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]

[1]University of Missouri – Department of Speech, Language and Hearing Sciences, [2]Arizona State University – Department of Speech and Hearing Science, [3]Hong Kong University – Academic Unit of Human Communication, Learning, and Development, [4]Waisman Center, University of Wisconsin, [5]University of Wisconsin – Department of Communication Sciences and Disorders

° indicates equal contribution

Abstract

Sensory feedback is crucial for maintaining accurate motor control. One process of movement correction is sensorimotor adaptation, or motor learning in response to perceived sensory errors. Recent studies have demonstrated that people can simultaneously adapt to opposing errors on a single movement (e.g., leftward and rightward errors on a reach) given some context that differentiates when each error occurs. In speech production, linguistic structure (e.g., the same vowel in different words) has been shown to provide sufficient context for adapting to opposing errors, but it is not clear whether this is restricted to the same effectors (i.e. lips, tongue, jaw in the oral cavity) or also includes movements of other effectors used in speech (i.e., the vocal folds in the larynx). While manual reaching studies have shown that contextual movements need not be produced with the same effector as the learning target, they have thus far only tested left-right effector pairs. We present the results of three simultaneous adaptation experiments in speech that examine whether laryngeal movements for pitch control can provide context for oral articulatory movements for vowels. In each experiment, the resonances that correlate with articulator position during vowels were perturbed in three different directions that were predictable given a pitch context. First, Mandarin speakers differentially adapted given pitch contexts that signaled differences in word meaning, suggesting that lexical uses of pitch provide context for vowels. Second, English speakers differentially adapted given different arbitrary pitch matching contexts on the word “head”, suggesting that even non-meaningful pitch movements provide context for vowels. Third, English speakers were unable to differentially adapt when simply listening to a contextual pitch, indicating that mere auditory input of pitch is insufficient. Together, these

results indicate that sensorimotor context for learning can be provided by different effectors than the learning target.

Introduction

Extensive research has shown that sensory feedback plays a crucial role in maintaining accurate motor control, including speech motor control (Houde & Jordan, 2002b, 2002a; J. A. Jones & Munhall, 2000; Tourville et al., 2008). 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 left than reality (Ghahramani et al., 1996; Krakauer et al., 1999; Simani et al., 2007); 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* /hɛd/ to result in a token sounding more like *hid* /hɪd/ (Munhall et al., 2009; Villacorta et al., 2007). Historically, perturbations have been applied consistently to either a single movement, or to all movements in a study. 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 vs. 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 (Gippert et al., 2023; Howard et al., 2010, 2012; Sheahan et al., 2016). 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 or when followed by a rightward reach, and perturbed to the left when associated with a distinct color 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, while purely sensory cues (such as different colored lights) do not (Gandolfo et al., 1996;

Howard et al., 2012, 2013; Sheahan et al., 2016). 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 if movements from a different type of effector can provide sensorimotor context for a target movement.

There has 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", Rochet-Capellan & Ostry, 2011), or simultaneous perturbation of the same syllable in different words (e.g. "*ped*igree" vs. "*ped*icure", Zeng et al., 2023). 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 segmental kinematics could be providing 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. F_0 provides crucial insight on the question of specific sensorimotor adaptation because it creates different motor contexts without directly involving the articulators targeted by the auditory perturbations. In Experiment 1, we test the hypothesis that f_0 as lexical tone in Mandarin can serve as context for differential adaptation 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 non-

lexical 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 non-motor cue. These follow-up studies show that pitch imitation, but not pitch cuing 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.

Experiment 1: Lexical tone in Mandarin Chinese

Mandarin Chinese (Mandarin) is a lexical tone language with four lexical tones. Tone has a high functional load in the language (Oh et al., 2015); minimal pairs where 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.

Methods

Participants

Twenty native speakers of Mandarin (16 women, 3 men, 1 non-binary), ranging in age from 18 to 31 years (median: 23, sd: 3.8), participated in Experiment 1. A sample size of 20 participants 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 (Rochet-Capellan & Ostry, 2011). No participant reported any history of hearing, speech, or neurological disorders. In addition, all participants passed an automated Hughson-Westlake hearing screening (pure-tone thresholds $\leq$ 25 dB HL in both ears at 250, 500, 1000, 2000, and 4000 Hz). 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 gave informed consent. All procedures were approved by the Institutional Review Board at the University of Wisconsin–Madison.

