories. *Neuron* 92: 773–779, 2016. doi:10.1016/j.neuron.2016.10.017.
14. Gandolfo F, Mussa-Ivaldi FA, Bizzi E. Motor learning by field approximation. *Proc Natl Acad Sci USA* 93: 3843–3846, 1996. doi:10.1073/pnas.93.9.3843.
15. Howard IS, Wolpert DM, Franklin DW. The effect of contextual cues on the encoding of motor memories. *J Neurophysiol* 109: 2632–2644, 2013. doi:10.1152/jn.00773.2012.
16. Rochet-Capellan A, Ostry DJ. Simultaneous acquisition of multiple auditory–motor transformations in speech. *J Neurosci* 31: 2657–2662, 2011. doi:10.1523/JNEUROSCI.6020-10.2011.
17. Zeng Y, Niziolek CA, Parrell B. Simultaneous acquisition of multiple auditory–motor transformations reveals supra-syllabic motor planning in speech production (Preprint). *PsyArXiv*, 2023. doi:10.31234/osf.io/ceqan.
18. Cai S, Boucek M, Ghosh SS, Guenther FH, Perkell JS. A System for Online Dynamic Perturbation of Formant Trajectories and Results From Perturbations of the Mandarin Triphthong /iau/. 8th International Seminar on Speech Production, 2008.
19. Okely JA, Deary IJ, Overy K. The Edinburgh lifetime musical experience questionnaire (ELMEQ): responses and non-musical correlates in the Lothian birth cohort 1936. *PLoS One* 16: e0254176, 2021. doi:10.1371/journal.pone.0254176.
20. Niziolek CA, Houde J. wave_viewer: First Release, 2015. doi:10.5281/zenodo.13839.
21. Jones JL, Zalewski C, Brewer C, Luckner J, Drayna D. Widespread auditory deficits in tune deafness. *Ear Hear* 30: 63–72, 2009. doi:10.1097/AUD.0b013e31818ff95e.
22. Hantzsch L, Parrell B, Niziolek CA. A single exposure to altered auditory feedback causes observable sensorimotor adaptation in speech. *eLife* 11: e73694, 2022. doi:10.7554/eLife.73694.
23. ISO 226. 2003(E). *Acoustics—Normal Equal-Loudness-Level Contours*. ISO, 2003. <https://www.iso.org/standard/34222.html>.
24. Takeshima H, Suzuki Y, Ozawa K, Kumagai M, Sone T. Comparison of loudness functions suitable for drawing equal-loudness-level contours. *Acoust Sci Tech* 24: 61–68, 2003. doi:10.1250/ast.24.61.
25. Boersma P, Weenink D. *Praat: Doing Phonetics by Computer* (Online). Version 6.1.35. <http://www.fon.hum.uva.nl/praat/> [2021 Sep 21].
26. Cai S, Beal DS, Ghosh SS, Tiede MK, Guenther FH, Perkell JS. Weak responses to auditory feedback perturbation during articulation in persons who stutter: evidence for abnormal auditory–motor transformation. *PLoS One* 7: e41830, 2012. doi:10.1371/journal.pone.0041830.
27. Parrell B, Agnew Z, Nagarajan S, Houde J, Ivry RB. Impaired feed-forward control and enhanced feedback control of speech in patients with cerebellar degeneration. *J Neurosci* 37: 9249–9258, 2017. doi:10.1523/JNEUROSCI.3363-16.2017.
28. Bates D, Maechler M, Bolker B, Walker S. lme4: Linear Mixed-Effects Models Using Eigen and S4. CRAN, 2014.
29. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2023. <https://www.R-project.org/>.
30. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. *J Stat Softw* 82: 1–26, 2017. doi:10.18637/JSS.V082.i13.
31. Lenth R. emmeans: Estimated Marginal Means, aka Least-Squares Means (Online). CRAN, 2025. <https://CRAN.R-project.org/package=emmeans> [2025 Jan 31].
