

Speakers monitor auditory feedback for temporal alignment and linguistically relevant duration

Robin Karlin and Benjamin Parrell

Citation: [The Journal of the Acoustical Society of America](#) **152**, 3142 (2022); doi: 10.1121/10.0015247

View online: <https://doi.org/10.1121/10.0015247>

View Table of Contents: <https://asa.scitation.org/toc/jas/152/6>



Published by the [Acoustical Society of America](#)

JASA
THE JOURNAL OF THE
ACOUSTICAL SOCIETY OF AMERICA

**Special Issue: Fish Bioacoustics:
Hearing and Sound Communication**

CALL FOR PAPERS

Speakers monitor auditory feedback for temporal alignment and linguistically relevant duration

Robin Karlin^{a)}  and Benjamin Parrell^{b)} 

Waisman Center, University of Wisconsin–Madison, Madison, WI 53705, USA

ABSTRACT:

Recent altered auditory feedback studies suggest that speakers adapt to external perturbations to the duration of syllable nuclei and codas, but there is mixed evidence for adaptation of onsets. This study investigates this asymmetry, testing three hypotheses: (1) onsets adapt only if the perturbation produces a categorical error; (2) previously observed increases in vowel duration stem from feedback delays, rather than adaptation to durational perturbations; (3) gestural coordination between onsets and nuclei prevents independent adaptation of each segment. Word-initial consonant targets received shortening perturbations to approximate a different phoneme (cross-category; VOT of /t/ > /d/; duration of /s/ > /z/) or lengthening perturbations to generate a long version of the same phoneme (within-category; /k/ > [k^{hh}]; /ʃ/ > [ʃː]). Speakers adapted the duration of both consonants in the cross-category condition; in the within-category condition, only /k/ showed adaptive shortening. Speakers also lengthened all delayed segments while perturbation was active, even when segment duration was not perturbed. Finally, durational changes in syllable onsets and nuclei were not correlated, indicating that speakers can adjust each segment independently. The data suggest that speakers mainly attend to deviations from the predicted timing of motor states but do adjust for durational errors when linguistically relevant. © 2022 Acoustical Society of America.

<https://doi.org/10.1121/10.0015247>

(Received 6 July 2022; revised 12 October 2022; accepted 3 November 2022; published online 2 December 2022)

[Editor: Susanne Fuchs]

Pages: 3142–3154

I. INTRODUCTION

A handful of recent studies indicate that speakers use auditory feedback about the duration of segments to correct for perceived temporal errors induced by experimentally altering segmental timing in the auditory feedback (Floegel *et al.*, 2020; Karlin *et al.*, 2021; Mitsuya *et al.*, 2014; Oschkinat and Hoole, 2020, 2022). However, it has been suggested that temporal adaptation (learning) is limited by syllable position. Specifically, both Oschkinat and Hoole (2020) and Karlin *et al.* (2021) reported no adaptation in syllable onsets, but did find adaptation in syllable nuclei (vowels). Oschkinat and Hoole (2020) additionally reported adaptation in syllable codas, indicating that the crucial distinction is syllable position rather than consonant status. These results contrast with the results reported by Mitsuya *et al.* (2014), who found changes in VOT production in syllable onsets in response to altered auditory feedback. We see three possible interpretations of these results.

The first hypothesis is that syllable onsets require a perturbation that crosses a categorical boundary in order to elicit a motor response, similar to effects of spectral change induced by altered auditory feedback reported by Niziolek and Guenther (2013), where speakers compensated more to perturbations of vowel formants that crossed a categorical boundary than to perturbations of the same magnitude that

stayed within category. The two target words used by Mitsuya *et al.* (2014) were “tipper” and “dipper,” where the contrast is largely carried by voice onset time (VOT) in the initial consonant; specifically, “tipper” has long-lag VOT, while “dipper” has short-lag VOT. In this study, speakers heard pre-recorded tokens of “tipper” when saying “dipper” and vice versa, and thus perceived themselves as saying a different word than intended. In contrast, neither Oschkinat and Hoole (2020) nor Karlin *et al.* (2021) used perturbations that crossed categorical boundaries: Oschkinat and Hoole (2020) lengthened the affricate /pf/ in the German word “Pfannkuchen”; Karlin *et al.* (2021) lengthened VOT on /k/ and the duration of friction noise in /s, z/ in the onset of American English words. Although duration is a secondary categorical cue that distinguishes /z/ and /s/ in American English, where /z/ is shorter than /s/ (Baum and Blumstein, 1987; Bjorndahl, 2018; Jongman, 1989), the phrasal context “a zipper” used in the experiment promotes voicing during the /z/ friction, which strongly encourages the /z/ percept even at a longer duration (Cole and Cooper, 1975; Francis *et al.*, 2008). As such, the durational perturbation did not effectively cross the categorical boundary. The need to cross a category boundary could arise from perceptual limitations. Previous perceptual studies have indicated that listeners are less sensitive to non-contrastive duration differences in syllable onsets than in either syllable nuclei or codas (Goedemans and van Heuven, 1995; Huggins, 1972). Perturbations were of equal magnitude in syllable onsets and nuclei in both Oschkinat and Hoole (2020) and

^{a)}Electronic mail: rkarlin@wisc.edu

^{b)}Also at: Department of Communication Sciences and Disorders, University of Wisconsin–Madison, Madison, WI 53705, USA.

Karlin *et al.* (2021)—i.e., if the syllable onset was perturbed by 40 ms, the vowel was also perturbed by 40 ms. Thus, the perceived error would not be as great for syllable onsets as for vowels, as the perturbation magnitude would be closer to the perceptual threshold. Crossing the categorical boundary increases perceptual sensitivity to duration differences, increasing the likelihood that speakers would perceive an error and adapt future utterances.

A second hypothesis is that speakers are actually largely reacting online to shifts in the end points of segments, rather than adapting to changes in duration. In all studies, lengthening a consonant target has been achieved by delaying the end of that consonant, rather than shifting the beginning of the consonant earlier. As such, the onset of the following vowel is delayed, creating a mismatch in temporal alignment between the acoustic feedback and the predicted state of the articulators. There is extensive evidence that speakers are sensitive to delayed auditory feedback (DAF). In these studies, auditory feedback from an entire utterance is shifted by anywhere from 20 to 800 ms. At lower delay magnitudes similar to those induced in temporal adaptation studies (e.g., 20–80 ms), typical speakers reduce speech rate, with the majority of this effect localized to vowel duration (Howell and Powell, 1987; Kalveram and Jäncke, 1989; Yates, 1963, *inter alia*), and without the repetition errors that commonly occur at larger delay magnitudes [e.g., 100–300 ms, Kalveram and Jäncke (1989)]. Given that DAF is typically applied globally, it is unclear whether the slower speech and lengthened vowels induced by DAF are driven by repeated, local delays of individual segments or if they reflect a more global response. As the delay of vowel onset in temporal adaptation papers has been squarely in the range that produces DAF-induced slowing, it is possible that the large lengthening effects documented in the delayed vowel by Mitsuya *et al.* (2014), Oschkinat and Hoole (2020), and Karlin *et al.* (2021) are due to this more general mechanism rather than being an adaptive response to the shortened vowel duration in the auditory feedback signal.

A third hypothesis for the observed asymmetry stems from the structure of motor plans for a syllable. In both Oschkinat and Hoole (2020) and Karlin *et al.* (2021), a lengthening perturbation of the syllable onset was also accompanied by an opposing, shortening perturbation of the syllable nucleus. Thus, in order to adapt both segments, speakers would have to shorten the syllable onset while also lengthening the syllable nucleus. Previous studies have shown that the movements for syllable onset consonants and vocalic nuclei are initiated at the same time (Browman and Goldstein, 1988, 1989), and it has been hypothesized that this reflects a tight in-phase coordinative relationship in the CV subsyllable. Under this view, the controller receives a tightly integrated CV unit; in order to alter the temporal dynamics of one part of this unit, the other is necessarily also affected. Thus, lengthening the vowel in response to a shortening perturbation would cause lengthening in the syllable onset as well, canceling out any attempt at shortening the syllable onset.

The present study explores these three alternative hypotheses. First, we examine the role of categorical boundaries by perturbing segments such that the end result is either across a categorical boundary (e.g., /t/ > /d/) or within the same category (e.g., /k/ with a longer VOT). If a categorical shift is necessary to trigger temporal adaptation in syllable onsets, then participants will only adapt their productions in the cross-category condition. Second, we test DAF as an explanation for extensive vowel lengthening by lengthening onset consonants without shortening the following vowel. If the speakers lengthen their vowels in the delay-only condition similarly to previous studies where the vowel was both shortened and delayed, this would suggest this lengthening is primarily driven by feedback delays rather than being an adaptive response to perceived vowel shortening. Finally, we test potential effects of the coordinative structures of syllables by examining changes in vowel duration when only the onset consonant is perturbed and the vowel is played back veridically. If speakers change their onset consonant and vowel productions in parallel even when the vowel is not perturbed, this would suggest that speakers control the duration of CV structures as one unit.