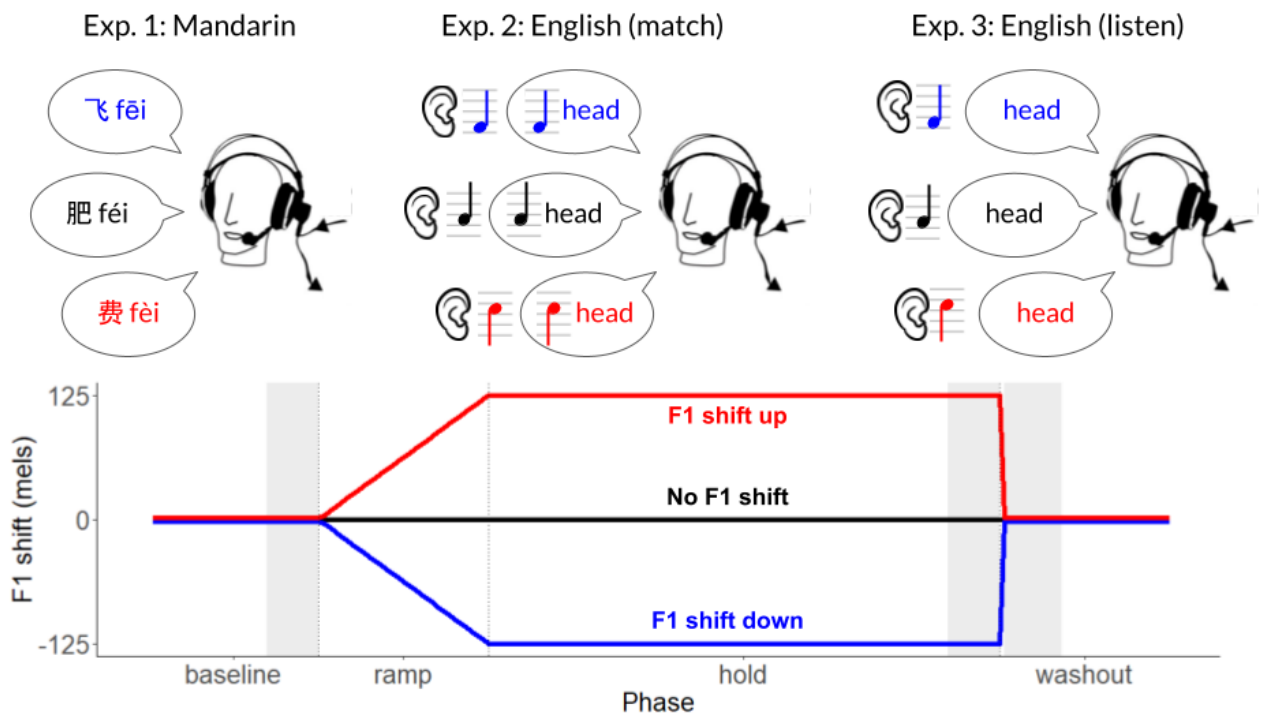
Task

There were three target words in this study, differing only in lexical tone: 飞 *fēi* "fly" (tone 1; high tone); 肥 *féi* "fat" (tone 2; rising tone), and 费 *fèi* "cost" (tone 4; falling tone) (Figure 1, 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 seconds, and there was 1.25 seconds 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 seconds between stimuli). The target words were pseudorandomly ordered within each phase (see below) such that no two sequential trials had the same target.

The vowel *ei* in each word received one of three perturbations to F1 (first resonant frequency of the vowel): F1 up, F1 down, or no perturbation. Formant frequency alterations were applied in mels, a logarithmic transformation of f_0 where equal differences in mels are judged by listeners to correspond to equal changes in pitch. The perturbation received by each word was counterbalanced across participants to the extent possible (each of the possible six permutations assigned to three or four participants). 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 Audapter (Cai et al., 2008). The measured latency of this system was ~19 ms.

The experiment had four phases (Figure 1, bottom): 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).

Figure 1: Schematic of all three experiments. Gray shaded areas indicate windows of analysis.



Data processing

Formant tracking was performed with wave_viewer (Niziolek & Houde, 2015), a MATLAB-based GUI that uses the Praat formant tracking algorithm (Boersma & Weenink, 2017). 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 based on specific parameters set for each participant (LPC order and pre-emphasis); 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). In order to focus on changes in F1 due to sensorimotor adaptation and avoid effects of formant transitions and online compensation, which is typically measurable ~100-150 ms after the onset of F1 perturbation (Larson et al., 2008; Tourville et al., 2008), 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 to 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 (Bates et al., 2014; R Core Team, 2019). Models were built incrementally with maximum likelihood comparisons using the anova function in the lmerTest package (Kuznetsova et al., 2015) 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 participant as a random effect. Post-hoc tests were performed with the emmeans package in R (Lenth, 2019), 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.

Results

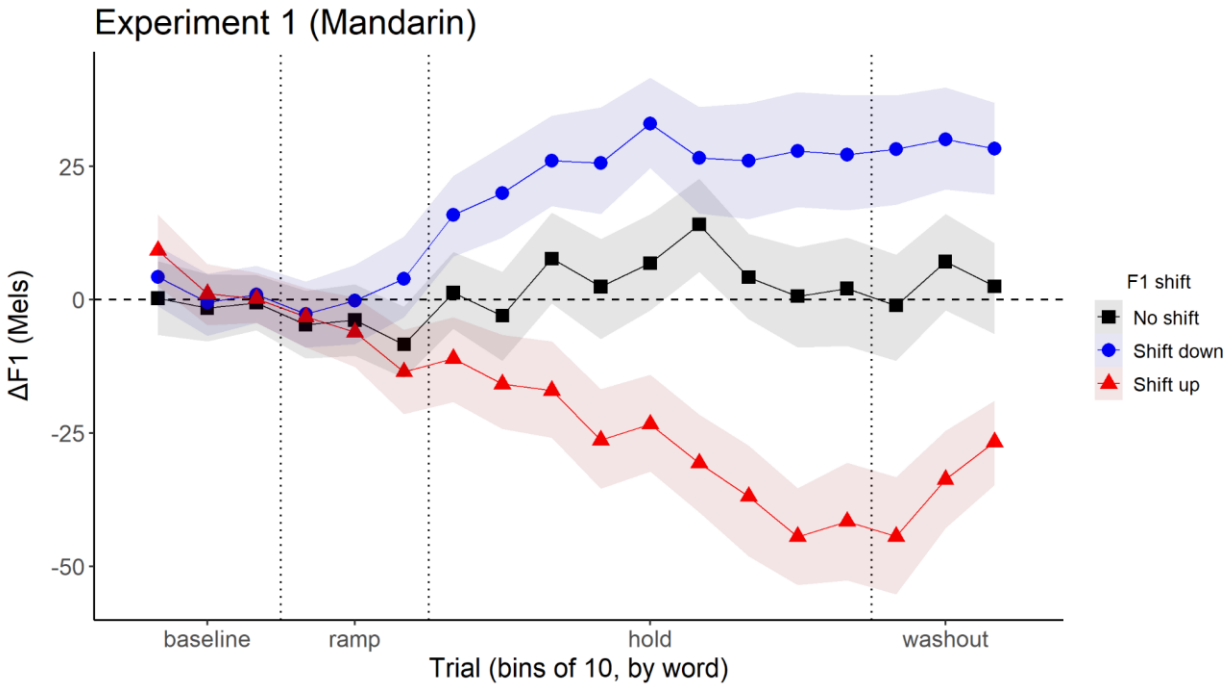
The results from Experiment 1 show that, as a group, Mandarin speakers learn simultaneous, opposing adaptations of the same segmental content using tone as context (Figure 2). During the hold phase, participants overall adapted their F1 up in opposition to a downward shift (26.8 ± 9.3 mels, $p < 0.0001$ compared to baseline), 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, $p < 0.0001$ compared to baseline), 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 that the lexical tone in Mandarin can provide context for sensorimotor learning of segments.

Figure 2: Difference from baseline F1 for each shift condition (Experiment 1).



Experiment 2: Arbitrary pitch in English

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 is universally planned with 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 (non-lexical) f_0 levels.

Methods

Participants

Twenty native speakers of American English (16 women, 4 men), ranging in age from 18 to 30 years (median: 20, sd: 3.9), participated in Experiment 2. Participants underwent the same screening procedures for neurological, speech, and hearing disorders as Experiment 1. 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 gave informed consent. All procedures were approved by the institutional review board at the University of Wisconsin–Madison.

