Simultaneous acquisition of multiple auditory-motor transformations reveals supra-syllabic motor planning in speech production

Yuyu Zeng[1], Caroline Niziolek[1, 2, *], Benjamin Parrell[1, 2, *]

[1] Waisman Center, University of Wisconsin–Madison, Madison, United States

[2] Department of Communication Sciences and Disorders, University of Wisconsin–Madison, Madison, United States

* Equal contribution

Abstract

Motor planning forms a critical bridge between psycholinguistic and motoric models of word production. While syllables are often considered the core planning unit in speech, growing evidence hints at supra-syllabic planning, but without firm empirical support. Here, we use differential adaptation to altered auditory feedback to provide novel, straightforward evidence for word-level planning. By introducing opposing perturbations to shared segmental content (e.g., raising the first vowel formant of "sev" in "seven" while lowering it in "sever"), we assess whether participants can use the larger word context to separately oppose the two perturbations. Critically, limb control research shows that such differential learning is possible only when the shared movement forms part of separate motor plans. We found differential adaptation in multisyllabic words, but of smaller size relative to monosyllabic words. These results strongly suggest speech relies on an interactive motor planning process encompassing both syllables and words.

Keywords

word production, speech motor planning, speech motor control, altered auditory feedback, sensorimotor adaptation

Statement of Relevance

One foundational question in spoken language research concerns the motor planning units that are combined to produce speech. We used alterations to speakers' auditory feedback to simultaneously drive opposing changes in the production of the same syllable depending on its word context (e.g., "pedigree" → "padigree"; "pedicure" → "pidicure"). This differential learning is incompatible with current psycholinguistic speech production models that assume a simple concatenation of syllables, but demonstrates the existence of supra-syllabic motor planning, potentially corresponding to words. However, differential learning was smaller in multisyllabic words relative to monosyllabic words, where learning did not conflict at the syllable level, implying an interactive motor planning process across syllabic and supra-syllabic levels. By revealing supra-syllabic level motor planning, our results call for an updated view of planning units in speech production, and are relevant to research on the mental representation of words, designing rehabilitation programs, and second-language production training.

General introduction

Motor planning is a critical bridge between psycholinguistic and motor control models of speech production. Psycholinguistic models typically end with selecting motor plans (e.g., Levelt, 1999), and motor speech models explain how these plans are translated into movements and acoustic signals (see Parrell et al., 2019 for an overview). Both classes of models frequently position syllables as the central unit of motor planning; one oft-cited example shows high-frequency syllables are produced faster than low-frequency counterparts (Cholin et al., 2011). However, there is evidence that planning may alternatively rely on smaller and larger units. Analyses of speech errors perceptually (e.g., *mell wade* for *well made*, involving sound exchanges; Shattuck-Hufnagel & Klatt, 1979) and articulatorily (e.g., co-producing /t/ and /k/ in the onset of *cop top*; Pouplier, 2007) imply planning sub-syllabically at a phonemic/gestural level. In contrast, anticipatory coarticulation can span multiple syllables (e.g., lip rounding in anticipation of a rounded vowel /u/ in *lee scoot;* Perkell & Matthies, 1992), suggesting planning across multiple syllables.

Despite its foundational role in psycholinguistic and motoric models of speech production, our understanding of the units and scope of speech motor planning is incomplete. Current evidence relies on relatively indirect measures that suffer from interpretive ambiguities (e.g., speech errors, reaction times, and timing differences). For example, the aforementioned exchange error could be attributed to psycholinguistic planning or motor execution. An alternative method that more directly assesses the scope of motor planning is needed. The sensorimotor adaptation procedure, which evokes learned changes to stored movement plans, is well-suited for this task. In speech, this procedure induces real-time perturbations to speakers' formants (the primary acoustic correlates of vowels; F1, F2, and F3 in Figure 1), so they hear a different vowel quality (e.g., "bed" is heard more like "bad"). Over repeated exposures, speakers adapt to counteract the perturbation (e.g., by changing the production of "bed" to be more like "bid"). This learning transfers to untrained syllables sharing the trained vowel (e.g., training on "bé" transfers to "pé"), offering indirect evidence for sub-syllabic planning (Caudrelier et al., 2018).

A more direct test of planning units occurs when opposing perturbations are applied to the same movement in different contexts (e.g., increasing F1 in "bed" but decreasing F1 in "head"). In these cases, learning is only possible when the motor system associates each perturbation with a unique context; without such an association, the opposing perturbations cancel out, and no learning occurs. Critically, this differentiating context must form part of the motor plan for the perturbed movement (Sheahan et al., 2016), as arbitrary cues (e.g., target color) and even kinematic contextual differences unrelated to planning are insufficient to drive differential learning (Howard et al., 2013). Put differently, motor plan differences are necessary and sufficient to drive differential adaptation under opposing perturbations.

Therefore, the presence or absence of differential learning can characterize the planning scope: i.e., whether or not a given context is part of the motor plan.

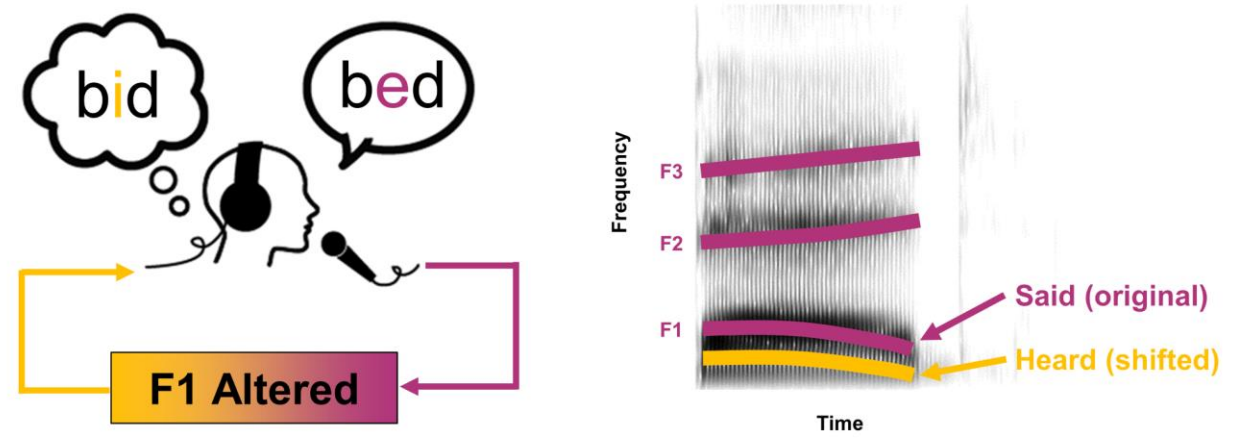

Figure 1. *The speech sensorimotor adaptation procedure. Left: a speaker's produced signal is altered and played back via headphones in real time. Right: details of the modification of the speech signal. The darker bands are energy concentrations at specific frequencies, corresponding to vowel formants (labeled F1, F2, and F3), the primary acoustic correlates of vowels. In the example, the F1 of the word "bed" is perturbed downward to sound more like "bid".*

In speech, speakers exposed to opposing perturbations in "bed" and "head" adapted separately in each word (Rochet-Capellan & Ostry, 2011), suggesting these monosyllabic words form planning units. If the vowel were the only planning unit, what was learned from one word would be canceled out by the other. Nevertheless, because monosyllabic words are distinct syllables as well as distinct words, it is unclear which context enabled the differential adaptation. More problematically, Osu et al. (2004) found that in a random perturbation schedule, which allowed exposure to the same perturbation on sequential trials (instead of alternating the perturbations), contextual differences unrelated to planning could drive differential adaptation in reaching, though this result has not been replicated (Howard et al., 2013). Correspondingly, the adaptation in Rochet-Capellan & Ostry (2011) might be explained by the sequential exposure to the same perturbation rather than the different syllable contexts.

In brief, previous research indicates phonemes (consonants & vowels) and syllables as motor planning units. However, firm support for planning above the syllable is still lacking. Here, we apply opposing perturbations to investigate the word as a potential unit of speech motor planning. After confirming that speakers adapt to opposing perturbations of the same vowel in distinct monosyllabic words without sequential exposure to the same perturbation (Experiment 1), we introduced opposing perturbations to the same syllable in different disyllabic and trisyllabic words (Experiments 2 & 3). If speech planning incorporates word-level information independently of the syllable/vowel, adaptation should occur when the same syllable forms part of distinct multisyllabic words (e.g., perturbations to

"ped" in "pedicure" and "pedigree" in opposite directions would induce differential learning). Conversely, if planning is purely syllabic, speakers will not adapt, as these words share a single syllabic plan, and learning from one word will nullify the other. Alternatively, if syllabic and word-level planning interact, speakers should exhibit differential adaptation, but adaptation should be reduced compared to monosyllabic words, due to the conflict between word- and syllabic-level planning (differential adaptation vs. canceled-out adaptation).

Methods

All experiments and analyses can be reconstructed using the code shared at https://osf.io/ytwqp/; this link also contains supplementary materials. The Institutional Review Board at the University of Wisconsin–Madison approved all procedures.

Participants

Fifteen speakers participated in Experiment 1 (14 female and 1 non-binary; mean age = 19.27, SD = 1.42, range = [19, 23]); twenty speakers in Experiment 2 (17 female and 3 male; mean age = 19.90, SD = 2.04, range = [18, 27]); and twenty speakers in Experiment 3 (17 female and 3 male; mean age = 21.30, SD = 2.91, range = [19, 27]). All participants were adult native English speakers, with some additionally speaking at least one other language (4 in Experiment 1, 7 in Experiment 2, and 8 in Experiment 3). All participants had no reported history of neurological, speech, or hearing disorders and passed a Hughson-Westlake audiology screening before participation (thresholds ≤ 25 dB hearing level bilaterally in the 250-4K Hz range). Participants were compensated either with course credit or monetary payment.

Procedure and equipment

Participants were seated in front of a computer screen where one word appeared per trial. Participants were instructed to read these stimuli aloud as they appeared on the screen. A Sennheiser MKE 600 microphone recorded the speakers' productions, which were digitized using a Scarlett 2i2 sound card, processed (and in some trials altered by shifting vowel formants, see below) using Audapter (Cai et al., 2008; Tourville et al., 2013), and played back over Beyerdynamic DT 770 PRO closed-back, circumaural headphones. The recording, processing, and playback occurred in near real time (~18 ms delay; measurement following Kim et al., 2020). Speech was processed similarly through Audapter on all trials, regardless of whether a formant perturbation was applied. All stimulus presentation and data collection were done using MATLAB (The MathWorks, Inc.).

