Auditory feedback is used for adaptive control of timing in speech

Robin Karlin,[1, a)] Chris Naber,[1] and Benjamin Parrell[2]

[1] *Waisman Center, UW-Madison*

[2] *Communication Sciences and Disorders, UW-Madison*

Real-time altered auditory feedback has demonstrated a key role for auditory feedback in both online feedback control and in updating feedforward control for future utterances. Much of this research has examined control in the spectral domain, and has found that speakers compensate for perturbations to vowel formants, intensity, and fricative center of gravity. The aim of the current study is to examine adaptation in response to temporal perturbation, using real-time perturbation of ongoing speech. Word-initial consonant targets (VOT for /k, g/ and fricative duration for /s, z/) were lengthened and the following stressed vowel (/æ/) was shortened. Overall, speakers did not adapt to lengthened consonants, but did lengthen vowels by nearly 100% of the perturbation magnitude in response to shortening. Vowel lengthening showed continued aftereffects during a washout phase when perturbation was abruptly removed. Although speakers did not adapt absolute consonant durations, consonant duration was reduced as a proportion of the total syllable duration. This is consistent with previous research that suggests that speakers attend to proportional durations rather than absolute durations. These results indicate that speakers actively monitor relative durations and update the temporal dynamics of planning units larger than a single segment.

[a)]rkarlin@wisc.edu

I. INTRODUCTION

The real-time altered auditory feedback experimental paradigm has provided ample evidence that auditory feedback plays a key role in both online feedback control (e.g., Burnett *et al.* 1998; Elman 1981; Purcell and Munhall 2006b) and in updating feedforward/predictive control (e.g., Houde and Jordan 1998; Jones and Munhall 2000; Patel *et al.* 2015; Purcell and Munhall 2006a) for future utterances. Perturbations of auditory feedback introduce auditory errors, or discrepancies between expected and perceived feedback. Inconsistent perturbations (in magnitude or direction) elicit online compensation for those perceived errors, where speakers change their motor behavior within an ongoing production in response to perturbed auditory feedback. Speakers can also learn from the errors produced by consistent perturbation and adapt their motor plans for future utterances to incorporate information from those errors. Adaptation is seen both while the feedback perturbation is present as well as after perturbation has been removed. These "aftereffects" are a clear sign that mismatches between sensory feedback and motor predictions caused changes in feedforward/predictive control. To date, much of this research has examined control in the spectral domain: correction for perceived error has been demonstrated for vowel formants (e.g., Houde and Jordan 1998, 2002; Purcell and Munhall 2006a; Tourville *et al.* 2008), f0 (e.g., Burnett *et al.* 1998; Jones and Munhall 2000), intensity (e.g., Patel *et al.* 2015), and fricative spectral center of gravity (e.g., Casserly 2011; Klein *et al.* 2019; Shiller *et al.* 2009). These studies have shown that spectral components of auditory feedback are used by the

sensorimotor control system for both online control as well long-term adjustments to speech actions.

However, relatively little is known about the role of temporal information in sensory feedback in speech motor control. Most studies that have examined the role of time in speech control invoke delayed auditory feedback (DAF), where speech is played back to the participants at varying delays ranging from 25 ms to 500 ms or more (Kalveram and Jäncke, 1989; Mitsuya *et al.*, 2017; Stuart and Kalinowski, 2015; Yates, 1963). In reaction to these perturbations, speakers slow their rate of speech and increase vowel length; large delays additionally induce speech errors and stuttering-like behavior. One study examined the effects of both delayed and advance auditory feedback: speakers respond to auditory feedback that is presented before onset of speech by speeding up the rate of articulation, though this effect was only reported at a lead time of -50 ms; larger lead times (-100 and -150 ms) and delayed feedback (lag times of 50, 100, and 150 ms) did not induce these temporal effects (Mochida *et al.*, 2010). Increased feedback delays also attenuate adaptation for formant perturbations, suggesting that auditory feedback is temporally specific (Max and Maffett, 2015; Mitsuya *et al.*, 2017; Shiller *et al.*, 2020).