Task

In Experiment 2, there was only one target word, *head* (Figure 1, center). On each trial, 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 Section 3.1.3 below). There was a 50 ms gap before the orthographic stimulus *head* appeared on the computer screen. The orthographic stimulus was on the screen for 1.5 seconds, and there was 1.25 seconds 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 seconds between stimuli). Participants were instructed to match the pitch of their production of the word *head* to the pitch 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 /ɛ/ in *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 word was counterbalanced across participants to the extent possible (each of the possible six permutations assigned to three or four participants). The experimental phases (baseline, ramp, hold, washout) were the same as in Experiment 1. All audio recording, perturbation, and playback equipment was similarly the same as in Experiment 1.

After the experiment was complete, participants completed an abbreviated version of the Edinburgh Lifetime Music Experience Questionnaire (ELMEQ, Okely et al., 2021) to collect information on instrumental and voice training.

Stimuli: Participant-specific pitch cues

For each participant, the frequencies of the pure tone targets were determined at the beginning of the experiment, based on their habitual pitch. In a pretest 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 sonorant and use a wide intonational range, thus providing an approximation of their habitual pitch. F_0 tracks were automatically extracted using wave_viewer (Niziolek, 2021). 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 3 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 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 3 semitones higher than the low tone, and the high tone was 3 semitones higher than the mid tone, for an overall difference of 6 semitones between high and low. 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 (J. L. Jones et al., 2009).

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 (Hantzsch et al., 2022), 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 (*ISO 226: 2003(E): Acoustics—Normal Equal-Loudness-Level Contours*, 2003; Takeshima et al., 2003).

After the tone values were determined, participants completed a practice phase with nine trials (three trials per pitch cue) that were identical in procedure to the remainder of the experiment, with no perturbation to the vowel formants. During the practice phase, 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. Participants were also able to repeat practice on request.

Data processing

The data processing and analysis used in Experiment 2 were identical to those used in Experiment 1.

Statistical analysis

The general statistical analysis for Experiment 2 was identical to that of Experiment 1.

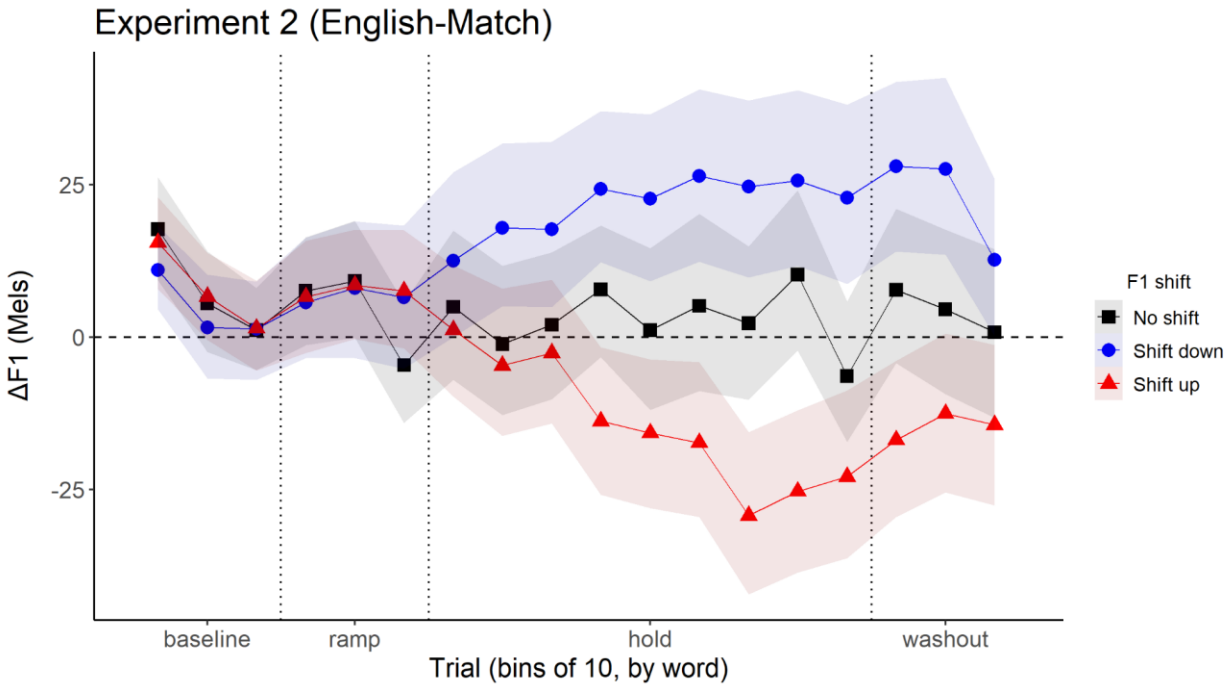
Results

Results from Experiment 2 show that, as a group, English speakers learn simultaneous, opposing adaptations with arbitrary produced f_0 as context (Figure 3). During the hold phase, participants overall adapted their F1 up in opposition to a downward shift (26.4 ± 11.6 mels, p = 0.003 compared to baseline); 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, p = 0.002 compared to baseline); 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.

Figure 3: Change from baseline F1 for each shift condition (Experiment 2).