On each trial, the target word appeared at the center of the screen for 1400 milliseconds (white text on

a black background). The inter-trial duration was 1250 ms, with a 250 ms jitter. If the produced speech in a trial did not meet a pre-specified intensity threshold, that trial was repeated. The intensity of speech playback varied with the intensity of the produced speech on each trial, but was calibrated for each participant with a targeted average level of ~80 dB SPL. The playback was mixed with speech-shaped noise at ~60 dB SPL to mask air- and bone-conducted feedback.

Before the main experiment (described below), participants completed a brief calibration phase to achieve more accurate formant tracking. The words *bid, bet,* and *bat* were each repeated 10 times (order randomized), and the recordings were used to determine a speaker-specific linear predictive coding (LPC) order, which was then used for identifying vowel formants in Audapter. Following the main experiment, participants completed a short survey of perturbation awareness and received an explanation of the study's purpose. The entire lab visit lasted ~1.5 hours.

Experiment design

All perturbed syllables contained the same vowel /ɛ/. Experiment 1 tested adaptation to opposing F1 perturbations in three monosyllabic words: *head*, *bed*, and *ted*. One word received an upward F1 perturbation, another a downward F1 perturbation, and the third no F1 perturbation. Experiment 2 tested adaptation to opposing F1 perturbations in two disyllabic words whose initial syllables were phonemically identical: *seven* and *sever*. The initial syllable received an upward F1 perturbation in one word and a downward F1 perturbation in the other; the second syllable in both words was unperturbed. One additional unperturbed word, *level*, was included as a filler. Experiment 3 tested adaptation to opposing F1 perturbations in two trisyllabic words whose initial syllables were phonemically identical: *pedigree* and *pedicure*. As in Experiment 2, the initial syllable received an upward F1 perturbation in one word and a downward F1 perturbation in the other; the remaining syllables in both words were unperturbed. Two other unperturbed words (*pedestal* and *carbonate*) were included as fillers, but only the word *pedestal* was analyzed; this way, all three experiments were comparable. Including fillers with different initial syllables in Experiments 2 and 3 minimized the possibility that participants would pre-plan the first syllable before the appearance of the stimulus word. For each experiment, the assignment of formant perturbations to experimental words was roughly counterbalanced across participants (see the supplementary materials for details). Figure 2A shows the schematics of the design.

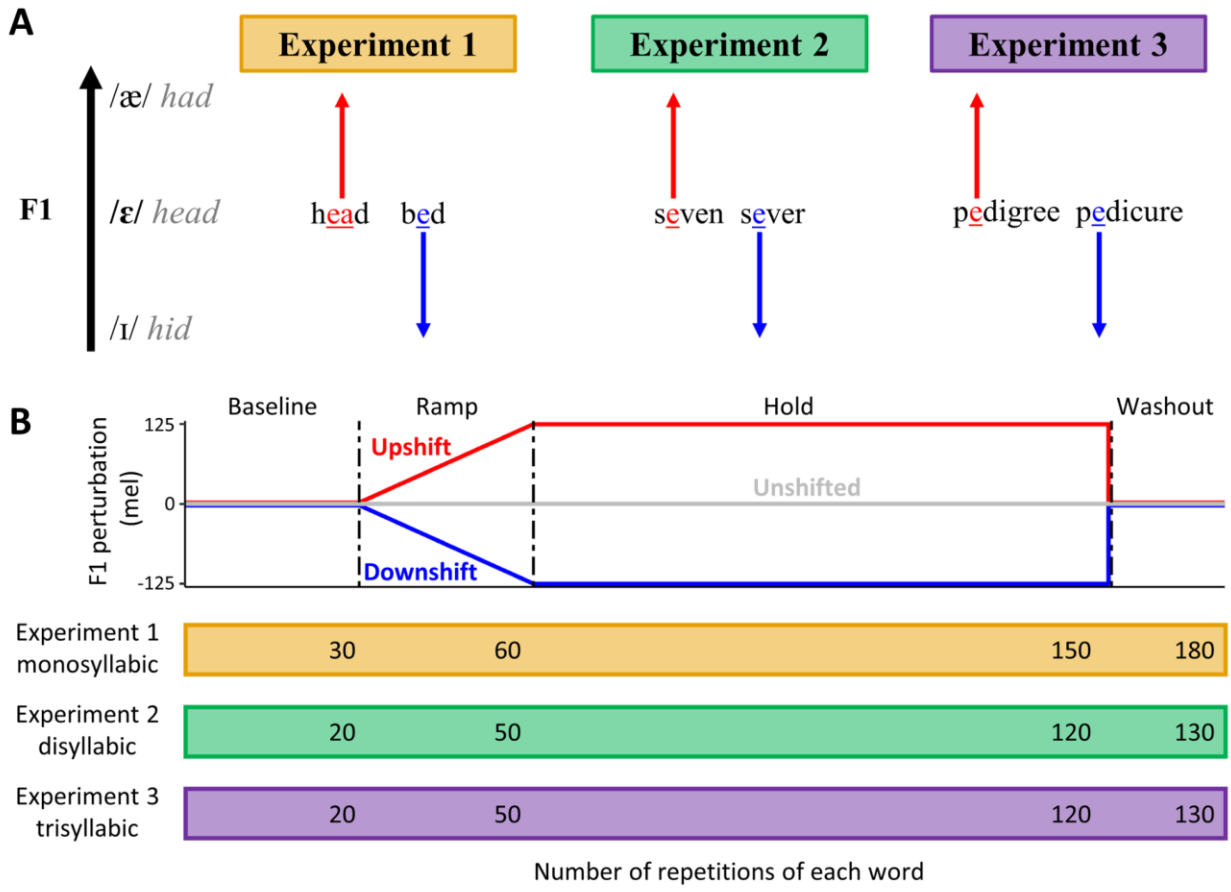

Figure 2. *Experimental design and procedure.* ***A****: Examples of F1 perturbation assignment to words in each experiment.* ***B****: Timeline of each experiment.*

For all experiments, stimulus words were presented in blocks, each containing one repetition of each word. The order of words within each block was randomized. An additional constraint ensured that no two adjacent trials contained the same word across blocks, thus preventing the effects of a random perturbation schedule, including sequential exposure to the same perturbation, as a potential confound for any observed adaptation (cf. Osu et al., 2004). Each experiment consisted of four phases: an unperturbed Baseline phase, a Ramp phase during which the magnitude of F1 perturbation in the auditory feedback was progressively increased across blocks (the F1 perturbation magnitude remained the same within each block), a Hold phase with a constant F1 perturbation of 125 mels, and an unperturbed Washout phase. The magnitude of all perturbations was calculated in mels, a perceptually adjusted frequency scale such that equal changes in mels are perceived as equally distant across all frequencies (Stevens et al., 2005). Experiment 1 consisted of 30 Baseline blocks, 30 Ramp blocks, 90 Hold blocks, and 30 Washout blocks. Experiments 2 and 3 had 20 Baseline blocks, 30 Ramp blocks, 70 Hold blocks, and 10 Washout blocks (Figure 2B). In all experiments, a self-timed break was included every 10 blocks.

Data processing

Vowel onset was identified as the point where periodicity was visible in the waveform and formants were visible in the spectrogram; vowel offset was identified as the point where formants, particularly F1 and F2, were no longer visible. For each vowel, F1 and F2 were tracked every 3 ms using Praat (Boersma & Weenink, 2023) via the wave_viewer package (Niziolek, 2015). Pre-emphasis values and LPC order were set for each participant individually. Errors in formant tracking (e.g., sudden jumps, tracking wrong formants) were corrected by minimal adjustments to these values. Trials with unresolvable formant tracking errors or unintended productions (e.g., coughing, hesitations) were excluded (Experiment 1: mean = 1.22%, SD = 1.46%, range = [0%, 3.52%]; Experiment 2: mean = 0.83%, SD = 1.05%, range = [0%, 3.33%]; Experiment 3: mean = 0.44%, SD = 0.57%, range = [0%, 1.92%]).

For most stimuli, formant measurements from 25% to 75% into a vowel were averaged and converted to the mel scale to obtain a single F1 value for each trial. The exception was the word *level*: in this case, the identified vocalic portion of the speech signal included both the target vowel and the initial consonant /l/ due to the difficulty in consistently segmenting these two sounds. To account for this, we used 40% to 75% of the vocalic segment to collect mean F1 values. This method included the steady-state portion of the vowel and excluded the initial /l/. To capture the F1 change for each speaker, the average F1 values were normalized relative to the mean F1 in the last 10 trials of the Baseline phase for each word.

We also measured fundamental frequency (the rate at which vocal folds vibrate, f_0) and intensity from trials with F1 data, as recorded by Audapter. Outliers, either 3 SDs away from the mean or an abrupt sample within a trial, were removed. Next, a single average value was calculated from 25% to 75% into a vowel (the same window used to calculate F1) for f_0 and intensity. To capture the change for each speaker, the average values were normalized as percentage change relative to the last 10 trials of the Baseline phase for each word.

Statistical analysis

Statistical analysis used mixed-effects models via the lme4 package (Bates et al., 2022; version 1.1-34) in R (R Core Team, 2023; version 4.3.1). Reported *p*-values were calculated by using the lmerTest package (Kuznetsova et al., 2020; version 3.1-3). The α level was set at 0.05 in all analyses. When there were more than two independent variables, the best model was selected by running the buildmer package (Voeten, 2023; version 2.9), which systematically compares models that differ in only one term (Matuschek et al., 2017). The maximal model that fed the backward fitting procedure included all possible interactions of the independent variables as fixed effects; the random effects included both participant and grouping variables (such as word) as random intercepts, plus a maximal random-slope

structure with all experimental variables to avoid inflated Type I error rate (Barr et al., 2013). Unless otherwise specified, discrete independent variables were sum-coded such that the intercept in a selected model corresponded to the average across conditions. The phia package (Rosario-Martinez et al., 2015; version 0.2-1) was used for post-hoc comparisons; the reported *p*-values used the correction procedure in Holm (1979).