While these studies have shown a role for the monitoring of temporal information in online speech control, there has been relatively little work on if and how temporal feedback is used in speech motor control. As for spectral feedback, this question can be probed by examining how speakers respond to auditory perturbations in the time domain, i.e. shortening or lengthening segments or portions of segments. Unlike spectral perturbation studies, temporal perturbation studies must either assess long-term adaptation of a perturbed segment,

or take into account the possibility of compensation in segments following the perturbed segment. This is due to the simple fact that a speaker cannot perceive the duration of a segment until after it has been completed, at which point it is impossible to enact online compensation of that particular segment's duration. One study (Ogane and Honda, 2014) found that speakers adapt to slowed formant transitions by increasing the velocity of both F1 and F2. A similar study using inconsistently applied perturbations (Cai *et al.*, 2011) reported that speakers delay subsequent articulatory movements in response to delayed formant transitions, but do not adjust for accelerated formant transitions.

Other studies have manipulated steady state portions of segments, rather than transitions. Mitsuya *et al.* 2014 pre-recorded instances of "tipper" (long-lag VOT) and "dipper" (short-lag VOT); for the perturbation phase of the study, participants heard their pre-recorded tokens of "tipper" when they said "dipper", and vice versa. In response, participants lengthened VOTs when they heard a shorter VOT than what they produced, and shortened VOTs when they heard a longer VOT. In addition, participants produced longer vowels when they heard shorter vowels (where "tipper" has a shorter stressed vowel than "dipper") but did not alter produced vowel duration when they heard longer vowels. Unlike the changes in VOT, the changes in vowel production disappeared immediately when perturbation was removed; this indicates that the VOT adjustments were learned adaptation, while the vowel effects may have resulted from purely compensatory mechanisms. As the onset of the shortened vowels was also delayed, it is possible that speakers were responding to delayed auditory feedback, rather than increasing their planned vowel durations.

Two recent studies employing online feedback manipulations have reported temporal adaptation in syllable nuclei and codas, but not in syllable onsets. Floegel *et al.* (2020) lengthened either the nucleus or coda of a CVC syllable. They reported that participants adaptively shortened both vowels and coda consonants, with no difference in magnitude of adaptation between the two. Oschkinat and Hoole (2020) compared adaptation between the complex consonant /pf/ in onset and coda position. In the 'onset condition' of their study, /pf/ in syllable onset position was lengthened and the following /a/ was shortened ("**pf***a*nnkuchen"); speakers did not significantly change their productions of /pf/ in onset position, but did lengthen the vowel /a/ by 8.8% of the baseline segment duration (raw change approximately 11.5 ms). In the 'coda condition', the /a/ nucleus was lengthened and the following coda /pf/ shortened ("n**a***pf*kuchen"); speakers both adaptively shortened the vowel by 10.3% of the baseline duration (raw change approximately 9 ms) and adaptively lengthened the consonant by 17.2% (raw change approximately 34 ms). Changes seen during application of the perturbation persisted into the washout phase of the experiment only in the coda condition. The authors suggested that the asymmetry adaptation reflects the higher stability of inter-segment timing between onset and nucleus, as compared to the timing between nucleus and coda, which would impede motor adaptation.

While these studies have demonstrated that temporal control of at least some aspects of speech relies on monitoring of auditory feedback, a number of open questions remain. First, unlike most studies of adaptation and compensation using perturbed auditory feedback, Mitsuya *et al.* did not produce altered feedback by perturbing productions on-line, but rather by playing pre-recording speech. Thus, any adjustments produced by participants did

not have a corresponding effect in the auditory feedback. The authors did not systematically assess whether or not participants were aware that they were not hearing their pre-recorded speech, but it is conceivable that participants noticed that their feedback was altered. It is unclear if and how the perception of auditory feedback as externally or self-generated affects its use in sensorimotor adaptation, but there is some evidence from reaching and pitch control that perturbations which are perceived as externally generated induce limited changes in the sensorimotor control system (Behroozmand and Larson, 2011; Korzyukov *et al.*, 2017; Liu *et al.*, 2010; Wei and Körding, 2009).

Second, Mitsuya *et al.* (2014) examined only temporal perturbations which caused the perception of a different phonemic category. They shortened VOT for /t/ (played "dipper" when participants produced "tipper") and lengthened VOT for /d/ (played "tipper" when participants produced "dipper"), thus crossing the category boundary in both cases. In contrast, the manipulations did not cross a phonemic boundary in either Oschkinat and Hoole (2020) or Floegel *et al.* (2020). Oschkinat and Hoole (2020) notes that this difference in methodology may have contributed to their reported lack of compensation in syllable onsets: the effects of perturbing the duration of speech segments may interact with phonemic boundaries, as has been shown for spectral perturbations (Mitsuya *et al.*, 2013; Niziolek and Guenther, 2013).