Post-hoc analysis: Ability to produce distinct pitch categories

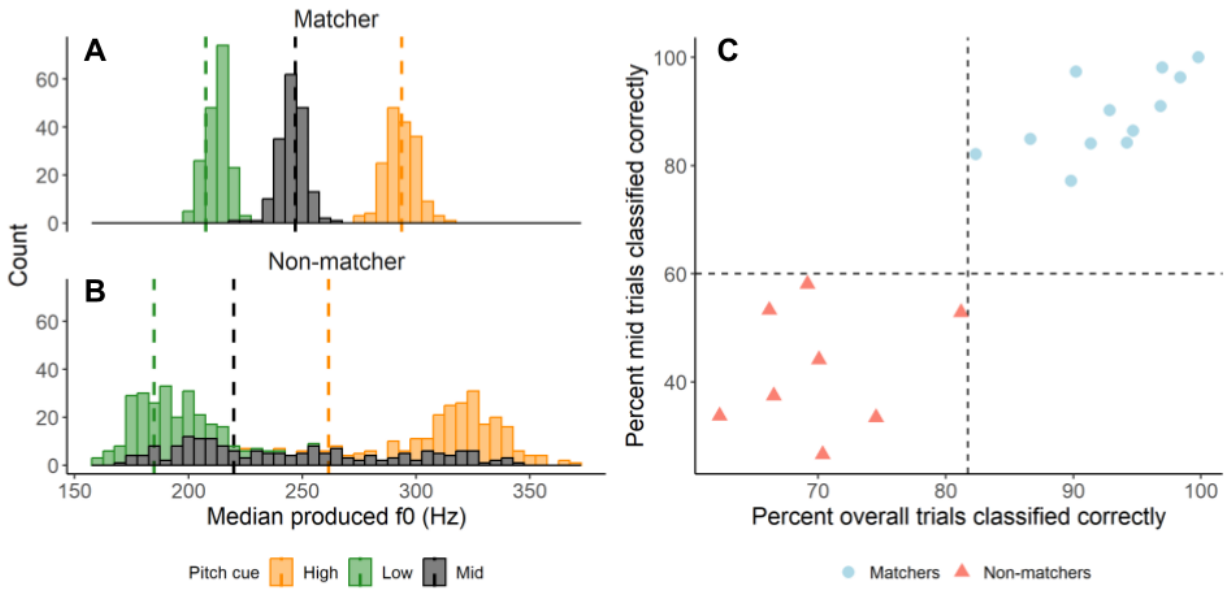
Although the pitch cues were three semitones apart, we observed that some participants had difficulty matching pitch. In particular, most participants had relatively clear high and low categories that aligned with the f_0 of the pure tone cue, but for several participants the mid tone was not distinct, with many trials produced identically to either high or low tones (see Figure 4a and 4b). 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 to 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 1000 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, as well as 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 Figure 4c); non-matchers (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 non-matchers using results from the post-experiment 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 non-matchers, none had vocal training or experience singing in a choir; five had experience with an instrument.

Figure 4: A: Distribution of produced f_0 for a participant who was very successful at matching. B: Distribution of produced f_0 for a participant who did not consistently differentiate the mid tone. C: Comparison of participants who were successful at matching and those who were not.



To determine whether differential adaptation was contingent on pitch-matching ability, we compared matchers and non-matchers 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 (Figure 5). Given the small dataset, 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 non-matchers. While the mean difference between matchers and non-matchers was numerically different (hold: 49.92 mels for matchers, 42.37 for non-matchers; washout: 52.67 mels for matchers, 34.53 mels for non-matchers), there was no statistically significant difference between matchers and non-matchers 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 to Mandarin speakers in Experiment 1 suggests that lexical tone and arbitrary pitch may differ in the extent to which they enable differential sensorimotor adaptation to vowel formant perturbations. To 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 (Figure 6) 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: Pitch as an external signal to adaptation direction

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 prior to 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 (Gandolfo et al., 1996; Howard et al., 2010, 2012, 2013). Thus, we predict that participants will not be able to adapt their speech according to external pitch cues.

Methods

Participants

21 native speakers of American English (14 women, 7 men), ranging in age from 18 to 43 years (median: 23, sd: 6.3), participated in Experiment 3. No participant who participated in Experiment 2 participated in Experiment 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. Participants underwent the same screening procedures for neurological, speech, and hearing disorders as Experiments 1 and 2. Participants were compensated for their participation monetarily. All participants gave informed consent. All procedures were approved by the Institutional Review Board at the University of Wisconsin–Madison.

Task

Experiment 3 was conducted using the same procedure as Experiment 2, in which 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 (Figure 1, right). If participants started to anticipate the presentation of the word *head* and start 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 demonstrated a tendency to inadvertently match pitch. The perturbation received by each word was counterbalanced across participants to the extent possible (each of the possible six permutations assigned to three or four participants).

In order to compare participants here to the subgroups in Experiment 2 who could and could not accurately match the target pitch, participants also completed a short section after the main experiment where they were asked to match pitch, with 10 trials for each of the high, mid, and low pitch cues. Matchers vs. non-matchers were determined from this task using the same procedure as Experiment 2; results from this analysis are included in Supplement A.

Data processing and analysis

Data was processed using the same procedures as Experiments 1 and 2.

Data analysis

Data was analyzed using the same procedures as Experiments 1 and 2.

Results

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 (Figure 4). 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.

Figure 5: Change from baseline F1 for each shift condition (Experiment 3, listening only).