Because the proper way to estimate effect size for mixed-effects models is still under development (Correll et al., 2022), we report R^2 (Nakagawa et al., 2017), representing the total variance explained by the selected model and the individual terms. Cohen's d (*d*) is reported as a standardized effect size measure for the difference between two group means. Summary statistics report mean values and standard errors.

The primary analysis for each experiment examines the F1 change in production (labeled ΔF1 in the formula below) across experimental phases to determine whether speakers adapt separately to the opposing perturbations. For Experiments 2 and 3, data from the last 10 blocks in the Hold phase and the first 10 blocks in the Washout phase were analyzed. Because Experiment 1 had more Hold blocks, adaptation was measured from blocks 61-70, such that the number of Hold blocks was identical across experiments (see the shaded areas in Figures 3A, 4A, and 5A). The maximal model included the independent variable Direction that coded the F1 perturbation (Upshift, Downshift, and Unshifted), and the independent variable Phase (Hold, Washout); its formula is shown below. Adaptation should surface as a significant difference between the Upshift and Downshift conditions. Whether a Direction condition differed significantly from zero was also tested.

ΔF1 ~ Direction * Phase + (Direction * Phase | Speaker) + (Direction * Phase | Word)[1]

We also examined whether the word items could predict the adaptation magnitude within each experiment given that: (1) a word's acoustic realization is known to be affected by its lexical frequency (e.g., Pluymaekers et al., 2005), and the words in our experiments differ in their lexical frequency; (2)

[1] Formulae of statistical models in this paper follow R syntax, where the variable to the left of the tilde (~) is the dependent variable, and the independent variables are to the left. An interaction of two independent variables is joined by a colon (e.g., Direction:Phase). When two independent variables are joined by an asterisk (*), its full expression includes individual variables and all their interaction (e.g., Direction * Phase = Direction + Phase + Direction:Phase). For mixed-effects models, fixed effects do not use parentheses, and random effects are placed within parentheses. Within the parentheses, a random intercept is placed to the right of a pipe (|), and its corresponding random slope is specified to the left.

more frequent words are often hypothesized to form a single "chunk" in the mental syllabary (Guenther, 2016), which may influence the adaptation responses.[2] This analysis relates specifically to the adaptation magnitude under F1 perturbation, so only the Upshift and Downshift conditions were included. Since F1 is expected to decrease in the Upshift (negative F1 change) but increase in the Downshift (positive F1 change), we sign-flipped (multiplied by -1) the F1 change values in the Upshift such that the expected adaptation that opposed the perturbation was always positive (labeled $\Delta F1_{word}$ in the formula below). The potential effect of word frequency was examined by placing Word as a fixed effect. Lexical frequency reports were taken from Google Ngram Viewer (Orwant & Brockman, 2019), using its most recent estimates in 2019. The maximal model fed to model selection has the following formula.

$$\Delta F1_{word} \sim \text{Word} * \text{Phase} * \text{Direction} + (\text{Word} * \text{Phase} * \text{Direction} \mid \text{Speaker})$$

In addition to establishing whether adaptation occurred in response to opposing perturbations in each experiment, a critical comparison concerns adaptation size across experiments, as adaptation size differences across experiments would suggest an interaction between syllable and word planning. To compare a single measure across studies, we use differential adaptation, the difference of F1 change between the Upshift and Downshift conditions (Downshift - Upshift; labeled $\Delta F1_{diff}$ in the formula below). For each participant, a single measure of differential adaptation was calculated across 10 blocks during the Hold and Washout phases. The analysis windows were identical to the modeling of adaptation in individual experiments as described above, where the amount of perturbation exposure during the Hold phase was identical across experiments. This analysis started with the maximal model below.

$$\Delta F1_{diff} \sim \text{Phase} * \text{Experiment} + (\text{Phase} * \text{Experiment} \mid \text{Speaker})$$

To preview, the magnitude of differential adaptation in the monosyllabic words in Experiment 1 was larger than that found in the multisyllabic words in Experiments 2 and 3. However, we also expect shorter vowel durations in multisyllabic words than in monosyllabic words (Umeda, 1975), which might affect the magnitude of adaptation. For example, studies of sensorimotor adaptation in reaching report that learning is greater with continuous visual feedback of hand position compared to more limited feedback about the reach endpoint (e.g., Schween et al., 2014). In other words, longer exposure to a perturbation may lead to greater adaptation. Therefore, the duration difference could be a confound for the adaptation size difference across experiments if similar effects hold in speech production. We conducted three supplementary analyses to test this possibility. First, we verified the duration difference across

[2] It is worth noting that even if there is a lexical frequency effect on the adaptation behavior, the adaptation results of the primary analyses should not be affected by lexical frequency, as our experimental design employed counterbalancing of the stimuli.

experiments. The analysis windows were identical to the modeling of adaptation in individual experiments as described above, where the amount of perturbation exposure during the Hold phase was identical across experiments. The dependent variable was the vowel duration in milliseconds (ms). This analysis started with the maximal model below.

$$\text{Duration} \sim \text{Direction} * \text{Phase} * \text{Experiment} + (\text{Direction} * \text{Phase} * \text{Experiment} \mid \text{Speaker}) + (\text{Direction} * \text{Phase} * \text{Experiment} \mid \text{Word})$$

Second, after verifying that perturbed vowels in Experiment 1 indeed had a longer duration than those in Experiments 2 and 3, we added duration to the selected model of differential adaptation across experiments to test if duration predicted differential adaptation. In this case, duration was the average vowel duration for the perturbed syllables. If duration as an independent variable predicts the magnitude of differential adaptation, we need to check whether the difference in differential adaptation across experiments is still attributable to our experimental manipulation in the updated model (i.e., a significant main effect of Experiment). Third, we re-ran the model selection procedure with duration in the maximal model to evaluate differential adaptation size across experiments; its formula is shown below.

$$\Delta F1_{diff} \sim \text{Phase} * \text{Experiment} * \text{Duration} + (\text{Phase} * \text{Experiment} * \text{Duration} \mid \text{Speaker})$$

Lastly, Experiments 2 and 3 showed a global tendency to increase F1 across stimuli, regardless of the perturbation direction. Even in Experiment 1, the Unshifted condition has a positive F1 change, a tendency also seen in previous work (e.g., Rochet-Capellan & Ostry, 2011). Global increases in F1 often accompany clearly articulated speech ("clear speech", henceforth), as produced, for example, in the presence of background noise or when speaking to a listener with hearing loss or from a different language background. Since clear speech often occurs when there is communication difficulty, the masking noise and formant perturbations in our experiments may have induced similar behavior. To test this possibility, we additionally report changes in other speech parameters often associated with clear speech, including f_0, intensity, and duration (Krause & Braida, 2003).

Results

We report results from three experiments testing the scope of speech motor planning via adaptation to altered auditory feedback. In Experiment 1, opposing perturbations are applied to the same vowel in different monosyllabic words. In Experiments 2 and 3, opposing perturbations are applied to the same initial syllable in different multisyllabic words. If the word is a planning unit independent of the syllable, differential adaptation should occur in all three experiments to an equal degree. Conversely, if speech planning relies mainly on syllables and the word is not a planning unit, speakers will exhibit differential

adaptation in Experiment 1 but not in Experiments 2 and 3, where the learning from opposing perturbations to the same syllable would cancel out. Lastly, if the word is a planning unit that interacts with syllable planning, we expect differential adaptation in all experiments but a reduced adaptation size in Experiments 2 and 3 relative to Experiment 1, as syllable-level learning (no adaptation) conflicts with word-level differential adaptation in these two experiments.

Experiment 1

As expected, participants adapted differentially to the opposing perturbations applied to distinct monosyllabic words (cf. Rochet-Capellan & Ostry, 2011). When F1 was perturbed upwards, speakers lowered their produced F1; when F1 was perturbed downwards, speakers raised their produced F1. Adaptation reached a plateau towards the end of the Hold, and then declined after removing the perturbation during the Washout (Figure 3A). All individual speakers' responses mirrored the group average (Figure 3B): 100% produced a higher mean F1 in the Downshift condition compared to the Upshift condition in the Hold phase, and 93% maintained this pattern in the Washout phase. The final selected model (R^2 = 0.22) was:

$$\Delta F1 \sim Direction + Phase + Direction{:}Phase + (Direction + Phase | Speaker)$$[3]

Model results indicated that the F1 change was significantly different between the perturbation conditions (a main effect of Direction, $\chi^2(2) = 45.06$, $R^2 = 0.345$, $p < 0.001$). All three perturbation conditions differed significantly from each other (all $p < 0.001$; Upshift vs. Downshift: $d = 1.31$). Adaptation differed from zero in the Upshift (-22.85 ± 2.52 mels, $p = 0.005$) and the Downshift (33.31 ± 2.54 mels, $p < 0.001$) conditions, but not in the Unshifted (5.09 ± 2.27 mels, $p = 0.55$).

Adaptation magnitude declined from the Hold to the Washout as evidenced by a significant Direction by Phase interaction ($\chi^2(2) = 11.10$, $R^2 = 0.008$, $p = 0.004$), although the Hold-Washout difference was not significant in any of the three perturbation conditions individually (all $p > 0.24$). Nonetheless, adaptation remained significantly different from zero in both the Upshift (Hold: -27.54 ± 3.28 mels, $p = 0.01$; Washout: -18.24 ± 3.04 mels, $p = 0.018$) and the Downshift (Hold: 36.60 ± 2.69 mels, $p < 0.001$; Washout: 30.06 ± 3.64 mels, $p = 0.002$) conditions, but not in the Unshifted condition (Hold: 6.14 ± 2.72 mels, $p = 0.99$; Washout: 4.04 ± 3.03 mels, $p = 0.99$). Post-hoc comparisons found that all perturbation conditions were significantly different from each other during both the Hold and Washout (all $p < 0.011$; Upshift vs. Downshift: $d = 1.50$, 1.12, respectively). No other selected term in the model was significant

[3] The independent variables of the selected models reported in this paper are listed according to their contribution to the overall model fit in descending order, measured in the significance of log-likelihood change.

($p = 0.98$).