32. Oh YM, Coupé C, Marsico E, Pellegrino F. Bridging phonological system and lexicon: insights from a corpus study of functional load. *J Phon* 53: 153–176, 2015. doi:10.1016/j.wocn.2015.08.003.
33. Ran Q, Gao K, Liang Y, Xia Q, Wichmann S. Phonetic differences between nouns and verbs in their typical syntactic positions in a tonal language: evidence from disyllabic noun–verb ambiguous words in Standard Mandarin Chinese. *J Phon* 98: 101241, 2023. doi:10.1016/j.wocn.2023.101241.
34. Wu Y, Adda-Decker M, Lamel L. Mandarin lexical tone duration: impact of speech style, word length, syllable position and prosodic position. *Speech Commun* 146: 45–52, 2023. doi:10.1016/j.specom.2022.11.001.
35. Avraham G, Taylor JA, Breska A, Ivry RB, McDougale SD. Contextual effects in sensorimotor adaptation adhere to associative learning rules. *eLife* 11: e75801, 2022. doi:10.7554/eLife.75801.
36. Smith MC, Coleman SR, Gormezano I. Classical conditioning of the rabbit's nictitating membrane response at backward, simultaneous, and forward CS-US intervals. *J Comp Physiol Psychol* 69: 226–231, 1969. doi:10.1037/h0028212.
37. Heald JB, Lengyel M, Wolpert DM. Contextual inference underlies the learning of sensorimotor repertoires. *Nature* 600: 489–493, 2021. doi:10.1038/s41586-021-04129-3.
38. Wolpert DM, Kawato M. Multiple paired forward and inverse models for motor control. *Neural Netw* 11: 1317–1329, 1998. doi:10.1016/S0893-6080(98)00066-5.
39. Ivry RB. The representation of temporal information in perception and motor control. *Curr Opin Neurobiol* 6: 851–857, 1996. doi:10.1016/S0959-4388(96)80037-7.
40. Chen J-Y. The representation and processing of tone in Mandarin Chinese: evidence from slips of the tongue. *Appl Psycholinguist* 20: 289–301, 1999. doi:10.1017/S0142176499002064.
41. Roelofs A. The WEAVER model of word-form encoding in speech production. *Cognition* 64: 249–284, 1997. doi:10.1016/S0010-0277(97)00027-9.
42. Roelofs A. Modeling of phonological encoding in spoken word production: from Germanic languages to Mandarin Chinese and Japanese. *Jpn Psychol Res* 57: 22–37, 2015. doi:10.1111/jpr.12050.
43. Alderete J, Chan Q, Yeung HH. Tone slips in Cantonese: evidence for early phonological encoding. *Cognition* 191: 103952, 2019. doi:10.1016/j.cognition.2019.04.021.
44. Hickok G, Venezia J, Teghipco A. Beyond Broca: neural architecture and evolution of a dual motor speech coordination system. *Brain* 146: 1775–1790, 2023. doi:10.1093/brain/awac454.
45. Wan I-P, Jaeger J. Speech errors and the representation of tone in Mandarin Chinese. *Phonology* 15: 417–461, 1998. doi:10.1017/S0952675799003668.
46. Weerathunge HR, Voon T, Tardif M, Cilento D, Stepp CE. Auditory and somatosensory feedback mechanisms of laryngeal and articulatory speech motor control. *Exp Brain Res* 240: 2155–2173, 2022. doi:10.1007/s00221-022-06395-7.
47. Zeng Y. Spoken Word Production of Mandarin Monosyllabic Words: from Lexical Selection to Form Encoding (PhD thesis). University of Kansas, 2022.
48. Lametti DR, Vaillancourt GL, Whitman MA, Skipper JI. Memories of hand movements are tied to speech through learning. *Psychon Bull Rev*, 1–10, 2025. doi:10.3758/s13423-024-02618-w.