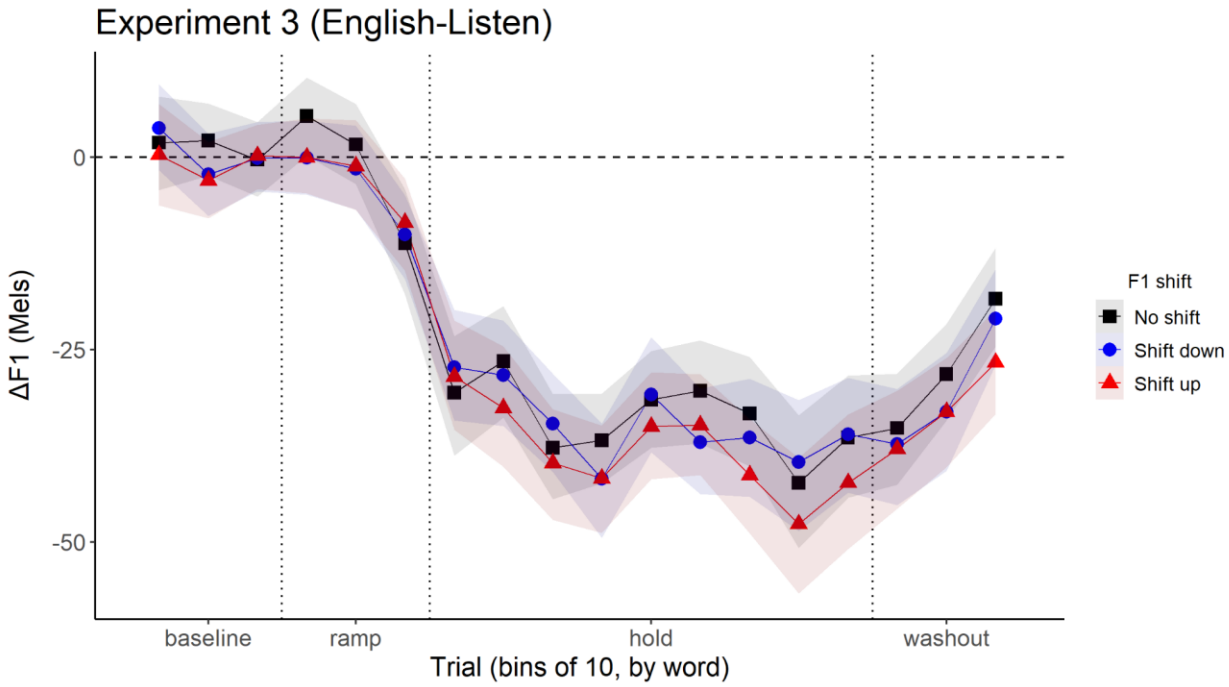
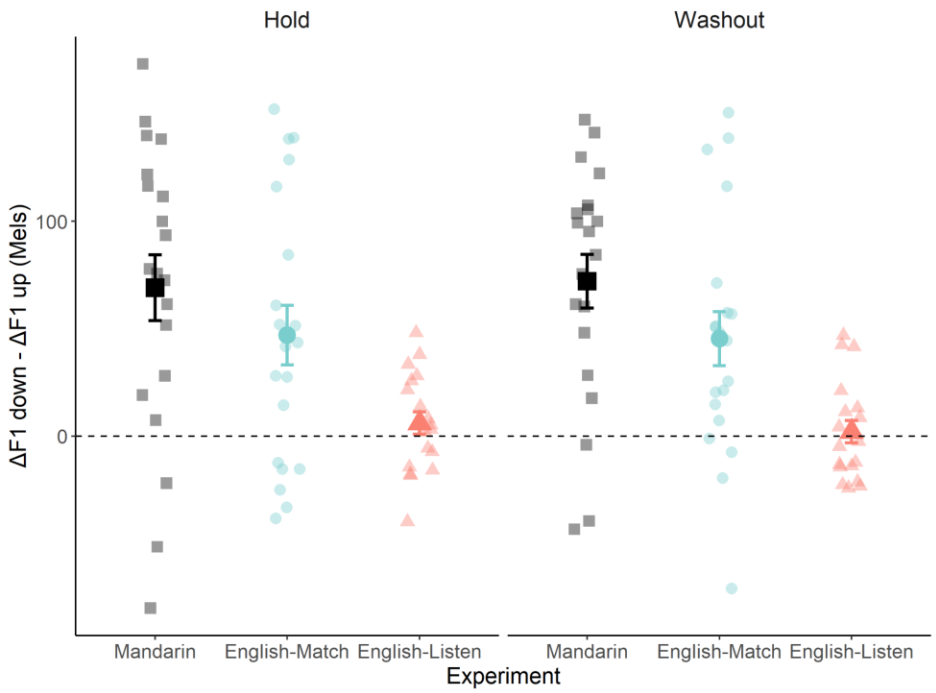


Figure 6: Difference between shift down and shift up for hold and washout phases, all three experiments.



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 lexical information. Finally, in Experiment 3, we showed that English speakers that 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 context for simultaneous adaptation to opposing sensory perturbation (Gandolfo et al., 1996; Howard et al., 2012; Sheahan et al., 2016), 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 (Ran et al., 2023). Second, although the segments are phonologically the same, there are consistent phonetic differences between the words (e.g. shorter duration in tone 4 words; Wu et al., 2023) 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 if the same pattern would hold for naturally planned, non-lexical intonation contours, which often stretch over many segments or even many words. Additional studies that examine intonation as 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), as well as 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 towards high tone. This contrasts with English pitch matching, which were 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.

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 actual execution of the contextual movement (Howard et al., 2010, 2012; Sheahan et al., 2016). This indicates that planning alone can generate distinct sensorimotor neural states. As a result, Sheahan et al. (2016) proposed that differential adaptation is licensed not by contextual movement per se, but by the generation of separate sensorimotor neural states at the moment of executing the target movement. Under this framing, one could conclude that f_0 is planned either before or simultaneously with segments. This runs contrary to models that posit that tone planning occurs after segmental planning (J.-Y. Chen, 1999; Roelofs, 1997, 2015), and lends some support to models that posit that segments and f_0 are planned in parallel streams (Alderete et al., 2019; Hickok et al., 2023; Wan & Jaeger, 1998; Weerathunge et al., 2022; Zeng, 2022).

However, it could be the case that mere simultaneous execution—with no need for simultaneous planning whatsoever—also generates different sensorimotor states. That is, while the differential adaptation of segments with distinct f_0 could reflect the relative time course of segmental vs. 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 movement has to be in order to affect the relevant sensorimotor state. More concretely: is context-specific adaptation only possible when the contextual movement 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 (Howard et al., 2010) and when the contextual hand precedes the target hand (Gippert et al., 2023). However, humans frequently use both hands in concert for a single task, which may

promote any manual task as relevant for 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 (Lametti et al., 2023), 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.

Conclusion

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 motor context.

Data availability statement

Data, experimental scripts, and analysis scripts are available on OSF (https://osf.io/v7zaf/).

Author contribution statement

Robin Karlin: Software; Methodology; Formal analysis; Writing - Original Draft Preparation, Review & Editing;
Emily Tesch: Software; Investigation
Ding-Lan Tang: Methodology; Writing - Review & Editing;
Yuyu Zeng: Writing - Review & Editing
Caroline A. Niziolek: Conceptualization; Funding acquisition; Writing - Review & Editing
Benjamin Parrell: Conceptualization; Funding acquisition; Writing - Review & Editing

Acknowledgments

Funding information

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

Appendix: Supplemental analyses (Experiments 1 and 2)

Previous work has shown that the supraglottal articulation of both monophthongs and diphthongs is influenced by the tone in Mandarin (Erickson et al., 2004; Hoole & Hu, 2004; Li et al., 2023). Specifically, high tones are associated with higher (lower F1) and more front (higher F2) vowels, while lower tones are associated with lower (higher F1) and more back (lower F2) vowels. This parallels the relationship described by *intrinsic* f_0, where vowels articulated higher in the mouth tend to be produced with a higher f_0 and vice versa (W.-R. Chen et al., 2021; Shadle, 1985; D. Whalen et al., 1995; D. H. Whalen & Levitt, 1995). These relationships may affect both the baseline articulation and formant adaptation of the target vowels in experiments 1 and 2, which both used different f_0 targets. We thus further separated the analysis by target tone, examining both baseline F1 productions for each tone, as well as how each tone responded to different F1 shifts.