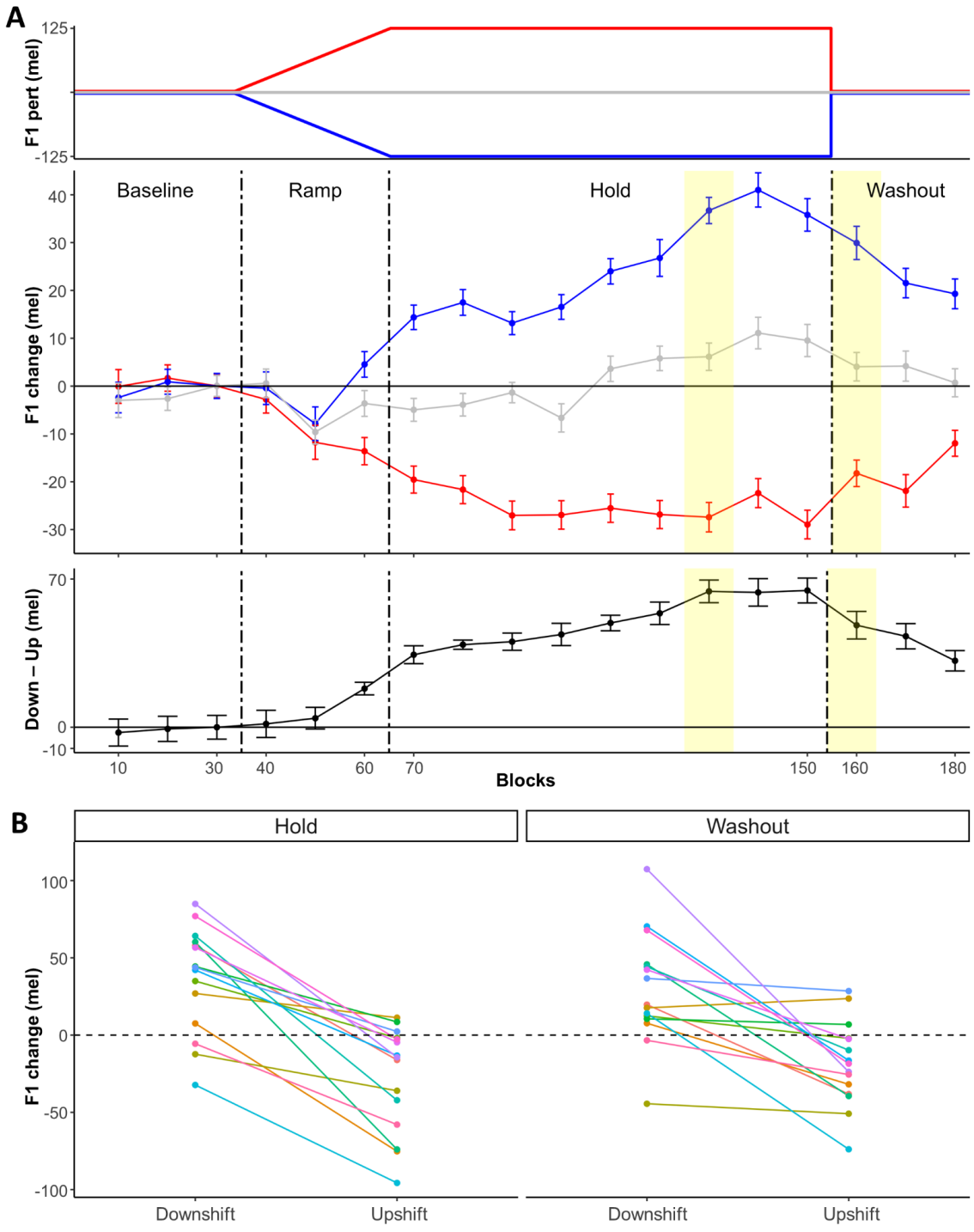

Figure 3. *Experiment 1 results. Individual data points show means and standard errors across 10 blocks.* ***A****: the F1 perturbation, F1 change, and differential adaptation (Downshift - Upshift). Colors show different perturbation conditions: red shows the Upshift, blue shows the Downshift, and gray shows the Unshifted. Shading indicates windows used in statistical analysis; blocks 61-70 were selected to achieve an identical amount of perturbation exposure in the Hold phase across experiments.* ***B****: individual speakers' F1 change in the analyzed Hold and Washout windows. Colors represent individual speakers.*

We also examined whether lexical items influenced the adaptation magnitude in Experiment 1. The selected model (R^2 = 0.089) was:

$$\Delta F1_{word} \sim \text{Word} + \text{Phase} + \text{Direction} + \text{Word:Direction} + \text{Word:Phase} + \text{Phase:Direction} + \text{Word:Phase:Direction} + (\text{Direction} + \text{Phase} + \text{Direction:Phase} \mid \text{Speaker})$$

The model detected no significant main effect involving Word (all $p > 0.17$), and post-hoc comparisons showed no significant difference between any pair of words (all $p > 0.30$). Numerically, words with higher lexical frequency seemed to be associated with larger adaptation: *head* (lexical frequency = 0.038%; adaptation = 44.37 ± 3.96 mels, $p = 0.001$), *bed* (lexical frequency = 0.013%; adaptation = 22.62 ± 2.95 mels, $p = 0.008$), and *ted* (lexical frequency = 0.0007%; adaptation = 21.09 ± 3.28 mels, $p = 0.17$).

In sum, Experiment 1 replicates the principal findings of Rochet-Capellan & Ostry (2011): speakers adapt separately to opposing perturbations in monosyllabic words containing the same vowel. Importantly, we found this behavior even when adjacent trials did not have identical perturbation/word; the ability to adapt to the opposing perturbations, in this case, cannot be attributed to sequential exposure to the same perturbation, a feature that could drive differential adaptation even in the absence of planning cues (Osu et al., 2004). We found no significant effect of lexical frequency on the adaptation.

Experiment 2

Experiment 2 tested whether participants could adapt to opposing perturbations applied to the same initial syllable (*"sev"*) in different disyllabic words (*seven* and *sever*). Speakers in Experiment 2 responded differently to the Upshift and Downshift perturbations. However, they tended to increase F1 regardless of the perturbation direction. Still, the F1 increase in the Downshift was greater than in the Upshift. Intriguingly, in Experiment 2, while speakers' F1 production in the Upshift condition initially increased in parallel with the Downshift condition, F1 switched to the predicted direction (from positive to negative) towards the end of Hold and during Washout (Figure 4A). Conversely, in Experiment 1, the F1 change diverged early in the Ramp phase and reached a plateau by the end of the Hold phase. Compared to the early separation of the Upshift and Downshift in Experiment 1, the differential adaptation in Experiment 2 took longer to emerge, potentially due to conflict between syllable-level learning (no adaptation) and word-level learning (differential adaptation). Speakers were largely consistent in their response to the auditory perturbations (Figure 4B): 95% produced a higher mean F1 in the Downshift condition compared to the Upshift condition in the Hold phase, and 85% in the Washout phase.

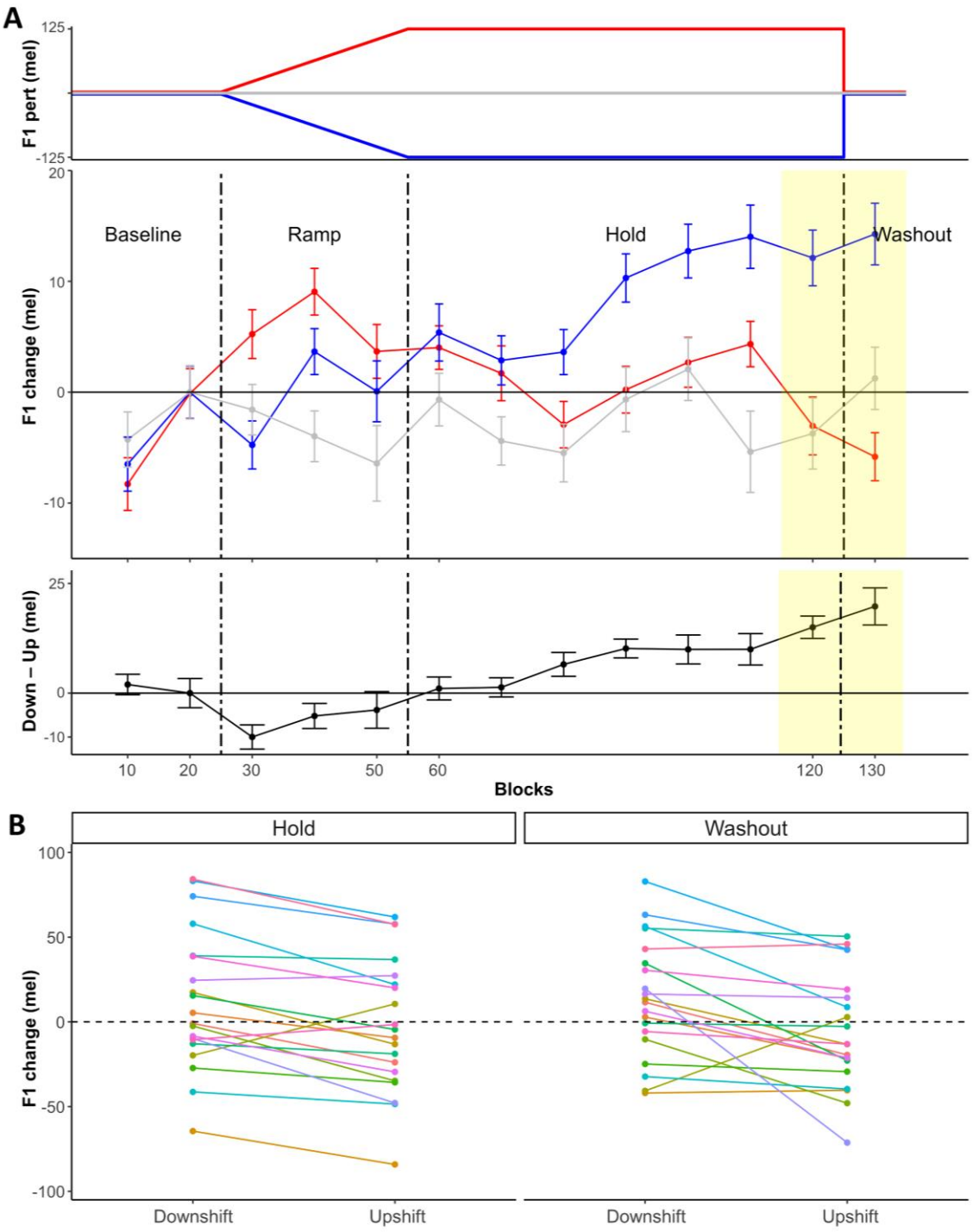

Figure 4. *Experiment 2 results. Individual data points show means and standard errors across 10 blocks.* ***A****: the F1 perturbation, F1 change, and differential adaptation. Colors show different shift conditions: red shows Upshift, blue shows Downshift, and gray shows Unshifted. Shading indicates windows used in statistical analysis.* ***B****: individual speakers' F1 change in the analyzed Hold and Washout windows. Colors represent individual speakers.*