Third, responses to different segments (stops, fricatives, ffricates or stop/fricative clusters) in different studies may vary because neuromotor control of time and timekeeping is not uniform. A diverse body of research has indicated two distinct types of timekeeping in the brain: absolute timing, which is based on the duration of a single event, and relative timing,

which occurs between multiple events (Zelaznik *et al.*, 2005). Some work has shown that these two types of timing are centered in distinct areas of the brain, with absolute timing involving a cerebello-olivary circuit, and relative timing relying on a circuit between the basal ganglia, thalamus, and supplementary motor area (Teki *et al.*, 2011). Correspondingly, the ability to judge intervals based on absolute timing and adapt to perturbations of absolute timing is impaired in people with cerebellar damage, while relative timing abilities remain intact (Grube *et al.*, 2010; Ivry *et al.*, 2002). In speech, both types of timing are present, and as such may produce different adaptation effects. VOT in stops is the result of the relative timing and coordination of stop constrictions and glottal adduction and abduction, while the duration of a fricative or a vowel involves the absolute timing of a single constriction. This suggests that there may be different temporal control mechanisms at work for these two aspects of speech timing.

Finally, there is evidence from both perception (Denes, 1955; Port and Dalby, 1982) and production (Boucher, 2002; Kessinger and Blumstein, 1998) that speakers may not control the absolute duration of individual segments, but rather attend to proportional relationships relative to other segments in some higher level unit, such as the syllable (Fowler, 1981; Munhall *et al.*, 1992). For example, there is some evidence that the effect of vowel duration in the perception of coda voicing in English is affected by the proportional durations of the vowel and coda closure (Denes, 1955; Port and Dalby, 1982); similarly, it has been argued that VOT duration is compared to vowel duration (Boucher, 2002; Port and Dalby, 1982). It has also been shown that the ratio of VOT to syllable duration remains constant over changes in speech rate (Boucher, 2002; Kessinger and Blumstein, 1998). For this reason, Mitsuya

147 *et al.* (2014) did not endeavor to alter single segments, but rather the whole utterance:
148 by playing back "tipper" when participants said "dipper", both the VOT and the vowel
149 duration changed, both of which contribute to the perception of voicing. Thus, speakers
150 may compensate for temporal perturbations to one segment by partially adjusting adjacent
151 segments to preserve a desired relative duration, instead of directly adjusting the perturbed
152 segment to preserve absolute duration. Segments within a syllable are tightly coordinated
153 with one another (Browman and Goldstein, 1986; Fowler, 1981; Gafos, 2002). Thus, as
154 suggested by Oschkinat and Hoole (2020), adjusting the movements for a single segment
155 would effectively adjust the timing relationships of an entire coordinative structure.

156 The present study addresses outstanding questions regarding the role of auditory feedback
157 in adaptation of temporal control in speech. First, this study attempts to replicate the study
158 reported in Mitsuya *et al.* 2014, which found temporal adaptation for VOT in syllable onsets,
159 with the addition of real-time perturbation of timing, contingent on ongoing production.
160 Second, we test adaptation in two potentially different types of speech timing: relative timing
161 between two distinct actions (VOT, i.e. the relationship between consonant closure and the
162 onset of voicing), and the absolute timing of a single action (the duration of fricatives and
163 vowels). Finally, we assess adaptation both in proportional timing of segments in a syllable
164 and in absolute timing of individual segments.

II. METHODS

A. Participants

20 speakers (18 F, 2 M) participated in the study, ranging in age from 18 to 66 (mean 26.0 years, median 20.5 years). No participant reported any history of speech, hearing, or neurological disorders. All participants gave informed consent. Participants were compensated for their participation either monetarily or through extra credit in a course in the UW-Madison Communication Sciences and Disorders department. All procedures were approved by the Institutional Review Board at UW-Madison.

B. Task

There were four target consonants, /g, k, z, s/, as well as one target vowel, /æ/. There were thus four stimulus words: gapper, capper, zapper, and sapper. All words are either nonce words or highly infrequent, minimizing potential effects of word frequency on compensation magnitude. Each participant completed the experiment in two sessions, conducted on two different days. The experiment was split into two sessions in part to keep the duration of an experimental session manageable, and in part to minimize potential carryover effects of adaptation between different words. Each session had two word blocks, which were formed of one stop and one fricative, and one voiced and one voiceless consonant—e.g., session 1 may include the blocks with gapper and sapper, with session 2 including blocks for capper and zapper. Each session lasted approximately 50 minutes. The order of the sessions and the word blocks within each session was counterbalanced across participants.