A.1 Experiment 1: Mandarin

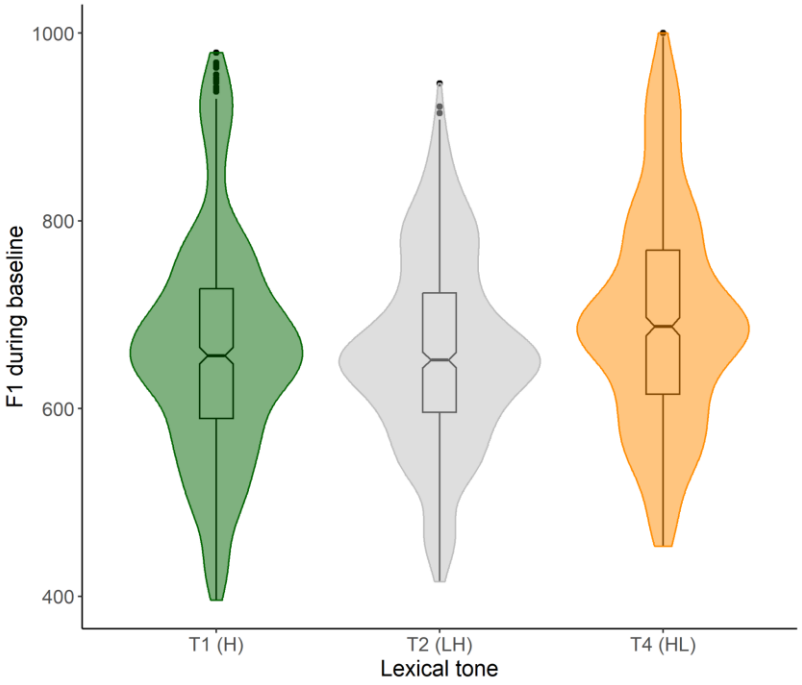
Results are illustrated in Figure A1. Production during the baseline phase will only be affected by intrinsic f_0-vowel height relationships. Here we examine the same vowel window as in the main analysis (25-100 ms). Based on the results of previous studies on the relationship between f_0 and F1 in Mandarin, we would expect that Tone 1 (high level tone) and Tone 4 (falling tone) would have the lowest F1 during this interval, as they have they highest f_0, while Tone 2 (mid-rising tone) would have the highest F1, as it has the lowest f_0. During the baseline phase, there is a significant effect of lexical tone on F1 during baseline trials ($p < 0.0001$). Contrary to prediction, F1 in tone 4 (falling; 696 ± 23.8 mels) is significantly higher than F1 in either tone 1 (high level; 664 ± 23.8 mels) or tone 2 (rising; 660 ± 23.8 mels).

The pairing of lexical tone and shift direction was counterbalanced (to the extent possible) in this study; as such, there would be no overall interference effect of adaptation direction and intrinsic F1. However, we can still examine the magnitude of change in F1 as a function of both shift direction and tone. Production during the hold and washout phases, on the other hand, will also include effects of learning. It may be the case that shifting F1 will have different effects on vowel adaptation, depending on if the shift counters the relationship between F1 and tone, or enhances it.

There is a significant three-way interaction between phase, shift direction, and lexical tone identity ($p < 0.0001$). Specifically, speakers did not adapt F1 in Tone 1 words at all, but showed significant adaptation in both Tone 2 and Tone 4 words. In Tone 1, there is no statistically significant difference between baseline and hold in either the shift up or shift down conditions (both $p > 0.83$). For Tone 2, there is a significant difference between baseline and hold for both shift up and shift down (both $p < 0.0001$); both changes oppose the perturbation. Tone 4 patterns similarly, showing significant difference between baseline and hold for both shift up and shift down (both $p < 0.001$). Tones 2 and 4 are also significantly different from Tone 1 during hold for both shift up (both $p < 0.0001$); only Tone 2 is significantly different from Tone 1 for shift down ($p < 0.0001$). This pattern counters any prediction of adaptation being influenced by enhancing or reducing intrinsic f_0-F1 relationships.

Thus, in Experiment 1, the overall adaptation response to the upward and downward shifts were driven by Tones 2 and 4, while T1 remained stable throughout. This suggests that T1 has some resistance to change, but not one that can be explained by formant shifts enhancing or reducing an intrinsic f_0-F1 relationship.

Figure A1. A: Baseline F1 of each lexical tone. B: changes in F1 compared to baseline, separated by lexical tone.



A.2 Experiment 2: English (Match)

Results from this analysis are illustrated in Figure A2. There is a significant effect of pitch cue on F1 during the baseline phase ($p < 0.0001$). Contrary to what would be predicted by intrinsic

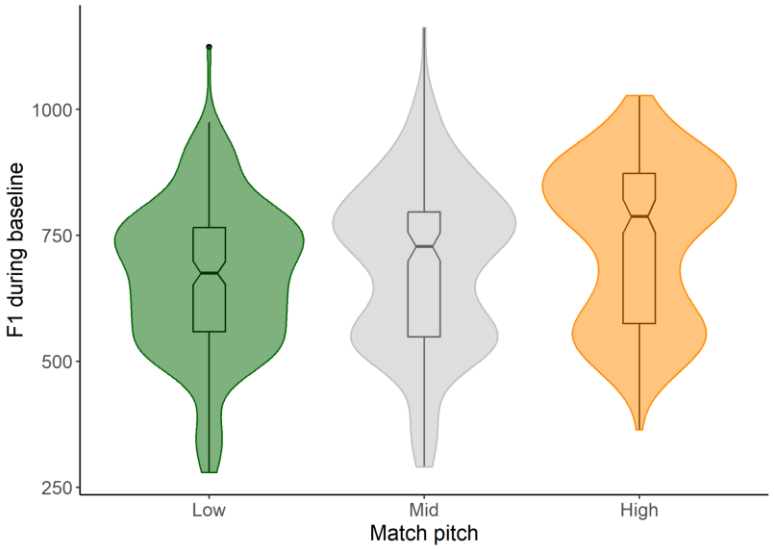
f_0-F1 relationships, *head* with the high pitch cue has the highest F1 (740 ± 32.4 Hz), followed by the mid pitch cue (694 ± 32.5 Hz), and the low pitch cue with the lowest F1 (667 ± 32.4 Hz). All pitch cues are significantly different from each other ($p \leq 0.0001$).