II. METHODS

A. Participants

Twenty-five participants (18 women and seven men) participated in the study, ranging in age from 19 to 42 years (mean = 26.3 years, median = 24 years). No participant reported any history of speech, hearing, or neurological disorders. In addition, all participants passed an automated Hughson-Westlake hearing screening (25 dB HL or lower 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.

B. Task

There were two different consonant target types: VOT in /k, t/, and the duration of frication in /f, s/. These two consonant target types were selected to examine potential effects of cue primacy: VOT is the primary cue for voicing category in /k, t/, but duration is a secondary cue for voicing category in /f, s/. In all, there were four stimulus words: *copper*, *tapper*, *shopper*, *sapper*. The words *copper* and *shopper* were used for the within-category condition, and the words *tapper* and *sapper* were used for the cross-category condition (temporal contrasts with the words “dapper” and “zapper,” respectively).

Participants completed the experiment in one session of four word blocks. The possible orders of word blocks were constrained such that they alternated between the cross-category and within-category condition, e.g., cross-within-cross-within but not within-cross-cross-within. The eight

possible orders of word blocks were counterbalanced, to the extent possible, across participants. Each word block consisted of four phases: a 30-trial baseline phase with veridical feedback, a 30-trial ramp phase where the perturbation was monotonically increased from 0 ms perturbation to maximum perturbation, a 60-trial hold phase where every trial received maximum perturbation, and a 30-trial washout phase with veridical feedback. On each trial, the participant produced the phrase “my [TARGET WORD].” At the end of the session, participants completed a survey to assess their awareness of the applied perturbation.

C. Temporal perturbation

1. Implementation

Temporal perturbation of specific segments was achieved using Audapter’s online status tracking (OST) capability, which uses root mean square intensity to detect changes in the acoustic signal, such as segment boundaries (Cai *et al.*, 2008). OST settings were individualized for each participant during a pretest phase that preceded each word block.

During the experiment, detected segment boundaries were used to trigger time-warping events in each perturbed trial. Time-warping events in Audapter are strictly causal and as such must start with an initial “time dilation” period, which lengthens a specified portion of audio by a factor that is less than 1 (indicating dilation). For example, a 20 ms portion of vowel may be warped by a factor of 0.25 to produce an 80 ms portion (generating a 60 ms lengthening perturbation). This portion is followed by an optional “stasis” period, where audio is played back at the original speed, but at a delay produced by the initial dilation. Finally, a “catch-up” period can produce shortening, where audio is played back at a specified rate greater than 1 (indicating acceleration) until the samples return to real time. Only the initiation of the time-warping event is triggered by segment detection; all subparts of the time-warping event are scheduled by predetermined durations.

For the cross-categorical condition, the consonant target (VOT for /t/; fricative duration for /s/) was shortened. In order to generate this shortening using Audapter, the vowel /aI/ in “my” was lengthened. To place the catch-up (shortening) phase of the time-warping event during the target segment, the duration of the stasis period was based on an estimate of the time interval between the lengthening event during /aI/ and the beginning of the target segment (estimated for each participant using OST values). Thus, for /t/ the stasis period was based on the duration of /aI/ + closure of /t/, and for /s/ the stasis period was based on the duration of /aI/ only. As speakers show both random variability in segment durations and can potentially show overall changes in segment durations as the experiment progresses, the stasis period was determined for each trial using a running average of the previous ten trials.

For the within-category condition, perturbation was much more straightforward. The consonant target (VOT for

/k/; fricative duration for /f/) was lengthened during the time dilation phase of the time-warping event. In order to maintain durations for the remainder of the utterance, the stasis period was set based on the duration of the remainder of the phrase + three standard deviations, again using a running average of the previous ten trials. Thus, the catch-up period occurred after the participant had finished the target utterance (see Fig. 1 for illustration of perturbations).

It may be noted here that we are exclusively working with the long member of the consonant targets (i.e., all voiceless, /t, k, s, f/), and that the shortening and lengthening perturbations are thus correlated with cross- vs within-category conditions, respectively. This is due to the technical difficulty in warping short targets: there is a very small amount of signal available to detect and warp in the shorter segments, particularly for VOT, where there may be only 10–20 ms of burst [see Karlin *et al.* (2021) for more detail on the difficulties in implementing this perturbation].

2. Participant awareness of perturbations

Nineteen of the 20 participants included in the analysis noted that they had noticed manipulation of the feedback. Of these 19, 15 specifically described temporal effects (e.g., “slower/drawn out”) or effects that were likely references to the actual perturbation (e.g., “the pronunciation of the beginning letters”). The remaining participant also described temporal effects after being informed that they had received manipulated feedback. As a weighty majority of the participants described a specific awareness of the perturbation, we do not conduct any analyses of effects of awareness.

3. Achieved perturbations

For the within-category perturbation, the target perturbation was 60 ms; consonant targets were lengthened by 60 ms and every segment thereafter was delayed by the same amount.

For the cross-category condition, perturbation magnitude was based on the distance between categories for each participant. For *sapper*, participants also produced “my zipper” during a pretest phase to provide a baseline for the duration of /z/. If the difference between their /s/ and /z/ categories was not at least 60 ms, the perturbation was set to 60 ms to match perturbation magnitude in the within-category condition. The median target perturbation for *sapper* was 60 ms, with range from 60 ms to 80 ms. For *tapper*, the magnitude of the perturbation was based on the participant’s mean VOT in *tapper* during the pretest phase, relative to a /d/-like VOT of 10 ms. The minimum of 60 ms was not applied in this case, as many speakers did not have sufficiently long VOT to support such a large perturbation. The median target perturbation for *tapper* was 64.5 ms, with range 41 to 98 ms.

Perturbation was overall successful, with the overwhelming majority of trials achieving at least 75% of the intended perturbation (*shopper*: 99% of trials; *copper*:

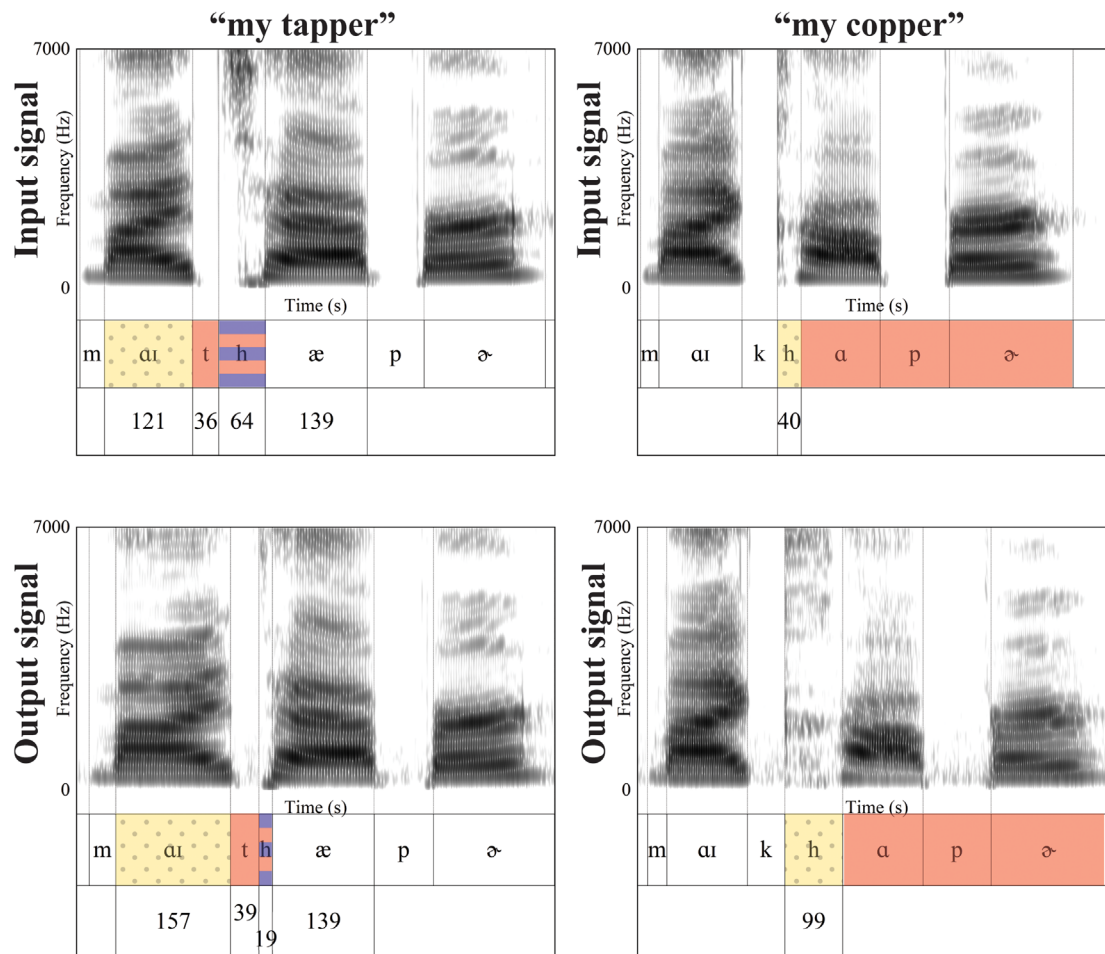


FIG. 1. (Color online) examples of cross-category (shortening) perturbation (“my taper”) and within-category (lengthening) perturbation (“my copper”); durations of perturbed segments are indicated in ms. Top panels are what the speaker said; bottom panels are what was played back over headphones. Yellow (with dots) indicates segments that were lengthened; red (solid) indicates segments that were delayed; blue/red striped indicate segments that were both delayed and shortened.

91.5%; *sapper*: 87.9%). The exception was *tapper*, which had two participants excluded for insufficient perturbation (see exclusions below). After these exclusions, 50% of trials achieved at least 75% of the intended perturbation; however, 85.6% of trials achieved at least 50% of intended perturbation. As we are principally interested in motor learning for the consonant targets, we did not eliminate individual trials for insufficient perturbation. Instead, we include additional analyses for the behavior of *tapper*, particularly the vowel, which was frequently unintentionally delayed. Means and standard deviations for achieved duration perturbations for

each segment are presented in Table I below. Means and standard deviations for feedback delays incurred for each segment are presented in Table II.

D. Exclusions

Five participants were excluded prior to data analysis: Two participants were excluded due to technical difficulties in the cross-category condition during data collection; one participant was excluded due to technical difficulties during the washout phase of *shopper*. Another participant had highly

TABLE I. Duration perturbations by segment (M = mean, SD = standard deviation). Negative values indicate shortening; positive values indicate lengthening. Lengthened segments are indicated in bold; shortened segments are indicated with italics. Very small (absolute value < 1 ms) values are due in large part to inherent imprecision in hand-correction. All values in ms.

		/m/		/ai/		Stop closure		Consonant target		Stressed vowel		/p/		/æ/	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
W-C	shopper	-0.3	2.1	<0.1	2.1	—	—	58.0	2.9	1.7	2.8	-1.3	3.1	-1.2	10.5
	copper	-0.2	2.0	<0.1	1.9	1.5	8.5	54.4	16.3	3.0	11.9	<0.1	5.0	-1.5	10.2
X-C	sapper	1.2	8.5	56.5	14.1	—	—	-57.8	10.2	<0.1	3.2	<0.1	2.6	-0.9	3.7
	tapper	4.3	16.9	55.7	24.0	-13.1	17.4	-41.7	19.3	-5.4	11.6	<0.1	1.5	-0.3	2.9

TABLE II. Feedback delays of the onset of each by segment, adjusted by subtracting the mean hardware lag (13.3 ms) (M = Mean, SD = Standard deviation). Positive values indicate delay in feedback. Bold cells indicate intentional delay. Very small (absolute value < 1 ms) values are due in large part to inherent imprecision in hand-correction. All values in ms.

		/m/		/ai/		Stop closure		Consonant target		Stressed vowel		/p/		/ə/	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
W-C	shopper	-0.2	1.2	<0.1	2.0	—	—	<0.1	1.4	57.9	2.5	59.5	1.6	58.2	2.7
	copper	<0.1	1.2	<0.1	1.6	<0.1	1.3	12.1	8.4	55.6	14.3	58.8	7.4	58.6	6.7
X-C	sapper	-0.1	2.6	1.0	8.8	—	—	67.1	10.0	<0.1	2.2	-0.6	2.2	-0.5	2.0
	tapper	<0.1	<0.1	4.3	16.9	60.0	19.9	46.9	22.1	5.2	11.6	<0.1	0.7	-0.3	1.3

inconsistent pause structure and speech rate in the *copper* and *tapper* conditions, such that the stop closure portion could not be reliably segmented. One additional participant was excluded due to visible intoxication during the study.

One participant was excluded from the cross-category condition only after data processing due excessive pausing between words during washout (*tapper*), and from *sapper* as they were a clear outlier (>1.5 IQRs below the first quartile). The *tapper* data from two additional participants were excluded from analysis due to insufficient perturbation (-6 ms and -21 ms, both ≥ 1.5 IQRs below the first quartile). As we focus here on the effects of consistent perturbation on motor learning, we did not exclude individual trials due to insufficient perturbation.

E. Data labeling

The audio from participants' productions was hand-segmented to obtain the duration of each segment in the target utterance. Raters followed a segmentation guideline and were trained by the first author; the first author also performed spot checks to ensure accuracy throughout the experiment.

F. Analysis

Linear mixed effects models were run using the lme4 package in R (Bates *et al.*, 2014). Models included random intercepts for participant. Random slopes were not included due to singularity of fit. Models were built incrementally and compared using the analysis of variance function (Kuznetsova *et al.*, 2015). *Post hoc* comparisons were corrected using the Tukey method and were run using the EMMEANS package (Lenth, 2019). For the analysis of changes in segment duration, potential fixed effects were phase (hold, baseline, washout), consonant type (VOT vs fricative duration), and perturbation condition (cross-category vs within-category, or the accompanying effects in segments other than the consonant target). Supplementary materials, data, and associated code are available at <https://osf.io/fq785/>.

The last ten trials of the baseline phase serve as a baseline of comparison for changes across phases of the experiment. The last ten trials of hold serve as a measure of the combined effects of adaptation and online reactions to auditory feedback. The first ten trials of washout serve as the measure of adaptation alone, as there is no perturbation for

speakers to react to. Reported estimated means are the change in production compared to baseline.

For analyses of consonant targets (i.e., the initial consonants in the target words), changes in production are normed to the mean perturbation received in the last ten trials of hold, to allow direct comparison both between *tapper* and the other words, as well as different participants with different magnitudes of perturbation in *tapper* and *sapper*. Estimated means are thus reported as change from baseline in percent of perturbation. Furthermore, models for the consonant targets use sign-flipped data such that shortening (for *copper*, *shopper*) and lengthening (*tapper*, *sapper*) can be directly compared. Changes in production in *copper* and *shopper* trials are multiplied by -1 so that positive values indicate changes in opposition to the perturbation.

For analyses of all other segments, models are conducted based on raw change in milliseconds, and estimated means are reported as difference from baseline in milliseconds. For these analyses, data are not sign-flipped.

III. RESULTS

A. Hypothesis 1: Effects of categorical boundaries on onset consonant adaptation

Changes in consonant production are illustrated in Fig. 2. Overall, participants consistently adapted their productions of syllable onsets in the cross-category condition, but only adapted the production of /k/ in the within-category condition.

The addition of phase significantly improves the fit of the model compared to the null model [$\chi^2(2) = 232.42$, $p < 0.0001$]; overall, participants changed the production of syllable onset consonants in opposition to the perturbation. Both the hold phase and washout are significantly different from baseline (hold: $23.7 \pm 1.9\%$; washout: $11.5 \pm 1.9\%$, both $p < 0.0001$); there is significantly greater opposing change during the hold phase than during washout ($p < 0.0001$).