The final selected model ($R^2 = 0.025$) was:

ΔF1 ~ Direction + Phase + Direction:Phase + (Direction + Phase | Speaker)

Speakers adapted according to the perturbation they received (a main effect of Direction ($\chi^2(2) = 16.65$, $R^2 = 0.033$, $p < 0.001$). The Upshift differed significantly from the Downshift ($d = 0.39$, $p < 0.001$), but the other two differences were not significant (Downshift vs. Unshifted: $d = 0.30$, $p = 0.20$; Upshift vs. Unshifted: $d = 0.07$, $p = 0.62$). When compared against zero, none of the three perturbation conditions differed significantly from zero (Upshift = -4.50 ± 1.90 mels, $p = 0.99$; Downshift = 13.08 ± 2.11 mels, $p = 0.34$; Unshifted = -1.38 ± 2.53 mels, $p = 0.99$).

There was no significant difference between the Hold and Washout phases (Direction by Phase interaction, $\chi^2(2) = 3.15$, $R^2 = 0.001$, $p = 0.21$). Numerically, for the Upshift and Downshift conditions, the adaptation magnitude increased from the Hold (Upshift: -3.04 ± 2.57 mels; Downshift: 12.02 ± 2.43 mels) to the Washout (Upshift: -5.95 ± 2.22 mels; Downshift: 14.14 ± 2.90 mels); little difference was present in the Unshifted condition (Hold: -3.78 ± 3.33 mels; Washout: 1.07 ± 3.07 mels). No perturbation condition in either phase differed significantly from zero (all $p > 0.40$). Nevertheless, post-hoc comparisons showed significant separations between the Upshift and the Downshift within the Hold ($d = 0.32$, $p = 0.019$) and the Washout ($d = 0.46$, $p < 0.001$). No other selected term in the model was significant ($p > 0.20$).

Again, we examined whether lexical items influence adaptation magnitude in Experiment 2. The selected model had the following formula ($R^2 = 0.017$):

$$\Delta F1_{word} \sim \text{Direction} + \text{Word} + \text{Direction:Word} + \text{Phase} + \text{Word:Phase} + \text{Direction:Phase} + \text{Direction:Word:Phase} + (1 + \text{Word} \mid \text{Speaker})$$

Word was not significant in the model ($\chi^2(1) = 0.07$, $R^2 = 0.002$, $d = 0.09$, $p = 0.79$). No other selected term in the selected model was significant ($p > 0.06$). Regarding frequency, *seven* has a higher frequency than *sever* (0.007% vs. 0.0001%). Numerically, the higher-frequency *seven* (10.99 ± 3.21 mels, $p = 0.43$) tended to have a larger adaptation than *sever* (6.86 ± 3.20 mels, $p = 0.43$). The direction of this non-significant trend of lexical frequency mirrors that of Experiment 1, where higher lexical frequency was associated with larger adaptation.

In sum, we tested differential adaptation to the identical initial syllable of disyllabic words in Experiment 2. Participants tended to increase their F1 regardless of the formant perturbation direction initially; the production in the Upshift lowered and separated from the Downshift by the end of the Hold phase. Overall, speakers do learn to adapt differentially to the two perturbations. We found no significant effect of lexical frequency on the adaptation.

Experiment 3

Experiment 3 tested whether participants could adapt to opposing perturbations applied to the same initial syllable (*"ped"*) in different trisyllabic words (*pedigree* and *pedicure*). Like Experiment 2, speakers in Experiment 3 tended to increase F1 initially regardless of the perturbation direction. Despite this global trend, the F1 increase in the Downshift was greater than in the Upshift, thus showing differential adaptation (Figure 5A). Most speakers mirrored the group average (Figure 5B): 95% produced a higher mean F1 in the Downshift condition compared to the Upshift condition in the Hold phase, and 90% in the Washout phase. The final selected statistical mode (R^2 = 0.042) was:

ΔF1 ~ Direction + Phase + Direction:Phase + (Direction + Phase | Speaker)

Speakers adapted according to the perturbation they received (a main effect of Direction ($\chi^2(2)$ = 13.36, R^2 = 0.05, p = 0.001). The Downshift differed significantly from the Upshift (d = 0.43, p = 0.001) and the Unshifted (d = 0.38, p = 0.004), but the difference between the Upshift and the Unshifted was not significant (d = 0.09, p = 0.20). When compared against zero, only the Downshift differed significantly from zero (Upshift = 0.98 ± 1.71 mels, p = 0.87; Downshift = 15.56 ± 1.44 mels, p = 0.34; Unshifted = 4.05 ± 1.47 mels, p = 0.66).

Like Experiment 2, Experiment 3 speakers' F1 production in the Upshift initially increased in parallel with the Downshift, and then switched to the predicted direction (from positive to negative) during the Washout (Direction by Phase interaction, $\chi^2(2)$ = 5.04, R^2 = 0.002, p = 0.081). However, there was no significant difference between the Hold and the Washout within each perturbation condition (all p > 0.059). Numerically, for the Upshift and Downshift, the adaptation magnitude barely changed from the Hold (Upshift: 3.81 ± 2.03 mels; Downshift: 15.72 ± 1.85 mels) to the Washout (Upshift: -1.87 ± 2.27 mels; Downshift: 15.40 ± 1.81 mels); the Unshifted dropped its adaptation magnitude from the Hold to the Washout (Hold: 7.76 ± 1.74 mels; Washout: 0.33 ± 1.92 mels). Only the Downshifted condition differed significantly from zero in both phases (both p < 0.035; all other p > 0.25). Still, post-hoc comparisons found significant separations between the Upshift and the Downshift within the Hold (d = 0.36, p = 0.028) and the Washout (d = 0.49, p < 0.001). The Downshift differed significantly from the Unshifted during the Washout (d = 0.49, p = 0.001), but not during the Hold (d = 0.27, p = 0.15). No other selected term in the model was significant (p > 0.07).

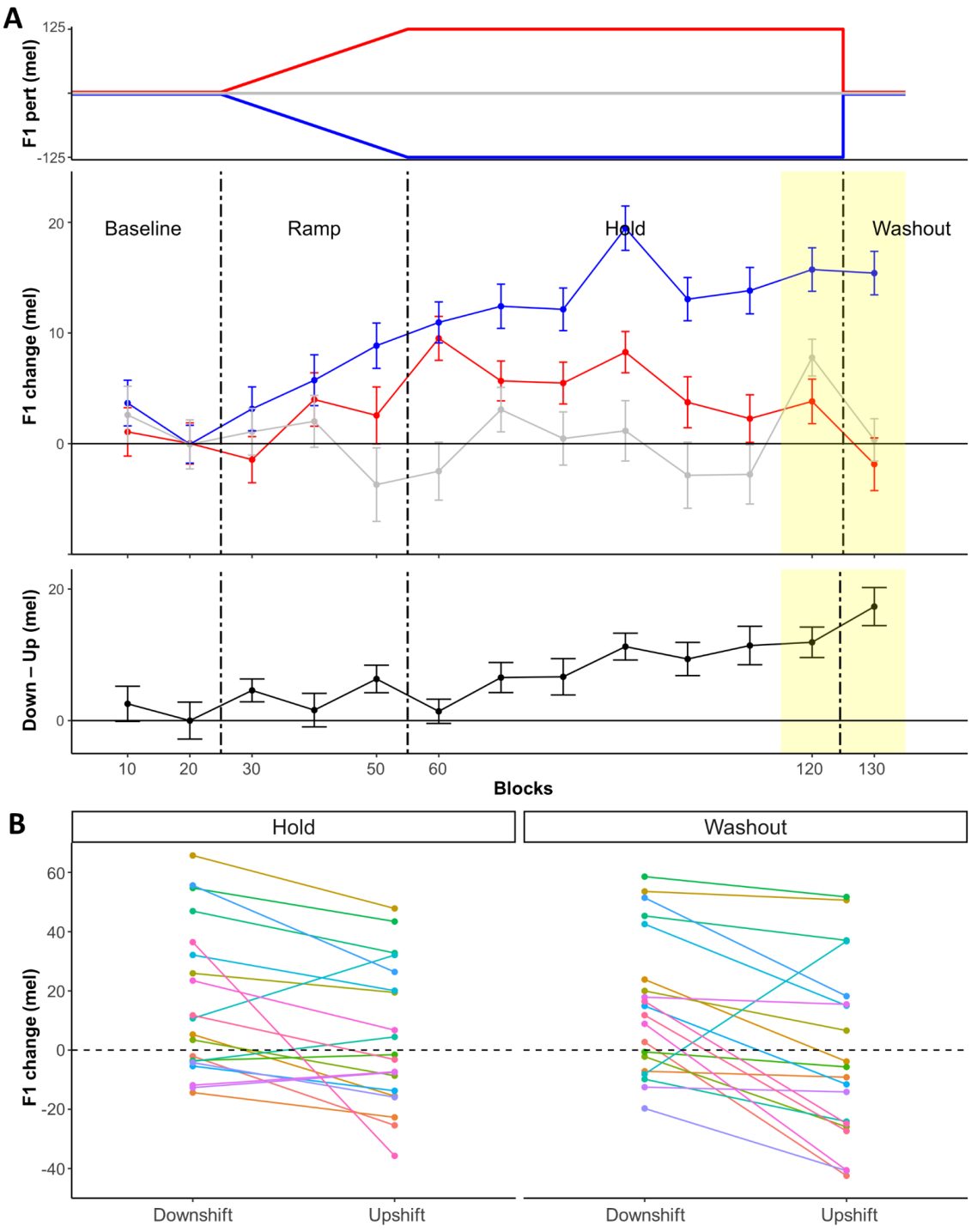

Figure 5. *Experiment 3 results. Individual data points show means and standard errors across 10 blocks.* ***A****: the F1 perturbation, F1 change, and differential adaptation. Colors show different shift conditions: red shows Upshift, blue shows Downshift, and gray shows Unshifted. Shading indicates windows used in statistical analysis.* ***B****: individual speakers' F1 change in the analyzed Hold and Washout windows. Colors represent individual speakers.*