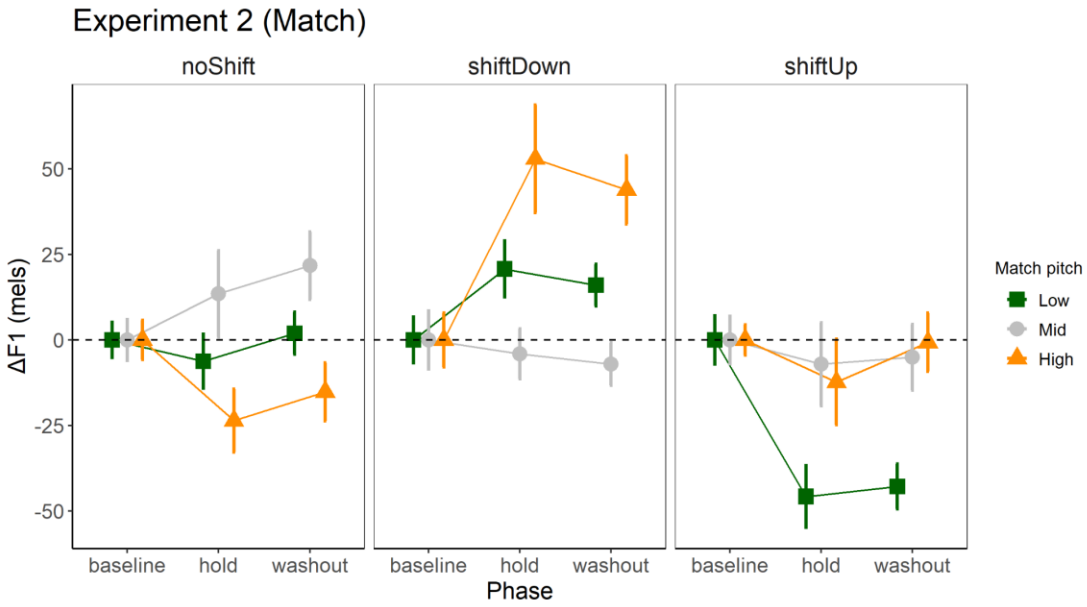
The three-way interaction between shift direction, phase, and pitch cue significantly improves the fit of the model for change in F1. Specifically, the mid pitch cue did not show adaptation in any shift condition, but both the high and low pitch cues do. During the hold phase of the downward shift condition, change in F1 for the high pitch cue is numerically greater (52.9 ± 13.1 mels) than the low pitch cue (22.5 ± 14.0 mels). However, this difference is not statistically significant ($p = 0.29$). In addition, the high pitch cue shows significantly more change than the mid pitch cue ($p = 0.0002$), but the low pitch cue does not ($p = 0.49$).

In contrast, during the hold phase of the upward shift condition, change in F1 for the low pitch cue is numerically lower (-37.8 ± 13.6 mels) than the high pitch cue (-14.4 ± 14.0). Only the low pitch cue shows significant difference between hold and baseline ($p = 0.004$, compared to $p = 0.99$ for both mid and high pitch cues); however, there are no significant differences between any of the pitch cues (all $p > 0.38$).

Together, these results indicate that the pitch cues did not show uniform adaptation in English, and may suggest some interaction of intrinsic f_0-F1 relationships and shift direction. That only the low pitch cue seems to have triggered adaptation in the upward shift condition provides insight into the overall result for this study, which was that there was no statistically significant adaptation in the upward shift condition.

Figure A2. A: Mean F1 of the vowel during baseline, separated by match pitch. B: changes in F1 from baseline, separated by match pitch.





A.3 Discussion

Interestingly, adaptation did not appear to be uniform across different lexical tone or arbitrary pitch cue. In Mandarin, *ei* did not show adaptation to either upward or downward shifts when produced with a high level tone (T1), but did adapt, regardless of shift direction, when produced with a rising tone (T2) and falling tone (T4). Similarly, *head* produced with the mid

pitch cue in Experiment 2 did not show adaptation to either upward or downward shifts, but both high and low pitch cues did. One possible explanation for the mid tone results in Experiment 2 is that the mid pitch cue had the most variable reproduction. As previously noted, differential adaptation depends on the use of distinct motor plans; if some participants did not develop a distinct motor plan for the mid pitch cue, then we would expect that this pitch in particular would have highly impaired adaptation. Any observed differences between low and high pitch cues could then be related to intrinsic f_0-F1 relationships. This explanation does not work for the Mandarin results, however, as participants reliably produced all target tones. These results also come with a major caveat, which is that it is possible that some of the patterns would change with additional participants. Due to the counterbalancing of tone/pitch cue and shift direction, there are either six or seven participants per unique tone-shift combination; this is likely insufficient to detect any true effect of intrinsic f_0 or lexical tone on formant adaptation. Future studies are needed to examine the effects of f_0 on the adaptation of other aspects of production.

References

Alderete, J., Chan, Q., & Yeung, H. H. (2019). Tone slips in Cantonese: Evidence for early phonological encoding. *Cognition*, *191*, 103952.

Bates, D., Maechler, M., Bolker, B., Walker, S., & others. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R Package Version*, *1*(7), 1–23.

Boersma, P., & Weenink, D. (2017). *Praat: Doing phonetics by computer* (Version 6.0.26) [Computer software]. http://www.fon.hum.uva.nl/praat/

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2008). A system for online dynamic perturbation of formant trajectories and results from perturbations of the Mandarin triphthong /iau/. *Proceedings of the 8th ISSP*, 65–68.

Chen, J.-Y. (1999). The representation and processing of tone in Mandarin Chinese: Evidence from slips of the tongue. *Applied Psycholinguistics*, *20*(2), 289–301.

Chen, W.-R., Whalen, D. H., & Tiede, M. K. (2021). A dual mechanism for intrinsic f0. *Journal of Phonetics*, *87*, 101063.

Erickson, D., Iwata, R., Endo, M., & Fujino, A. (2004). Effect of tone height on jaw and tongue articulation in Mandarin Chinese. *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*.

Gandolfo, F., Mussa-Ivaldi, F. A., & Bizzi, E. (1996). Motor learning by field approximation. *Proceedings of the National Academy of Sciences*, *93*(9), 3843–3846.

Ghahramani, Z., Wolpert, D. M., & Jordan, M. I. (1996). Generalization to local remappings of the visuomotor coordinate transformation. *Journal of Neuroscience*, *16*(21), 7085–7096.

Gippert, M., Leupold, S., Heed, T., Howard, I. S., Villringer, A., Nikulin, V. V., & Sehm, B. (2023). Prior movement of one arm facilitates motor adaptation in the other. *Journal of Neuroscience*, *43*(23), 4341–4351.