The addition of perturbation condition also significantly improves model fit [$\chi^2(1) = 148.32$, $p < 0.0001$], as does the interaction between category condition and phase [$\chi^2(2) = 167.52$, $p < 0.0001$]. Both conditions were significantly different during both hold and washout, though at different magnitudes: the cross-category words were significantly different during hold and showed large changes ($42.7 \pm 2.1\%$, $p < 0.0001$), with reduced but still significant differences during washout ($16.5 \pm 2.1\%$, $p < 0.0001$); the within-category words

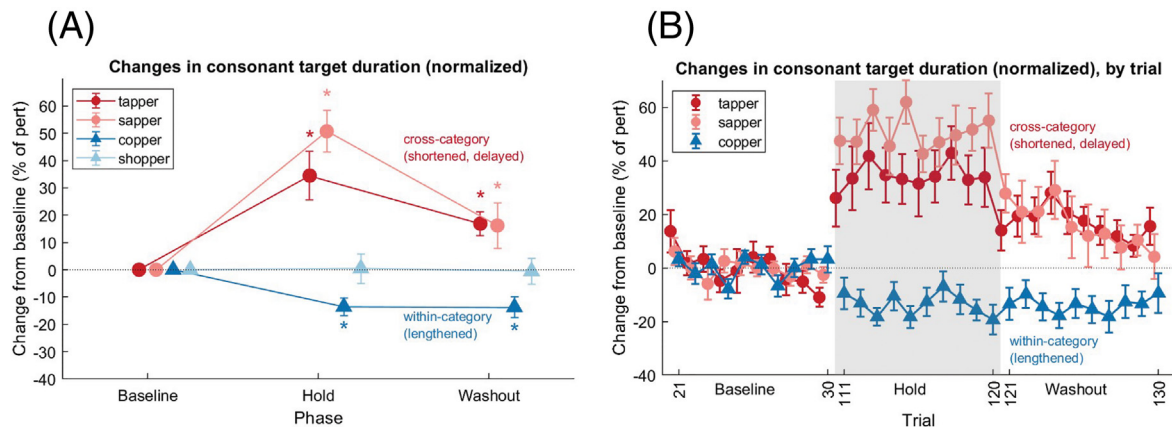


FIG. 2. (Color online) Changes in consonant production, normalized as percentage of perturbation (data not sign-flipped). (A) Production by phase (ramp excluded); asterisks indicate significant difference from baseline. (B) Production by trial within each phase (only trials used in the analysis) for words with significant differences from baseline.

were also significantly different from baseline in hold ($6.5 \pm 2.1\%$, $p = 0.01$) and washout ($7.1 \pm 2.2\%$, $p = 0.005$) but with more similar magnitudes of difference between the two phases.

Adding consonant type to the model does not significantly improve model fit [$\chi^2(1) = 3.64$, $p = 0.06$]. However, the interaction between consonant type and phase does significantly improve model fit [$\chi^2(2) = 9.07$, $p = 0.01$]. Both consonant types are significantly different from baseline during hold (VOT: $24.4 \pm 2.1\%$; fricative duration: $24.8 \pm 2.1\%$; both $p < 0.0001$), and there is not a significant difference between consonant types ($p = 0.99$). Both consonant types are also significantly different from baseline during washout (VOT: $15.5 \pm 2.1\%$; fricative duration: $8.3 \pm 2.1\%$; both $p < 0.001$), where VOT consonant targets show more adaptation ($p = 0.005$).

The addition of the interaction between consonant type and perturbation condition also significantly improves model fit [$\chi^2(1) = 39.02$, $p < 0.0001$], suggesting that the individual words behave differently from each other. The three-way interaction between phase, perturbation condition, and consonant type significantly improves model fit [$\chi^2(2) = 29.29$, $p < 0.0001$]. During hold, *sapper* (fricative duration, cross-category: $50.7 \pm 2.5\%$, $p < 0.0001$; raw change 30.6 ± 1.4 ms) shows the most change, followed by *tapper* (VOT, cross-category: $34.3 \pm 2.6\%$, $p < 0.0001$; raw change 16.8 ± 1.4 ms), and then *copper* (VOT, within-category: $13.6 \pm 2.6\%$, $p = 0.0003$; raw change -7.7 ± 1.4 ms); *shopper* (fricative duration, within-category) is the only word that does not significantly change from baseline ($-0.7 \pm 2.5\%$, $p = 1.00$; raw change 0.4 ± 1.4 ms). The magnitude of change is significantly different between each word (all $p < 0.001$).

Similarly, all words show significant after-effects in washout except for *shopper* ($0.55 \pm 2.5\%$, $p = 1.00$; raw change -0.3 ± 1.4 ms). For the two cross-category words, the magnitude of change is significantly smaller during washout than during hold (*tapper*: $16.8 \pm 2.5\%$, raw change 7.7 ± 1.4 ms; *sapper*: $16.2 \pm 2.6\%$, raw change 11.7 ± 1.4 ms; all $p < 0.0001$ compared to both hold and baseline). This aligns

with the hypothesis that delaying the onset of the consonant in the two cross-category words would trigger lengthening due to DAF effects, which would disappear when perturbation is removed, leaving only the effects of learning from the duration perturbation during washout. This effect is illustrated in Fig. 2(B), where there is a large drop in duration across the phase boundary (trial 120 is perturbed; trial 121 is not). This contrasts with a more gradual decline in duration as the washout phase continues and speakers de-adapt.

For the within-category word *copper*, the magnitude of the learning in the washout phase is different from baseline, but not significantly different from hold ($13.8 \pm 2.5\%$, raw change -7.7 ± 1.4 ms; $p < 0.0001$ relative to baseline, $p = 1.00$ relative to hold). Since adaptation in this case is a shortening response, it is impossible that it should be triggered by DAF—which in any case did not occur for the onset of this consonant. Thus, the shortening response during hold is reflective of learning, which carries over to the washout phase.

1. Effects of perturbation on closure duration in *tapper* and *copper*

Changes in the closure duration and total duration of /t, k/ are illustrated in Fig. 3. For this analysis, we are only considering *copper* and *tapper*, as there is no distinct closure phase for the fricative targets. The addition of phase significantly improves model fit compared to the null model [$\chi^2(2) = 43.91$, $p < 0.0001$]. Consonant closure is longer than baseline during hold (6.7 ± 2.1 ms, $p < 0.001$), but not during washout (1.7 ± 2.1 ms, $p = 0.10$). Adding perturbation condition (no perturbation for *copper*; delay for *tapper*) does not significantly improve model fit [$\chi^2(2) = 0.42$, $p = 0.52$], nor does the interaction between phase and perturbation condition [$\chi^2(2) = 2.54$, $p = 0.28$]. Closure duration increases during hold for both consonants (*tapper* 8.0 ± 2.3 ms; *copper* 5.6 ± 2.3 ms, both $p < 0.01$ compared to baseline), and returns to baseline in washout (*tapper* 1.9 ± 2.3 ms; *copper* 1.5 ± 2.3 ms, both $p > 0.40$).

For consonant duration overall (closure plus aspiration), the addition of phase significantly improves model fit

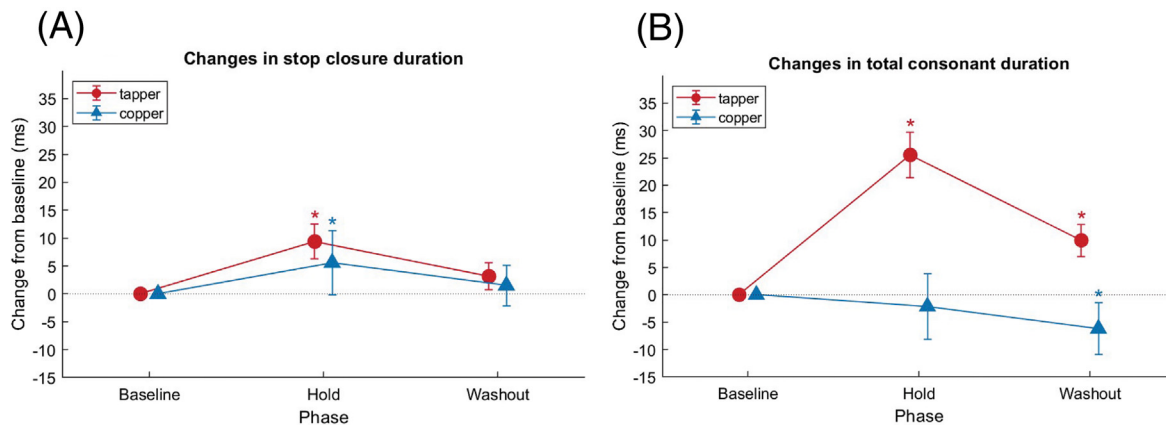


FIG. 3. (Color online) Changes in the closure (A) and total consonant duration (B) of /t, k/. Asterisks indicate significant differences from baseline.

$[\chi^2(2) = 70.20, p < 0.0001]$, as does the addition of perturbation condition $[\chi^2(2) = 130.46, p < 0.0001]$ and the interaction between phase and perturbation condition $[\chi^2(2) = 117.5, p < 0.0001]$. As suggested by the independent descriptions of closure duration and aspiration, the overall duration of /t/ in *tapper* was significantly longer during both hold (24.0 ± 2.9 ms, $p < 0.0001$) and washout (8.8 ± 2.9 ms, $p < 0.0001$). This is consistent with a DAF effect for both closure and aspiration in /t/ during hold, with adaptive lengthening remaining only in the aspiration during washout. In contrast, the overall duration of /k/ in *copper* was not significantly different during hold (-2.1 ± 2.9 ms, $p = 0.84$), but was significantly shorter during washout (-6.2 ± 2.9 ms, $p = 0.006$). This is indicative of adaptive shortening in the aspiration portion only, with potentially different strategies for achieving that shortening in hold and washout.