Again, we examined whether lexical items influence adaptation magnitude in Experiment 3. The selected model (R^2 = 0.084) was:

$$\Delta F1_{word} \sim \text{Direction} + \text{Word} + \text{Phase} + \text{Word:Phase} + \text{Direction:Phase} + \text{Direction:Word} + (\text{Word} \mid \text{Speaker})$$

Word was not significant in the model ($\chi^2(1) = 1.12$, $R^2 = 0.024$, $d = 0.31$, $p = 0.29$). Although the interaction of Word and Phase was significant ($\chi^2(1) = 8.31$, $R^2 = 0.005$, $p = 0.004$), this was driven by *pedicure* having larger adaptation during Washout than during Hold ($d = 0.21$, $p = 0.003$), rather than by the difference between the words. Regarding lexical frequency, *pedigree* has a higher frequency than *pedicure* (0.0001% vs. 0.00002%). Numerically, the lower-frequency *pedicure* (12.70 ± 2.42 mels, $p = 0.069$) tended to have a larger adaptation than the higher-frequency *pedigree* (1.89 ± 2.29 mels, $p = 0.70$). In other words, in Experiment 3, higher lexical frequency is associated with smaller adaptation, contradicting the trends in Experiments 1 and 2. No other term involving Word in the selected model was significant ($p > 0.78$).

To summarize, we tested adaptation to opposing perturbations in trisyllabic words in Experiment 3. Participants tended to increase their F1 regardless of the formant perturbation direction initially; the production in the Upshift lowered and separated from the Downshift during the Washout. Overall, speakers do learn to adapt differentially to the two perturbations. We found no significant effect of lexical frequency on the adaptation.

Comparisons of differential adaptation across experiments

Regarding the size of differential adaptation, there is a stark distinction between the monosyllabic words in Experiment 1 and the multisyllabic words in Experiments 2 and 3 (Figure 6). In the monosyllables in Experiment 1, differential adaptation plateaued at approximately 60 mels towards the end of the Hold phase and decreased to about 45 mels after removing the perturbation during the Washout (Figure 3A). In contrast, in the multisyllabic words in Experiments 2 and 3, differential adaptation did not plateau during the Hold phase and even increased slightly during Washout; the maximal values were approximately 20 mels (Figures 4A and 4A). In brief, the multisyllabic words had a substantially reduced magnitude of differential adaptation compared to the monosyllabic words. Comparing the magnitude of differential adaptation across experiments yielded the following model ($R^2 = 0.418$):

$$\Delta F1_{diff} \sim \text{Experiment} + \text{Phase} + \text{Experiment:Phase} + (1 \mid \text{Speaker})$$

The magnitude of differential adaptation differed across experiments (a significant main effect of Experiment, $\chi^2(2) = 63.58$, $R^2 = 0.493$, $p < 0.001$). Experiment 1 (56.18 ± 6.77 mels) had a larger differential adaptation than Experiment 2 (17.39 ± 3.58 mels; $d = 1.31$, $p < 0.001$) and Experiment 3 (14.62 ± 3.15 mels; $d = 1.46$, $p < 0.001$), but the difference between Experiments 2 and 3 was not significant ($d = 0.13$, $p = 0.23$).

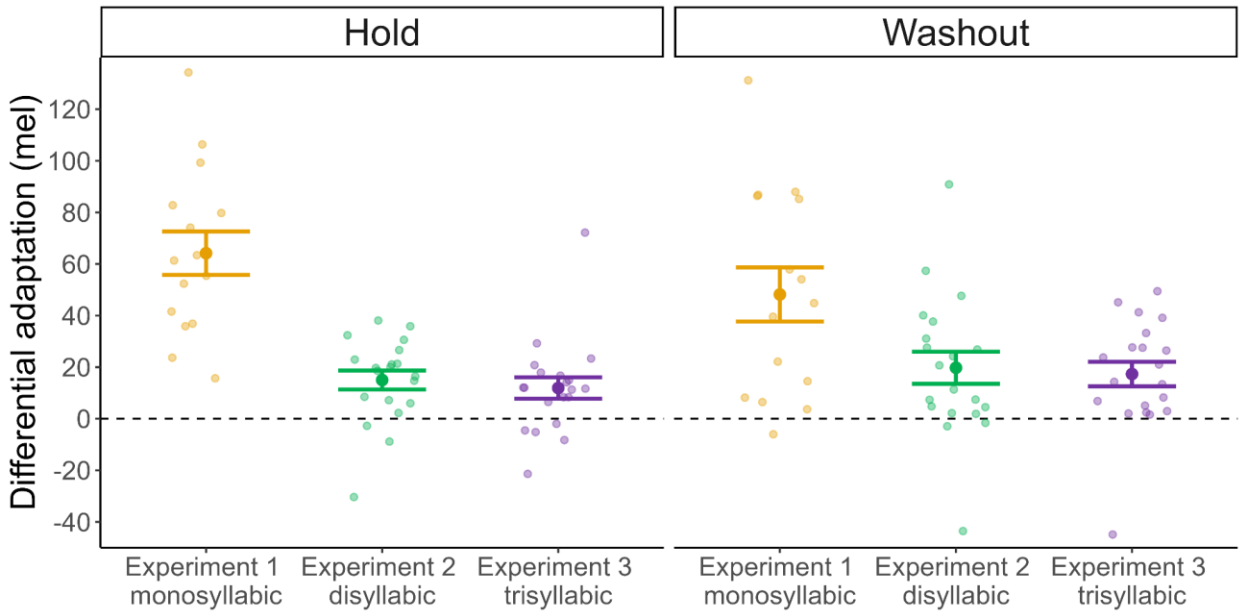

Figure 6. *Differential adaptation (Downshift - Upshift) across experiments in the analyzed Hold and Washout windows. The transparent dots show individual participant means. Error bars show standard errors.*

The interaction of Experiment and Phase was also significant ($\chi^2(2) = 9.86$, $R^2 = 0.038$, $p = 0.007$), which was primarily driven by Experiment 1's significant decrease in differential adaptation from Hold (64.18 ± 8.42 mels) to Washout (48.19 ± 10.48 mels; $d = 0.43$, $p = 0.016$). In contrast, differential adaptation increased slightly, though not significantly, from Hold to Washout in Experiment 2 (Hold: 15.01 ± 3.68 mels; Washout: 19.77 ± 6.22 mels; $d = 0.21$, $p = 0.55$) and Experiment 3 (Hold: 11.91 ± 4.15 mels; Washout: 17.32 ± 4.76 mels; $d = 0.27$, $p = 0.55$). Within each phase, differential adaptation in Experiment 1 was larger than in Experiments 2 and 3 (all $d > 0.99$, all $p < 0.001$), with no difference between Experiments 2 and 3 (both $d < 0.18$, both $p > 0.53$). No other selected term in the model was significant ($p = 0.52$).

Overall, these results show a larger differential adaptation in the monosyllabic words than in the multisyllabic words, with no difference between the multisyllabic words. However, it is possible that this difference is driven not by the mono- vs. multi-syllabic distinction, but by longer vowel durations in the monosyllabic words than in the multisyllabic words (Umeda, 1975). Longer vowel durations would result in longer exposure to the perturbation, which alone could explain the adaptation size difference (Schween et al., 2014). To check this, we first tested whether the vowel duration difference indeed existed in the analyzed window of F1 change across experiments, yielding the following model ($R^2 = 0.562$):

Duration ~ Experiment + Direction + Experiment:Direction + Phase + Experiment:Phase + (Phase + Direction + Phase:Direction | Speaker) + (Direction | Word)

As expected, vowel duration differed across experiments (a significant main effect of Experiment,

$\chi^2(2) = 452.47$, $R^2 = 0.569$, $p < 0.001$). Durations were the longest in Experiment 1 (233.96 ± 3.15 ms), intermediate in Experiment 2 (122.04 ± 1.06 ms), and the shortest in Experiment 3 (73.40 ms ± 0.64 ms). The difference between each pair of experiments was significant (all $d > 1.97$, all $p < 0.001$).

The interaction of Experiment and Phase was also significant ($\chi^2(2) = 6.46$, $R^2 < 0.001$, $p = 0.04$), primarily driven by the small duration decrease in Experiment 3 from Hold to Washout (Hold: 234.22 ± 4.49 ms; Washout: 233.71 ± 4.43 ms; $d = 0.007$, $p = 0.077$), yet the difference within Experiments 1 and 2 was not significant (both $d < 0.018$, both $p > 0.85$). The duration relationship between experiments described above still held within each phase (Experiment 1 > Experiment 2 > Experiment 3; all $d > 1.91$, all $p < 0.001$). No other selected term in the model was significant ($p > 0.66$).

Next, we tested whether the vowel duration difference could explain the magnitude of differential adaptation difference across experiments. First, we tested whether adding duration and/or its interaction to the selected statistical model improved overall fit (see the supplementary material for a list of all models during this forward-fitting procedure). None of the models tested with added duration term(s) improved overall fit compared to the original model (all $p > 0.42$).

Second, we ran a separate model selection procedure with duration, its interactions, and its random effects included in the maximal initial formula, resulting in the following model ($R^2 = 0.277$):

$$\Delta F1_{diff} \sim \text{Experiment} + \text{Duration} + \text{Phase} + \text{Experiment:Phase} + (\text{Duration} \mid \text{Speaker})$$

Comparing this model with the original differential adaptation model without duration yielded a significant difference ($p = 0.018$). However, this new model with added duration explained less variance than the original model (R^2 decreased from 0.418 to 0.277), suggesting that this model with duration was a worse fit to the data than the original. Although the estimated relationship between the duration and the magnitude of differential adaptation is in the predicted direction ($\beta = 0.071$; longer vowel duration was linked to larger adaptation size), duration was not a significant predictor in this new model ($\chi^2(1) = 0.40$, $R^2 = 0.01$, $p = 0.53$).