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

Hickok, G., Venezia, J., & Teghipco, A. (2023). Beyond Broca: Neural architecture and evolution of a dual motor speech coordination system. *Brain*, *146*, 1775–1790. https://doi.org/10.1093/brain/awac454

Hoole, P., & Hu, F. (2004). Tone-vowel interaction in standard Chinese. *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*.

Houde, J. F., & Jordan, M. I. (2002a). Sensorimotor adaptation of speech I. *Journal of Speech, Language, and Hearing Research*.

Houde, J. F., & Jordan, M. I. (2002b). Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research*, *45*, 295–310.

Howard, I. S., Ingram, J. N., Franklin, D. W., & Wolpert, D. M. (2012). Gone in 0.6 seconds: The encoding of motor memories depends on recent sensorimotor states. *Journal of Neuroscience*, *32*(37), 12756–12768.

Howard, I. S., Ingram, J. N., & Wolpert, D. M. (2010). Context-dependent partitioning of motor learning in bimanual movements. *Journal of Neurophysiology*, *104*(4), 2082–2091.

Howard, I. S., Wolpert, D. M., & Franklin, D. W. (2013). The effect of contextual cues on the encoding of motor memories. *Journal of Neurophysiology*, *109*(10), 2632–2644.

ISO 226: 2003(E): Acoustics—Normal Equal-Loudness-Level Contours. (2003).

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *The Journal of the Acoustical Society of America*, *108*(3), 1246–1251.

Jones, J. L., Zalewski, C., Brewer, C., Lucker, J., & Drayna, D. (2009). Widespread auditory deficits in tune deafness. *Ear and Hearing*, *30*(1), 63.

Krakauer, J. W., Ghilardi, M.-F., & Ghez, C. (1999). Independent learning of internal models for kinematic and dynamic control of reaching. *Nature Neuroscience*, *2*(11), 1026–1031.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmerTest.' *R Package Version*, *2*(0).

Lametti, D. R., Vaillancourt, G. L., Whitman, M., & Skipper, J. I. (2023). *Memories of hand movements are tied to speech through learning*.

Larson, C. R., Altman, K. W., Liu, H., & Hain, T. C. (2008). Interactions between auditory and somatosensory feedback for voice F 0 control. *Experimental Brain Research*, *187*, 613–621.

Lenth, R. (2019). *emmeans: Estimated Marginal Means, aka Least-Squares Means*.

https://CRAN.R-project.org/package=emmeans

Li, C., Al-Tamimi, J., & Wu, Y. (2023). TONE AS A FACTOR INFLUENCING THE DYNAMICS OF DIPHTHONG REALIZATIONS IN STANDARD MANDARIN. *20th International Congress of Phonetic Sciences (ICPhS)*.

Munhall, K. G., MacDonald, E. N., Byrne, S. K., & Johnsrude, I. (2009). Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *The Journal of the Acoustical Society of America*, *125*(1), 384–390.

Niziolek, C. A. (2021). *wave_viewer: First release* (Version 1.1) [Computer software]. https://doi.org/10.5281/zenodo.593003

Niziolek, C. A., & Houde, J. (2015). *Wave_Viewer: First Release* [Computer software]. https://doi.org/10.5281/ZENODO.13839

Oh, Y. M., Coupé, C., Marsico, E., & Pellegrino, F. (2015). Bridging phonological system and lexicon: Insights from a corpus study of functional load. *Journal of Phonetics*, *53*, 153–176.

Okely, J. A., Deary, I. J., & Overy, K. (2021). The Edinburgh lifetime musical experience questionnaire (ELMEQ): Responses and non-musical correlates in the lothian birth cohort 1936. *Plos One*, *16*(7), e0254176.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ran, Q., Gao, K., Liang, Y., Xia, Q., & Wichmann, S. (2023). 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. *Journal of Phonetics*, *98*, 101241.

Rochet-Capellan, A., & Ostry, D. J. (2011). Simultaneous acquisition of multiple auditory–motor transformations in speech. *Journal of Neuroscience, 31*(7), 2657–2662.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, *64*(3), 249–284.

Roelofs, A. (2015). Modeling of phonological encoding in spoken word production: From G ermanic languages to M andarin C hinese and J apanese. *Japanese Psychological Research*, *57*(1), 22–37.

Shadle, C. H. (1985). Intrinsic fundamental frequency of vowels in sentence context. *The Journal of the Acoustical Society of America*, *78*(5), 1562–1567.

Sheahan, H. R., Franklin, D. W., & Wolpert, D. M. (2016). Motor planning, not execution, separates motor memories. *Neuron*, *92*(4), 773–779.

Simani, M. C., McGuire, L. M., & Sabes, P. N. (2007). Visual-shift adaptation is composed of separable sensory and task-dependent effects. *Journal of Neurophysiology*, *98*(5), 2827–2841.

Takeshima, H., Suzuki, Y., Ozawa, K., Kumagai, M., & Sone, T. (2003). Comparison of loudness functions suitable for drawing equal-loudness-level contours. *Acoustical Science and Technology*, *24*(2), 61–68.

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

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America*, *122*(4), 2306–2319.

Wan, I.-P., & Jaeger, J. (1998). Speech errors and the representation of tone in Mandarin

Chinese. *Phonology*, *15*(3), 417–461.

Weerathunge, H. R., Voon, T., Tardif, M., Cilento, D., & Stepp, C. E. (2022). Auditory and somatosensory feedback mechanisms of laryngeal and articulatory speech motor control. *Experimental Brain Research*, *240*(7–8), 2155–2173.

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics*, *23*(3), 349–366.

Whalen, D., Levitt, A. G., Hsiao, P.-L., & Smorodinsky, I. (1995). Intrinsic F 0 of vowels in the babbling of 6-, 9-, and 12-month-old French-and English-learning infants. *The Journal of the Acoustical Society of America*, *97*(4), 2533–2539.

Wu, Y., Adda-Decker, M., & Lamel, L. (2023). Mandarin lexical tone duration: Impact of speech style, word length, syllable position and prosodic position. *Speech Communication*, *146*, 45–52.

Zeng, Y. (2022). *Spoken Word Production of Mandarin Monosyllabic Words: From Lexical Selection to Form Encoding* [PhD Thesis]. University of Kansas.

Zeng, Y., Niziolek, C. A., & Parrell, B. (2023). *Simultaneous acquisition of multiple auditory-motor transformations reveals supra-syllabic motor planning in speech production*. OSF. https://doi.org/10.31234/osf.io/ceqan