B. Hypothesis 2: Effects of delay on following vowel

The remainder of the analyses will be reported as raw change in milliseconds. Changes in vowel production are illustrated in Fig. 4. The addition of phase significantly improves the model fit compared to the null model

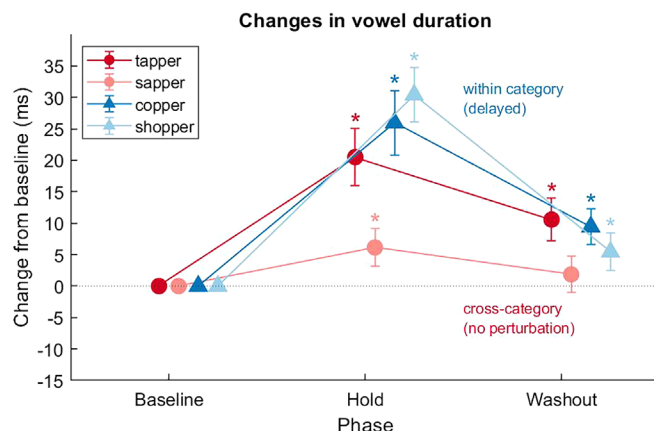


FIG. 4. (Color online) Changes in vowel production over the course of the experiment, reported in ms. Asterisks indicate significant differences from baseline.

$[\chi^2(2) = 582.67, p < 0.0001]$; participants produced longer vowels during the hold phase (20.7 ± 1.7 ms, $p < 0.0001$) and the washout phase (6.5 ± 1.7 ms, $p < 0.0001$). The addition of perturbation condition (delay vs no change) also significantly improves model fit $[\chi^2(1) = 81.39, p < 0.0001]$, as does the interaction between perturbation condition and phase $[\chi^2(2) = 104.38, p < 0.0001]$. Speakers increased the duration of their vowels more during the hold phase when the vowel was delayed (28.19 ± 1.8 ms) than when the vowel was unaltered (12.6 ± 1.8 ms, $p < 0.0001$).

Adding consonant type to the model also significantly improves model fit $[\chi^2(1) = 32.07, p < 0.0001]$, as does the interaction between phase and consonant type $[\chi^2(2) = 16.96, p = 0.0002]$. Vowels lengthen overall for both types of consonants; however, there are greater changes in production in VOT targets (22.8 ± 1.8 ms) than in fricative duration targets (18.0 ± 1.8 ms, $p = 0.0003$). There are also aftereffects in both cases, again larger in VOT target (VOT: 9.6 ± 1.8 ms; fricative: 3.3 ± 1.8 ms).

The interaction between consonant type and perturbation condition significantly improves model fit $[\chi^2(1) = 36.87, p < 0.0001]$, as does the three-way interaction between phase, perturbation condition, and consonant type $[\chi^2(2) = 39.88, p < 0.0001]$. During hold, participants lengthened vowels for all words, showing the most lengthening in the two words where the onset of the vowel was delayed (*shopper*: 30.4 ± 2.0 ms; *copper*: 26.0 ± 2.0 ms; both $p < 0.0001$ relative to baseline), followed by *tapper* (19.8 ± 2.0 ms, $p < 0.0001$), and with *sapper* showing the least amount of lengthening (5.4 ± 2.0 ms, $p = 0.006$). Vowel durations remained somewhat elevated during washout for all words but *sapper*, but showed significant decreases between hold and washout. The two VOT target words showed the greatest after-effects (*copper*: 9.5 ± 2.0 ms; *tapper*: 9.7 ± 2.0 ms), followed by *shopper* (delayed vowel during hold; 5.4 ± 2.0 ms), with *sapper* returning to baseline (no perturbation on vowel; 1.1 ± 2.0 ms).

Overall, these results suggest that speakers are reacting online to the delay in the onset of the vowel, which largely goes away after this perturbation is removed. However, even the vowels in the unperturbed condition showed some lengthening, warranting further investigation. In particular,

the vowel in *tapper* showed a larger increase in duration than *sapper*. One possible source of the additional lengthening in *tapper* is that speakers were affected by unintentional perturbations to the vowel. Recall that the perturbation in the cross-category condition required estimating the time between the onset of the vowel and the onset of the consonant target. Since the VOT target in *tapper* is relatively short, the window of estimation was frequently imperfect, leading to inadvertent delay and shortening during the vowel target as well (see Tables I and II for more detail on the perturbations received by the segments surrounding the consonant target, and Fig. 5 for an illustration of the wide dispersal of delays in *tapper* trials). Thus, it is possible that speakers that received more delay also lengthened their vowels more. First, we test the trial-to-trial relationship between delayed feedback and lengthening during the hold phase in *tapper*. Adding delay magnitude to the model significantly improves model fit [$\chi^2(1) = 27.03$, $p < 0.0001$]; trials with more delay increase vowel duration more ($\beta = 0.51$ ms, $SE = 0.09$ ms). This relationship strongly indicates that speakers react to delay online, and are not implementing some global strategy when they are in some delayed auditory feedback mode.

To investigate whether speakers that experienced more shortening and delay of the vowel also showed greater lengthening during washout, we can also examine the by-participant relationship between mean perturbation magnitude during hold (here, the magnitude of shortening) and the magnitude of aftereffect (difference from baseline during first 10 trials of washout). This model is a simple linear regression, as each participant has one datapoint. There is not a significant relationship between perturbation magnitude and aftereffect magnitude [$R^2 = 0.0002$, $F(1, 17) = 0.004$, $p = 0.95$]. Thus, the aftereffects are not the result of learning in response to (unintentional) shortening of the vowel. Another possibility is that speakers retain

some effect of delay, similar to what was reported for the within-category condition above. However, there is not a significant relationship between mean delay magnitude and aftereffect either [$R^2 = 0.001$, $F(1, 17) = 0.02$, $p = 0.89$].

C. Hypothesis 3: Coordination between syllable onsets and nuclei

Finally, it is possible that the overall lengthening in both *tapper* and *sapper* is related to lengthening of the syllable onset. This could stem from a tight coordinative relationship between onset consonants and syllable nuclei, as hypothesized by Karlin *et al.* (2021) and Oschkinat and Hoole (2020). Such a relationship would predict that speakers that lengthen their consonants more will also lengthen the following vowel more. In the following analysis, we test if changes in stressed vowel duration during hold can be predicted by changes in the produced duration of the perturbed onset consonant.

Starting from a base model that includes both consonant type (as *tapper* and *sapper* have already been shown to have different magnitudes of change from baseline in the vowel), and vowel delay (as delay will independently cause lengthening of the vowel), adding change in consonant production to the model does not significantly improve model fit [$\chi^2(1) = 0.34$, $p = 0.56$]. There is also no improvement when considering change from baseline as a proportion of the duration of the segment [$\chi^2(1) = 0.65$, $p = 0.42$]. The interaction between word and change in consonant target does not significantly improve model fit either [$\chi^2(1) = 3.49$, $p = 0.06$ for change in milliseconds; $\chi^2(1) = 0.50$, $p = 0.48$ for change as proportion of baseline segment duration].

It may also be the case that trial-to-trial variability in implementation obscures the overall relationship between C and V duration. To address this, we also tested for a relationship between a speaker's mean change in C and V

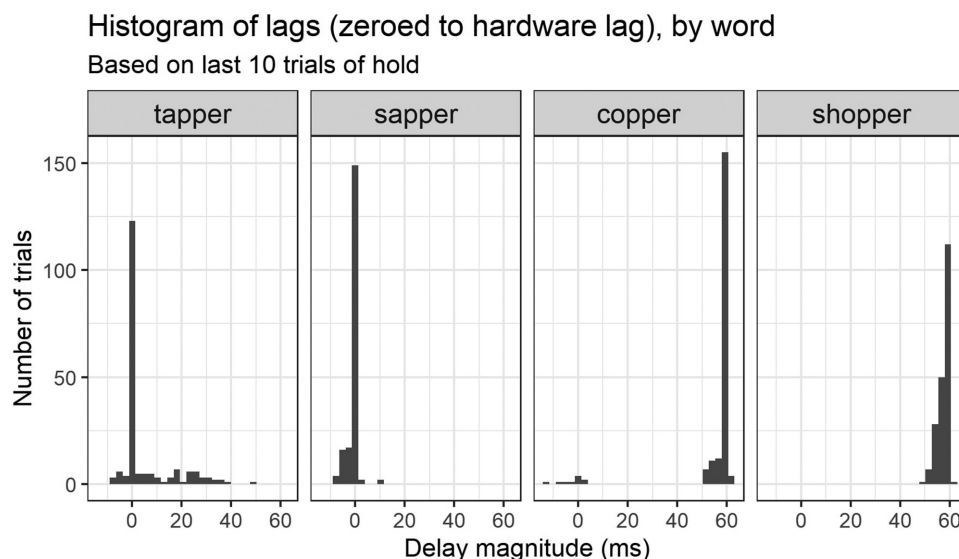


FIG. 5. Histograms of the feedback delay of the stressed vowel for each word, with hardware lag removed (~ 13.3 ms). Note the much wider spread in *tapper* than in the other target words.