The magnitude of differential adaptation still differed across experiments in this new model (a significant main effect of Experiment, $\chi^2(2) = 6.00$, $R^2 = 0.104$, $p = 0.05$); this difference across experiments was also modulated by Phase (a significant interaction of Experiment and Phase, $\chi^2(2) = 11.80$, $R^2 = 0.031$, $p = 0.003$). Between experiments, Experiment 1 still had a significantly larger differential adaptation than Experiment 2 ($p = 0.047$), but the difference between Experiment 1 and Experiment 3 was not significant ($p = 0.11$). A similar pattern existed within the Hold phase (Experiment 1 vs. Experiment 2, $p = 0.001$; Experiment 1 vs. Experiment 3, $p = 0.052$). No other post-hoc comparisons between experiments were significant (all $p > 0.11$). No other selected term in the model was significant

($p > 0.43$).

In brief, although we found duration differences across the three experiments, we found no evidence of a significant relationship between duration and the magnitude of differential adaptation over and above the primary effect of monosyllabic vs. multisyllabic words.

Changes in speech parameters related to clear speech

In Experiments 2 and 3, although speakers adapted separately in the Downshift and Upshift conditions, there was a global tendency to increase F1. Similarly, in Experiment 1, the Unshifted condition also increased F1 despite receiving no perturbation. Here, we evaluate clear speech as a potential cause of this overall F1 increase. Figure 7 shows the percent changes in the acoustic parameters often associated with clear speech across experiments (duration, f_0, and intensity). Data for Experiment 1 were adjusted to make equal comparison across experiments possible (omitting the first 10 Baseline blocks, and the last 20 blocks from the Hold and Washout phases). The supplementary materials contain figures showing these changes as a function of perturbation direction; generally, different perturbation directions within a single experiment were parallel.

The duration had a steady decline in all three experiments. This is expected given the repeated production of a small set of stimulus words (e.g., Parrell & Niziolek 2021), potentially nullifying any increase that would typically be associated with clear speech. In contrast, the f_0 increased steadily in all three experiments, in line with a clear speech account. However, past reports have shown that f_0 tends to increase across extended repetitions of simple stimuli, like those used here (e.g., Jones & Munhall, 2000). Therefore, although the f_0 increase is consistent with a clear speech account, causes other than clear speech are also likely. The intensity changes had a greater range and more fluctuations, which may be attributed to fatigue and the inclusion of breaks in our experiment: speakers may drop intensity due to fatigue but may resume intensity after a break. Still, the observed general increase in intensity does not contradict a clear speech account.

In brief, of the three acoustic parameters often associated with clear speech, the f_0 and intensity changes do not contradict the predictions of clear speech. The duration change is the opposite of what clear speech predicts but is expected due to the repeated production of a limited set of words. Therefore, the evidence is mixed regarding whether the global F1 increase in Experiments 2 and 3, as well as the F1 increase in the unperturbed stimuli here and in Rochet-Capellan & Ostry (2011), is attributable to clear speech.

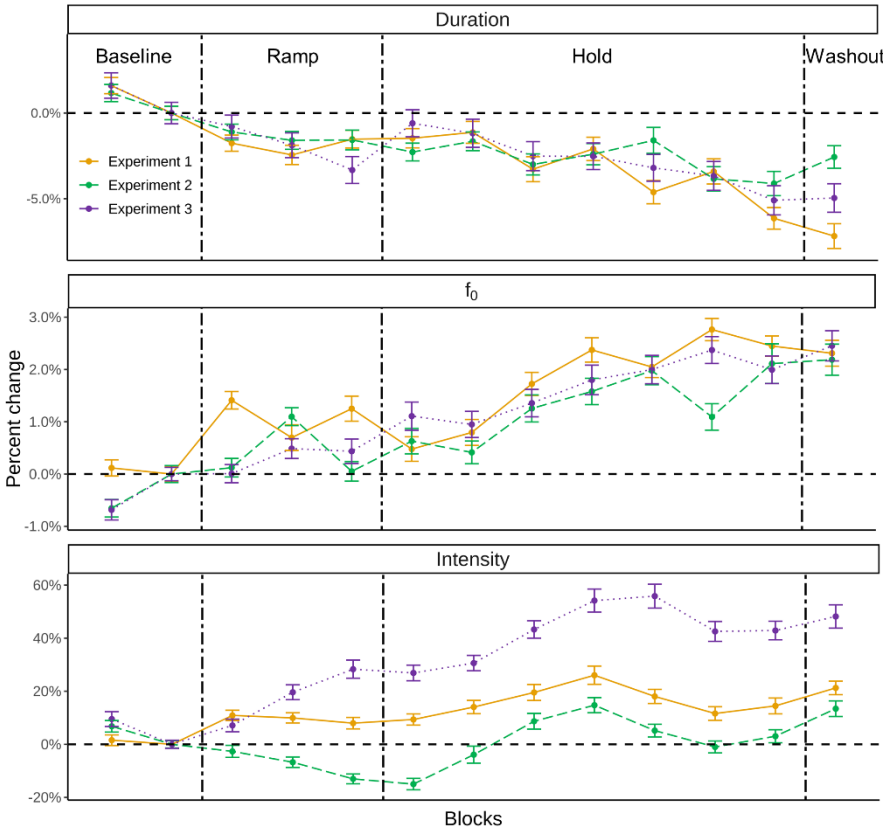

Figure 7. *Change in duration, fundamental frequency (f_0), and intensity in all experiments. Individual data points show means and standard errors across 10 blocks. Because Experiment 1 had more trials than Experiments 2 and 3, data for Experiment 1 excludes the first 10 Baseline blocks, the final 20 Hold blocks, and the final 20 Washout blocks.*

Perturbation awareness

After the main experiment, participants filled out a questionnaire, where they were first asked to recall and describe the task they did. They were then informed that there were two groups, one receiving true auditory feedback, and the other receiving manipulated auditory feedback. It was also highlighted that the manipulation was made to their voice. Participants were asked to choose which group they thought they were in. Most participants believed they received true auditory feedback (53.33% in Experiment 1, 75% in Experiment 2, and 70% in Experiment 3). Of the participants who reported they were in the manipulated group, some identified a change to their vowel characteristics (50% in Experiment 1, 25% in Experiment 2, and 50% in Experiment 3). There is strong evidence that speech adaptation is a largely implicit process and that even when participants are aware of the perturbation, they are unable to generate any strategies to oppose it (Kim & Max, 2021; Munhall et al., 2009). Therefore, participants developing conscious strategies is an unlikely cause for the differential adaptation in our study.

Discussion

In three experiments, we tested whether speakers could adapt differentially to opposing perturbations applied to words sharing potential planning units (Experiment 1: monosyllabic words sharing the same vowel; Experiments 2 and 3: multisyllabic words sharing the first syllable). All three experiments had a reliable separation between the two perturbation directions (all $p \leq 0.001$), showing differential adaptation. Importantly, adaptation in the multisyllabic words cannot be attributed purely to kinematic differences: carryover coarticulation was absent (identical initial syllables), and differences in anticipatory coarticulation were minimized (the same segment following the initial syllables). Indeed, baseline F1 did not differ between stimuli in any experiment (all $p > 0.09$). Therefore, the separation in multisyllabic words is possible only with supra-syllabic planning that encompasses the upcoming syllable(s), as these different planning contexts enable differential adaptation (Sheahan et al., 2016).

Notably, our finding shows word-level motor planning above and beyond what is shown by anticipatory coarticulation. Although anticipatory coarticulation must be attributed to speech motor planning at some level (Recasens, 2018), coarticulation itself does not provide strong evidence for supra-syllabic or word-level planning. First, the anticipatory spread of gestures is temporally limited to, at most, the vowel preceding the triggering segment, though potentially across multiple intervening consonants (e.g., Noiray et al. 2011; cf. the three-syllabic words in Experiment 3). Second, coarticulatory effects are due to either biomechanical inertia or effort optimization (Recasens, 2018). Third, coarticulation is not tied to specific words or syllable sequences, but occurs consistently whenever the triggering segments are found. Together, this suggests coarticulation reflects lower-level movement optimization, potentially related to motor programming (the motoric specification of movement plans in context; van der Merwe, 2021) instead of motor planning *per se*. Conversely, the differential adaptation in our experiments has no biomechanical motivation: it is unrelated to existing gestures in the following syllables, is far from the conditioning syllable (in Experiment 3), and is tightly tied to a specific segmental/syllabic context within a word.

Adaptation size in multisyllabic words (disyllables: $d = 0.39$; trisyllables: $d = 0.43$) was significantly smaller than in monosyllabic words ($d = 1.31$), a reduction that cannot be solely attributed to the shorter vowel durations in the multisyllabic words. This reduced adaptation is explainable if planning occurs at both the word and the syllable level. Each word forms an independent syllable in monosyllabic words, so there is no conflict between word- and syllable-level planning. In multisyllabic words, the perturbed syllable is identical; the conflict between syllable-level learning (opposing adaptive movements cancel out) and word-level learning (enabling differential adaptation) could explain the reduced adaptation size. This syllable-word conflict may also explain the later separation in multisyllabic words relative to

monosyllabic words.

Although word-level motor planning is included in some models of speech motor control, it is generally restricted to high-frequency words, forming pre-specified selection units akin to frequent syllables (Guenther, 2016). Because our multisyllabic stimuli are relatively low-frequency, our results challenge these models, suggesting rather that all words form independent planning units regardless of their lexical frequency. Nevertheless, it is possible that the repeated production of a limited set of words may have resulted in the *ad hoc* creation of new, word-level plans for the stimuli in our experiments. Such a novel-motor-plans account could also explain the late separation and smaller differential adaptation in the multisyllabic words, as these novel plans would take time to emerge. Crucially, adaptation in the multisyllabic words did not reach a plateau by the end of the Hold phase as in the monosyllabic words; it is therefore unclear whether the adaptation size would always be less in multisyllabic words (compared to monosyllables) or if this difference would disappear with further perturbation exposure. The novel-motor-plans account and the syllable-and-word planning account make different predictions in this regard: word-level motor plans, even if newly developed, should asymptote at the same level as monosyllabic words. Conversely, if planning occurs at both syllabic and word levels, this difference in adaptation size should persist even with a longer perturbation exposure, as there is fundamentally a conflict between the two levels. Future research with longer exposure in the Hold phase could resolve this question.