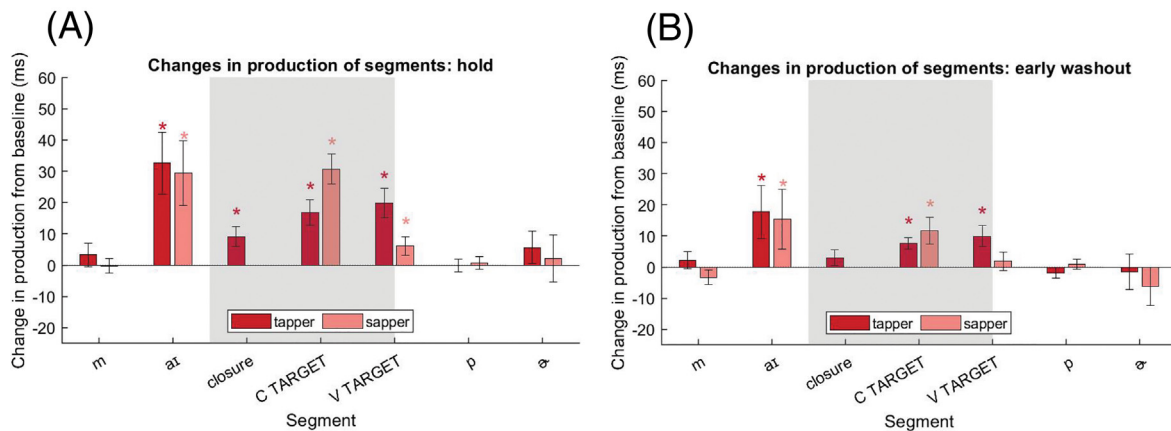


FIG. 6. (Color online) Changes in duration in each segment in the cross-category condition, during hold (A) and early washout (B). The gray panel indicates segments that were delayed.

duration in each word. Here, too, the addition of change in consonant target did not significantly improve on a model that already had word in it [$\chi^2(1)=0.46$, $p=0.23$]. Using change as proportion of original segment duration does not improve model fit either [$\chi^2(1)=3.47$, $p=0.06$]. In sum, the lengthening in the vowel is not explained by co-lengthening with the consonant.

D. Non-targeted segments

In this section, we report the behavior of the non-targeted segments: /m/ and /at/ in “my,” as well as the /p/ and /ø/. Although the remaining segments are not included in the main hypotheses of this experiment, they are still valuable to examine as they can provide insight on the effects of delay.

1. Duration of /p/

The /p/ in the target words was delayed in the within-category condition and played back veridically in the cross-category condition. The addition of phase significantly improves model fit [$\chi^2(2)=268.03$, $p<0.0001$]. Overall, /p/ is longer during hold ($p<0.0001$) but returns to baseline by washout ($p=0.37$).

Adding perturbation condition also significantly improves model fit [$\chi^2(1)=203.46$, $p<0.0001$], as does the interaction between phase and perturbation condition [$\chi^2(2)=280.87$, $p<0.0001$]. In words where /p/ was delayed, the duration of /p/ significantly increases during hold (16.4 ± 1.2 ms, $p<0.0001$), but returns to baseline in washout (1.8 ± 1.1 ms, $p=0.11$). In words where /p/ was not delayed, the duration of /p/ does not change in either phase (hold: -0.1 ± 1.2 ms; washout: 0.8 ± 1.1 ms; both $p=1.00$). The addition of consonant type does not significantly improve model fit, either as a single factor [$\chi^2(1)=2.26$, $p=0.13$], or in interaction with any other factors. This suggests that the main cause of lengthening in /p/ is an online reaction to delayed feedback, as both of the words with delayed onset of /p/ show lengthening during hold only, but there was no change in the words where participants received veridical feedback for /p/. The behavior of /p/ is illustrated in Fig. 6 (unperturbed) and Fig. 7 (delayed).

2. Duration of /ø/

The /ø/ in the target words was delayed in the within-category condition, and played back veridically in the cross-

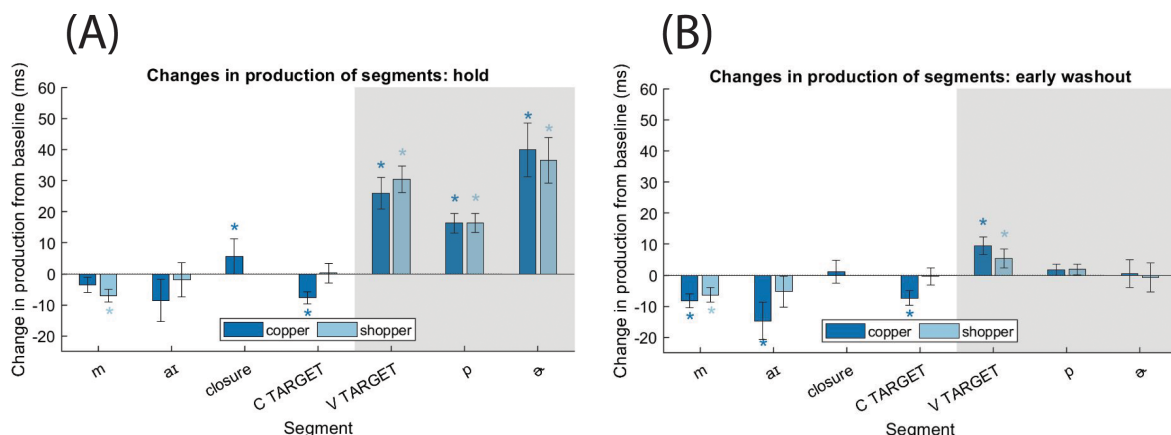


FIG. 7. (Color online) Changes in duration in each segment in the within-category condition, during hold (A) and early washout (B). The gray panel indicates segments that were delayed.

category condition. The addition of phase significantly improves model fit [$\chi^2(2)=287.16$, $p<0.0001$]. Overall, / ϑ / is longer during hold ($p<0.0001$), but not during washout ($p=0.50$).

Adding perturbation condition also significantly improves model fit [$\chi^2(1)=108.22$, $p<0.0001$], as does the interaction between phase and perturbation condition [$\chi^2(2)=165.88$, $p<0.0001$]. In words where / ϑ / was delayed, the duration of / ϑ / significantly increases during hold (38.2 ± 2.5 ms, $p<0.0001$), but returns to baseline in washout (-0.1 ± 2.5 ms, $p=1.00$). In words where / ϑ / was not delayed, the duration of / ϑ / does not significantly change in either phase (hold: 4.2 ± 2.5 ms; washout: -3.6 ± 2.5 ms; both $p>0.17$).

3. Duration of /m/

The /m/ in “my” was unperturbed in all words. Adding phase significantly improves model fit [$\chi^2(2)=72.04$, $p<0.0001$]. The duration of /m/ is significantly shorter during both hold ($p=0.04$) and washout ($p<0.0001$).

The addition of perturbation condition (following vowel lengthened vs no perturbation in following vowel) significantly improves model fit as well [$\chi^2(1)=38.84$, $p<0.0001$], as does the interaction between perturbation condition and phase [$\chi^2(2)=18.46$, $p<0.0001$]. The /m/ is significantly shorter only in words where the following vowel was not perturbed in both hold (5.3 ± 1.2 ms, $p=0.0001$) and washout (-7.2 ± 1.2 ms, $p<0.0001$); /m/ in words where the following vowel was lengthened are not significantly different in either phase (hold: 1.3 ± 1.2 ms, $p=0.93$; washout: -0.8 ± 1.2 ms, $p=0.97$). The behavior of /m/ is illustrated in Figs. 5 and 6.

4. Duration of /aɪ/

The /aɪ/ in “my” received a lengthening perturbation in the cross-category condition, and was not perturbed in the within-category condition. Adding phase significantly improves model fit [$\chi^2(2)=72.04$, $p<0.0001$]. Overall, the duration of /aɪ/ increases during hold ($p<0.0001$) and remains somewhat elevated during washout ($p=0.05$).

The addition of perturbation condition (lengthening vs no perturbation) also significantly improves model fit [$\chi^2(1)=387.53$, $p<0.0001$], as does the interaction between perturbation condition and phase [$\chi^2(2)=173.7$, $p<0.0001$]. Words where /aɪ/ was lengthened show lengthening during both hold (34.01 ± 3.1 ms, $p<0.0001$) and washout (18.7 ± 3.1 ms, $p<0.0001$). It should be noted that lengthening follows the perturbation, rather than opposing it. In contrast, words where /aɪ/ was not perturbed are shorter during washout (-9.9 ± 3.1 ms, $p<0.0001$) but not during hold (-5.3 ± 3.1 ms, $p=0.12$). These results are similar to the pattern observed for /m/ and suggest an overall shortening of the word “my” in words over the course of the study. The behavior of /aɪ/ is illustrated in Fig. 5 (unperturbed) and Fig. 6 (lengthened).

IV. DISCUSSION

In this study, we investigated the source of the apparent asymmetry between syllable onsets and vowels in temporal adaptation: previous studies have suggested that syllable onsets do not adapt to duration perturbations, but vowels are highly responsive to such perturbations. We found that speakers did adapt the duration of syllable onset targets, but did not adapt all segments equally. Specifically, both the VOT target and the fricative target in the cross-category condition showed adaptation, but only the VOT target of the within-category condition showed adaptation. We also found that speakers show local DAF effects in both consonants and vowels, lengthening specifically the segments that are delayed, even if the duration of that segment is unperturbed or if other segments in the word are not delayed.

Our first hypothesis was that speakers adapt to perceived cross-categorical errors, but not to perceived errors that do not cross a category boundary. This predicts that using a perturbation that moves the consonant target towards a categorical boundary would promote temporal adaptation of that consonant, but a perturbation that simply produces a longer version of that phoneme would not promote adaptation. The results partially support this hypothesis. We found that speakers lengthened the VOT in /t/ and the duration of frication in /s/ in response to a cross-category, shortening perturbation. In line with our hypotheses for the within-category condition, speakers did not adapt the duration of /f/; however, contrary to hypothesis, speakers did adaptively shorten VOT in /k/. Crucially, the duration of the consonant targets in /k/, /t/, and /s/ remained significantly different from baseline after perturbation was removed, indicating that the changes were not solely an online response to delay. Statistically, speakers showed the same magnitude of adaptation for all three segments (13.8%, 16.8%, and 16.2% of the perturbation or -7.7 , 7.7 , and 11.7 ms for /k, t, and s/, respectively).

It remains unclear why cross-category perturbations seem to promote temporal adaptation in syllable onsets, particularly in light of the result that speakers also adapted the VOT of /k/ under a within-category perturbation in this experiment. One possible explanation is that speakers attend more strongly to errors where the production is well outside the typical distribution [but not so different from the intended production so as to reduce a speaker’s perception of self-agency, cf. Subramaniam *et al.* (2018)]. In this conceptualization, categorical errors are not a necessary condition for adaptation, but rather a subset of productions that are very unusual for that segment. The distribution of durations is generally somewhat tighter on a population level in VOT (*copper*: M 59.6 ms, SD 17.3 ms; *tapper*: M 67.7 ms, SD 17.8 ms) than in fricative duration (*shopper*: M 131.3 ms, SD 28.9 ms; *sapper*: 131.8 ms, SD 20.2 ms). In addition, the mean perturbation of 54 ms in *copper* produced a 96% increase in duration over baseline, compared to an increase in only 46% for *shopper*. As such, it is possible that the perturbation on *copper* produces a more outlying token

than a perturbation of the same magnitude on *shopper*, thus encouraging adaptation. This could also potentially explain the lack of adaptation observed in “capper” in Karlin *et al.* (2021), where the perturbation was somewhat smaller and thus would not have produced such an unusual token (42 ms, or 65% increase from baseline VOT).

However, the data from the vowels in this study lend further support to the idea that onset consonants are not controlled or monitored fundamentally differently from vowels. Oschkinat and Hoole (2020) reported two types of vowel change: First, when a perturbation delayed and shortened the vowel, speakers lengthened the vowel during hold but did not show any aftereffects. Second, when a perturbation lengthened the vowel, speakers shortened the vowel during hold (which is necessarily adaptive, not compensatory), and showed aftereffects for several trials; this type of perturbation was also implemented by Floegel *et al.* (2020), with similar results. The asymmetry in adaptation between delayed/shortened vowels and lengthened vowels suggests that vowels too may be more likely to adapt when a perceived error crosses a categorical boundary. In both studies, which were conducted with German speakers, a perturbation that shortened the vowel did not risk impinging on another vowel category, but a perturbation that lengthened the vowel did. Even though the corresponding words with long vowels do not actually exist [e.g., [br:f] for Floegel *et al.* (2020), where the corresponding long vowel is /i/; /na:pfku:xən/ for Oschkinat and Hoole (2020)], German speakers may perceive an out-of-distribution error when those vowels are lengthened, possibly enhanced by a language-specific heightened sensitivity to vowel length differences. Thus, it could be the case that speakers monitor both syllable onsets and vowels for out-of-distribution errors, rather than vowels being monitored for all types of duration errors. Further investigation is necessary to investigate the role of category boundaries vs outlying tokens in error monitoring and adaptation.

One limitation of this study is that the two perturbation conditions in this study are not fully independent of other potential factors. For example, the cross-category condition specifically shortened consonant targets, and also delayed the onset. In contrast, the within-category condition lengthened consonant targets, without any delay. It may be the case that it is intrinsically more difficult to shorten segments than to lengthen them, in which case the within-category condition was at a distinct disadvantage. It is also possible that the lengthening observed in the cross-category condition is largely the result of the delay. However, the consonant durations remained elevated after perturbation was removed, at a statistically similar magnitude as the shortening observed in *copper* (16.2% of the perturbation for *tapper*, 16.8% for *sapper*, 13.8% for *copper*), suggesting that speakers did adapt based on the duration perturbation in addition to any potentially delay-induced lengthening. A study that fully crosses shortening and lengthening with category boundaries would provide additional insight.

Our second hypothesis was that speakers in Oschkinat and Hoole (2020) and Karlin *et al.* (2021) lengthened the vowel at least in part due to the incidental delay of the vowel onset, and not just in response to the shortening. We found that speakers lengthened the stressed vowel consistently in the within-category condition, when vowels were only delayed, and not shortened. This suggests that previous reports of speakers dramatically increasing vowel duration were also largely due to DAF effects, rather than adaptive lengthening in response to shortening. Interestingly, Oschkinat and Hoole (2022) reported that one participant had to be excluded as they repeatedly produced disfluencies when perturbation was active (producing “Tschetschenen” as “Tschetschenenen”). Since larger (continuous) delays have been reported to cause disfluencies and stuttering-like behavior in typical speakers, it is possible that the disfluencies were the result of the delay, even though only the vowel segment was delayed in that study.

Further evidence in favor of this hypothesis comes from the other segments in the phrase. Each perturbation condition only delayed some segments in the target word: the cross-category condition only delayed the first syllable onset (with some spillover into the stressed vowel in *tapper*), while the within-category condition only delayed segments after the first syllable onset. Speakers’ lengthening patterns largely follow these patterns: speakers lengthened /p/ and /ʌ/ when they were delayed (within-category condition), but did not change their productions when they received veridical feedback. In these segments, speakers returned to baseline production values when perturbation was removed, indicating that DAF effects can be local, online reactions to temporal displacement errors, and are not necessarily due to a global adjustment for the delay. Finally, in our analysis of *tapper*, where the perturbation was more inconsistent, the magnitude of lag on the stressed vowel was positively correlated with the magnitude of lengthening on a trial-to-trial basis. This further supports the idea that DAF effects are reactions to local temporal errors, and also indicates that DAF effects are not an all-or-nothing lengthening based on a binary distinction between detectable vs non-detectable lag.

One unexpected result in this study was that vowel duration remained slightly elevated in washout for all words but *sapper* (9.5 ms for *copper*, 9.7 ms for *tapper*, 5.4 ms for *shopper*). This may suggest that speakers may have some degree of DAF-related learning, which is inconsistent with the idea that DAF effects are local reactions to delay. One possibility is that the residual vowel lengthening is a product of some acclimatization to DAF and accompanying changes to the forward model (Malloy *et al.*, 2022). Further investigation is needed to probe why vowels maintained slightly elevated durations, while both the following /p/ and /ʌ/ returned to baseline immediately in washout.