Our finding that word context influences syllabic motor planning is at odds with most current models of both psycholinguistic word production and speech motor control. In both types of models, the link between word planning and speech articulation is frequently modeled as a hand-off or feed-forward activation of syllabic motor plans (e.g., Guenther, 2016; Levelt, 1999). Our results indicate that this interface is more complicated, suggesting that a multisyllabic or word representation remains present through articulatory planning. In a network or spreading activation model, this representation may take the form of interactive co-activation of both syllable and supra-syllabic linguistic information, or a cascade of partially activated upcoming syllables; in either case, the word context can exert influence on the pattern of activation at the level of articulatory planning, resulting in distinct motor planning states for the same syllable in different words. The degree to which such co-activation spans levels of the production hierarchy is a matter of ongoing debate in competing models of word production, which range from fully discrete and feed-forward connections (Levelt, 1999; segregated syllable and word activations), through limited interactivity (Goldrick et al., 2006; permitting syllable and word co-activation), to highly interactive (Strijkers, 2016; parallel activations of syllable and word). The co-activation of words and syllables during motor planning is necessary to explain our results; it also implies similar co-activation during upstream psycholinguistic planning, supporting limited/high interactivity

models over fully discrete models.

Differential adaptation in Experiments 2 and 3 resulted in the same syllable being produced differently in different multisyllabic words, suggesting that at some level, the representation of words involves fine-grained phonetic details, beyond simple concatenations of syllables. However, the reduced adaptation we observed in multisyllabic words compared to monosyllables, likely reflecting a conflict between whole-word and syllable-level planning, implies a shared (and potentially abstract) syllable-level plan. Together, these findings support recent hybrid models of word representation that incorporate abstract and episodic representations (see Goldrick & Cole, 2023 for a review), rather than purely episodic accounts such as early versions of the Exemplar Theory (e.g., Pierrehumbert, 2001) or purely abstract cognitivist accounts (e.g., Dell, 1986).

In sum, we found differential adaptation to opposing perturbations in monosyllabic and multisyllabic words with shared segmental content, though adaptation was reduced in the multisyllabic words relative to the monosyllables, implying an interaction of syllable-level and word-level planning over and above what has been shown by analyses of coarticulation. These results have broad implications for our understanding of speech production. For the cognitive representations of speech, our results are in line with recent hybrid solutions that include abstract and episodic representations (e.g., Pierrehumbert, 2016). For psycholinguistic planning, our results support models permitting word-syllable co-activation (Goldrick et al., 2006; Strijkers, 2016). For speech motor control, our results indicate motor plans are not restricted to syllables but also include word-specific planning. Broadly, our results suggest a tight integration between word and motor planning, suggesting revisions are needed to current models of both psycholinguistic word production and speech motor control.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., & Krivitsky, P. N. (2023). *lme4: Linear mixed-effects models using "Eigen" and S4* (1.1-34) [R]. https://CRAN.R-project.org/package=lme4

Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer* (6.3.18) [Computer software]. http://www.praat.org/

Cai, S., Boucek, M., Ghosh, S., Guenther, F., & Perkell, J. (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.

Caudrelier, T., Schwartz, J.-L., Perrier, P., Gerber, S., & Rochet-Capellan, A. (2018). Transfer of

learning: What does it tell us about speech production units? *Journal of Speech, Language, and Hearing Research*, *61*(7), 1613–1625. https://doi.org/10.1044/2018_JSLHR-S-17-0130

Cholin, J., Dell, G. S., & Levelt, W. J. M. (2011). Planning and articulation in incremental word production: Syllable-frequency effects in English. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 109–122. https://doi.org/10.1037/a0021322

Correll, J., Mellinger, C., & Pedersen, E. J. (2022). Flexible approaches for estimating partial eta squared in mixed-effects models with crossed random factors. *Behavior Research Methods*, *54*(4), 1626–1642. https://doi.org/10.3758/s13428-021-01687-2

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Goldrick, M. A., Miozzo, M., & Ferreira, V. S. (2006). Limited interaction in speech production: Chronometric, speech error, and neuropsychological evidence. *Language and Cognitive Processes*, *21*(7–8), 817–855. https://doi.org/10.1080/01690960600824112

Goldrick, M., & Cole, J. (2023). Advancement of phonetics in the 21st century: Exemplar models of speech production. *Journal of Phonetics*, *99*, 101254. https://doi.org/10.1016/j.wocn.2023.101254

Guenther, F. H. (2016). *Neural Control of Speech*. MIT Press.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

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. https://doi.org/10.1152/jn.00773.2012

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. https://doi.org/10.1121/1.1288414

Kim, K. S., & Max, L. (2021). Speech auditory-motor adaptation to formant-shifted feedback lacks an explicit component: Reduced adaptation in adults who stutter reflects limitations in implicit sensorimotor learning. *European Journal of Neuroscience*, *53*(9), 3093–3108. https://doi.org/10.1111/ejn.15175

Kim, K. S., Wang, H., & Max, L. (2020). It's about time: Minimizing hardware and software latencies in speech research with real-time auditory feedback. *Journal of Speech, Language, and Hearing Research*, *63*(8), 2522–2534. https://doi.org/10.1044/2020_JSLHR-19-00419

Krause, J. C., & Braida, L. D. (2003). Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America*, *115*(1), 362–378. https://doi.org/10.1121/1.1635842

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B., & Jensen, S. P. (2020). *lmerTest: Tests in linear mixed effects models* (3.1-3) [R]. https://CRAN.R-project.org/package=lmerTest

Levelt, W. J. M. (1999). Models of word production. *Trends in Cognitive Sciences*, *3*(6), 223–232. https://doi.org/10.1016/S1364-6613(99)01319-4

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

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. https://doi.org/10.1121/1.3035829

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, *14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Niziolek, C. (2015). *wave_viewer: First release* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.13839

Noiray, A., Cathiard, M.-A., Ménard, L., & Abry, C. (2011). Test of the movement expansion model: Anticipatory vowel lip protrusion and constriction in French and English speakers. *The Journal of the Acoustical Society of America*, *129*(1), 340–349. https://doi.org/10.1121/1.3518452

Orwant, J., & Brockman, W. (2019). *Google Ngram Viewer* [Computer software]. https://books.google.com/ngrams/

Osu, R., Hirai, S., Yoshioka, T., & Kawato, M. (2004). Random presentation enables subjects to adapt to two opposing forces on the hand. *Nature Neuroscience*, *7*(2), Article 2. https://doi.org/10.1038/nn1184

Parrell, B., Lammert, A. C., Ciccarelli, G., & Quatieri, T. F. (2019). Current models of speech motor control: A control-theoretic overview of architectures and properties. *The Journal of the Acoustical Society of America*, *145*(3), 1456–1481. https://doi.org/10.1121/1.5092807

Parrell, B., & Niziolek, C. A. (2021). Increased speech contrast induced by sensorimotor adaptation to a nonuniform auditory perturbation. *Journal of Neurophysiology*, *125*(2), 638–647. https://doi.org/10.1152/jn.00466.2020

Perkell, J. S., & Matthies, M. L. (1992). Temporal measures of anticipatory labial coarticulation for the vowel /u/: Within- and cross-subject variability. *The Journal of the Acoustical Society of America*, *91*(5), 2911–2925. https://doi.org/10.1121/1.403778

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. L. Bybee & P. J. Hopper (Eds.), *Frequency and the Emergence of Linguistic Structure* (pp. 137–158). John Benjamins. http://www.jbe.platform.com/content/books/9789027298034-tsl.45.08pie

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America*, *118*(4), 2561–2569. https://doi.org/10.1121/1.2011150

Pouplier, M. (2007). Tongue kinematics during utterances elicited with the SLIP technique. *Language and Speech*, *50*(Pt 3), 311–341. https://doi.org/10.1177/00238309070500030201

Recasens, D. (2018). Coarticulation. In *Oxford Research Encyclopedia of Linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.416

Rochet-Capellan, A., & Ostry, D. J. (2011). Simultaneous acquisition of multiple auditory–motor transformations in speech. *Journal of Neuroscience*, *31*(7), 2657–2662. https://doi.org/10.1523/JNEUROSCI.6020-10.2011

Rosario-Martinez, H. D., Fox, J., & R Core Team. (2015). *phia: Post-hoc interaction analysis* (0.2-1) [Computer software]. https://cran.r-project.org/web/packages/phia/index.html

Schween, R., Taube, W., Gollhofer, A., & Leukel, C. (2014). Online and post-trial feedback differentially affect implicit adaptation to a visuomotor rotation. *Experimental Brain Research*, *232*(9), 3007–3013. https://doi.org/10.1007/s00221-014-3992-z

Shattuck-Hufnagel, S., & Klatt, D. H. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal*

Behavior, *18*(1), 41–55. https://doi.org/10.1016/S0022-5371(79)90554-1

Sheahan, H. R., Franklin, D. W., & Wolpert, D. M. (2016). Motor planning, not execution, separates motor memories. *Neuron*, *92*(4), 773–779. https://doi.org/10.1016/j.neuron.2016.10.017

Stevens, S. S., Volkmann, J., & Newman, E. B. (2005). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185–190. https://doi.org/10.1121/1.1915893

Strijkers, K. (2016). A neural assembly-based view on word production: The bilingual test case. *Language Learning*, *66*(S2), 92–131. https://doi.org/10.1111/lang.12191

Tourville, J., Cai, S., & Guenther, F. (2013). Exploring auditory-motor interactions in normal and disordered speech. *Proceedings of Meetings on Acoustics*, *19*. https://doi.org/10.1121/1.4806503

Umeda, N. (1975). Vowel duration in American English. *The Journal of the Acoustical Society of America*, *58*(2), 434–445. https://doi.org/10.1121/1.380688

van der Merwe, A. (2021). New perspectives on speech motor planning and programming in the context of the four- level model and its implications for understanding the pathophysiology underlying apraxia of speech and other motor speech disorders. *Aphasiology*, *35*(4), 397–423. https://doi.org/10.1080/02687038.2020.1765306

Voeten, C. C. (2023). *buildmer: Stepwise elimination and term reordering for mixed-effects regression* (2.9) [Computer software]. https://CRAN.R-project.org/package=buildmer