A second unexpected result was that in the cross-category condition, speakers lengthened the vowel /aɪ/ in “my” in response to a lengthening perturbation. This is unexpected because Oschkinat and Hoole (2020) reported

that speakers shortened vowels in response to lengthening perturbations. However, it is possible that this too is the result of DAF effects, as the vowel /aɪ/ is a diphthong and thus has apparent temporal structure, unlike the monophthongs used in all previous studies (Floegel *et al.*, 2020; Karlin *et al.*, 2021; Oschkinat and Hoole, 2020, 2022). In the current study, perturbation began as soon as the vowel was detected, and lengthened only the first 20 ms of the detected vowel. As such, the [a] portion of the diphthong was lengthened, effectively delaying the onset of the [ɪ] portion and triggering DAF-like effects. However, the vowel duration also remained considerably elevated after perturbation was removed (18.7 ms longer), suggesting that the shift was not totally related to delay effects (compare residual lengthening of the delayed vowel during washout in *copper* and *shopper* at 9.5 and 5.4 ms, respectively). One possibility is that speakers were more likely to pause or hesitate before the targeted segments, perhaps related to increased cognitive load due to error processing (Adkins *et al.*, 2022; Jentzsch and Dudschig, 2009) or motor planning (Fox Tree and Clark, 1997; Seifart *et al.*, 2018). A full investigation of this phenomenon is beyond the scope of this study and is left for future work.

Our third hypothesis was that the gestural coordination between syllable onsets and vowels is such that when speakers plan to lengthen one, they also must lengthen the other. This predicts that vowels would increase in duration in parallel with adaptive lengthening in the syllable onset. However, this hypothesis was not supported by the data in this study. Although there was some vowel lengthening observed in the cross-category condition, when vowels were not perturbed, we found that there was no correlation between the degree of vowel lengthening and the degree of syllable onset lengthening in either hold (where the relationship may be obscured by delay-induced lengthening) or in washout (where all lengthening is adaptive). There was no relationship either on a trial level or on a by-participant level. This indicates that the coordination between onset and nucleus gestures does not prevent independent adaptation of each segment. Furthermore, it suggests that the adaptive shortening found in /k/ in this study [compared to the lack of adaptation in Karlin *et al.* (2021)] was not the result of freeing the consonant from contradictory adaptation in the vowel. Kinematic research on the temporal relationships between the articulatory gestures for the onset and vowel could shed further light on this issue.

V. CONCLUSION

Overall, this study provides evidence that speakers attend to both duration and the temporal alignment between the predicted timing and perceived timing of segment onsets/offsets when monitoring temporal aspects of speech. Crucially, this study provides evidence against the idea that syllable onsets are motorically inflexible while vowels are hyper-adaptive. Instead, we have shown that syllable onsets may adapt to perceived errors in duration if there is a

significant error, as mediated by the specific linguistic system of the speaker. In addition, we have shown that some of the previously reported behavior of vowels stems from DAF effects, rather than duration-based adaptation. Importantly, this study provides evidence that DAF effects can be a local reaction to delay, rather than a global strategy implemented across an entire utterance, suggesting that DAF effects may be a response of the motor system to a perceived mismatch from the predicted time course of the speech plan. Using altered auditory feedback to investigate asymmetries in temporal control provides insight on how linguistic and motoric aspects of speech timing are controlled and monitored.

ACKNOWLEDGMENTS

This work was supported by NIH Grant Nos. R01 DC017091 and F32 DC019535.

- Adkins, T. J., Zhang, H., and Lee, T. G. (2022). "What happens after an error?," *bioRxiv* 2022.03.17.484792.
- Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). "lme4: Linear mixed-effects models using Eigen and S4," R Package, version 1(7), pp. 1–23.
- Baum, S. R., and Blumstein, S. E. (1987). "Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English," *J. Acoust. Soc. Am.* **82**(3), 1073–1077.
- Bjorndahl, C. (2018). "A story of /v/: Voiced spirants in the obstruent-sonorant divide," Ph.D. thesis, Cornell University, Ithaca, NY.
- Browman, C. P., and Goldstein, L. (1988). "Some notes on syllable structure in articulatory phonology," *Phonetica* **45**(2–4), 140–155.
- Browman, C. P., and Goldstein, L. (1989). "Articulatory gestures as phonological units," *Phonology* **6**(02), 201–251.
- Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (2008). "A system for online dynamic perturbation of formant trajectories and results from perturbations of the Mandarin triphthong /iaʊ/, in *Proceedings of the 8th ISSP*, pp. 65–68.
- Cole, R. A., and Cooper, W. E. (1975). "Perception of voicing in English affricates and fricatives," *J. Acoust. Soc. Am.* **58**(6), 1280–1287.
- Floegel, M., Fuchs, S., and Kell, C. A. (2020). "Differential contributions of the two cerebral hemispheres to temporal and spectral speech feedback control," *Nat. Commun.* **11**(1), 1–12.
- Fox Tree, J. E., and Clark, H. H. (1997). "Pronouncing 'the' as 'thee' to signal problems in speaking," *Cognition* **62**(2), 151–167.
- Francis, A. L., Kaganovich, N., and Driscoll-Huber, C. (2008). "Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English," *J. Acoust. Soc. Am.* **124**(2), 1234–1251.
- Goedemans, R., and van Heuven, V. J. (1995). "Duration perception in sub-syllabic constituents," in *Proceedings of the 4th European Conference on Speech Communication and Technology*, pp. 1315–1318.
- Howell, P., and Powell, D. J. (1987). "Delayed auditory feedback with delayed sounds varying in duration," *Percept. Psychophys.* **42**(2), 166–172.
- Huggins, A. W. F. (1972). "Just noticeable differences for segment duration in natural speech," *J. Acoust. Soc. Am.* **51**(4B), 1270–1278.
- Jentzsch, I., and Dudschig, C. (2009). "Short article: Why do we slow down after an error? Mechanisms underlying the effects of posterror slowing," *Q. J. Exp. Psychol.* **62**(2), 209–218.
- Jongman, A. (1989). "Duration of frication noise required for identification of English fricatives," *J. Acoust. Soc. Am.* **85**(4), 1718–1725.
- Kalveram, K. T., and Jäncke, L. (1989). "Vowel duration and voice onset time for stressed and nonstressed syllables under delayed auditory feedback condition," *Folia Phoniatr. Logop.* **41**(1), 30–42.
- Karlin, R., Naber, C., and Parrell, B. (2021). "Auditory feedback is used for adaptation and compensation in speech timing," *J. Speech. Lang. Hear. Res.* **64**(9), 3361–3381.

- Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2015). "R Package Version," Package 'lmerTest,' version 2(0).
- Lenth, R. (2019). "emmeans: Estimated marginal means, aka least-squares means," <https://CRAN.R-project.org/package=emmeans> (Last viewed June 11, 2020).
- Malloy, J. R., Nistal, D., Heyne, M., Tardif, M. C., and Bohland, J. W. (2022). "Delayed auditory feedback elicits specific patterns of serial order errors in a paced syllable sequence production task," *J. Speech. Lang. Hear. Res.* **65**(5), 1800–1821.
- Mitsuya, T., MacDonald, E. N., and Munhall, K. G. (2014). "Temporal control and compensation for perturbed voicing feedback," *J. Acoust. Soc. Am.* **135**(5), 2986–2994.
- Niziolek, C. A., and Guenther, F. H. (2013). "Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations," *J. Neurosci.* **33**(29), 12090–12098.
- Oschkinat, M., and Hoole, P. (2020). "Compensation to real-time temporal auditory feedback perturbation depends on syllable position," *J. Acoust. Soc. Am.* **148**(3), 1478–1495.
- Oschkinat, M., and Hoole, P. (2022). "Reactive feedback control and adaptation to perturbed speech timing in stressed and unstressed syllables," *J. Phon.* **91**, 101133.
- Seifart, F., Strunk, J., Danielsen, S., Hartmann, I., Pakendorf, B., Wichmann, S., Witzlack-Makarevich, A., de Jong, N. H., and Bickel, B. (2018). "Nouns slow down speech across structurally and culturally diverse languages," *Proc. Natl. Acad. Sci. U.S.A.* **115**(22), 5720–5725.
- Subramaniam, K., Kothare, H., Mizuiri, D., Nagarajan, S. S., and Houde, J. F. (2018). "Reality monitoring and feedback control of speech production are related through self-agency," *Front. Hum. Neurosci.* **12**, 1–8.
- Yates, A. J. (1963). "Delayed auditory feedback," *Psychol. Bull.* **60**(3), 213–232.