Each block consisted of four phases: a 30-trial baseline phase with veridical feedback; a 30-trial ramp phase where the duration of the target segment was increased by ∼2 ms per trial; a 60-trial hold phase with the maximum perturbation (∼40 ms); and a 30-trial washout phase with veridical feedback. On each trial, the participant produced the phrase "a [TARGET WORD]". After the second session, participants completed a survey via Qualtrics to assess their awareness of the applied perturbations.

For /g, k/, VOT was lengthened, while for /s, z/ the fricative was lengthened; these targets will be referred to as "consonant targets". The consonant targets were chosen to examine differences in relative vs. absolute timing (stop VOT vs. fricative duration), as well as potential effects of moving towards a categorical boundary vs. moving away from the boundary (e.g., increasing VOT on a /g/ pushes the resulting consonant closer to the g/k boundary, while increasing VOT on a /k/ pushes it further from the g/k boundary). Due to inconsistencies in implementing temporal perturbation, /g/ was excluded from analysis (see Section II E below for details on exclusionary criteria). The vowel /æ/ immediately following the consonant was then shortened so that the overall syllable duration remained unchanged.

The experiment was presented in Matlab; time perturbation was achieved with Audapter (Cai *et al.*, 2008; Tourville *et al.*, 2013). Audio was recorded with an AKG 520 head-mounted microphone and played back over Beyer Dynamics DT 770 closed over-ear headphones at a level of ∼80 dB, mixed with noise at ∼60 dB. The noise served to mask participants' perception of their own, unaltered speech through either air or bone conduction. The delay between input and output signals was measured at ∼35 ms.

C. Implementation of temporal perturbation

Temporal perturbation of specific segments was achieved using Audapter's online status tracking (OST) capability, which uses heuristics based on root-mean-square (RMS) intensity to detect phonetic events such as segment boundaries (see Appendix A for sample OST values). OST settings were individualized for each participant during a pretest phase that preceded each word block. During this pretest phase, the participant read the target phrase nine times, using a standard set of OST parameters based on pilot testing to detect the segment boundaries in the target phrase. The experimenter then compared the placement of the detected boundaries relative to the actual segment boundaries and adjusted the OST parameters if necessary. Participants repeated this pretest phase using the adjusted parameters until the parameters did not have to be changed. These parameters were then used for the experimental phases of that word block.

During the experiment, detected segment boundaries were used to trigger time warping events in each trial. Time warping events in Audapter consist of an initial 'time dilation' period (where audio is resampled and played back more slowly), followed by a 'hold' period (audio played back at original speed, though at a delay depending on the magnitude of the time dilation), and finally a 'catch up' period (where audio is played back faster until the samples match incoming audio). The parameters of these periods were specified in a perturbation configuration file (PCF; see Appendix A for sample PCF values). The magnitude of time dilation and the speed used in the catch up period were standardized for all participants; the duration of the hold period was calculated for each participant based on

the pretest trials (for more detail on the parameters available to Audapter, see Cai 2014). For capper and gapper, the lengthening started when the release burst was detected, and for sapper and zapper, the lengthening started when the high frequency fricative noise was detected. The mean and standard deviation of the duration of the consonant target from the pretest trials were used to calculate the duration of the hold period of the time warp. This ensured that the catch up period would not re-shorten the lengthened consonant target, and would instead largely occur during the vowel target. Sample capper and sapper trials from the hold phase are provided in Figure 1. Examples of the parameters used for tracking and feedback alteration can be found in Appendix B.

There was some variability in the perturbation received for each block. Although the target perturbation magnitude was 60 ms for all participants and words, variability across participants and segments led to differences in perturbation received relative to the duration of the perturbed segment (see Table II). Proportional perturbation was calculated as the perturbation magnitude divided by the duration of the target segment in the same trial. Differences in mean proportional perturbation did not correlate with differences in mean compensation magnitude (see individual segment results for more detail). Variability within participant and word block also produced some inconsistency in perturbation received in each trial. As the detection of consonant targets relied on the presence of high frequency noise in the signal, trials where that noise was attenuated or where noise was introduced early (e.g., in the preceding vowel) caused some initial variation in how early the consonant target could be detected. In the case that the consonant target is detected late, there may not be sufficient material to lengthen; in the case that the consonant target is detected too

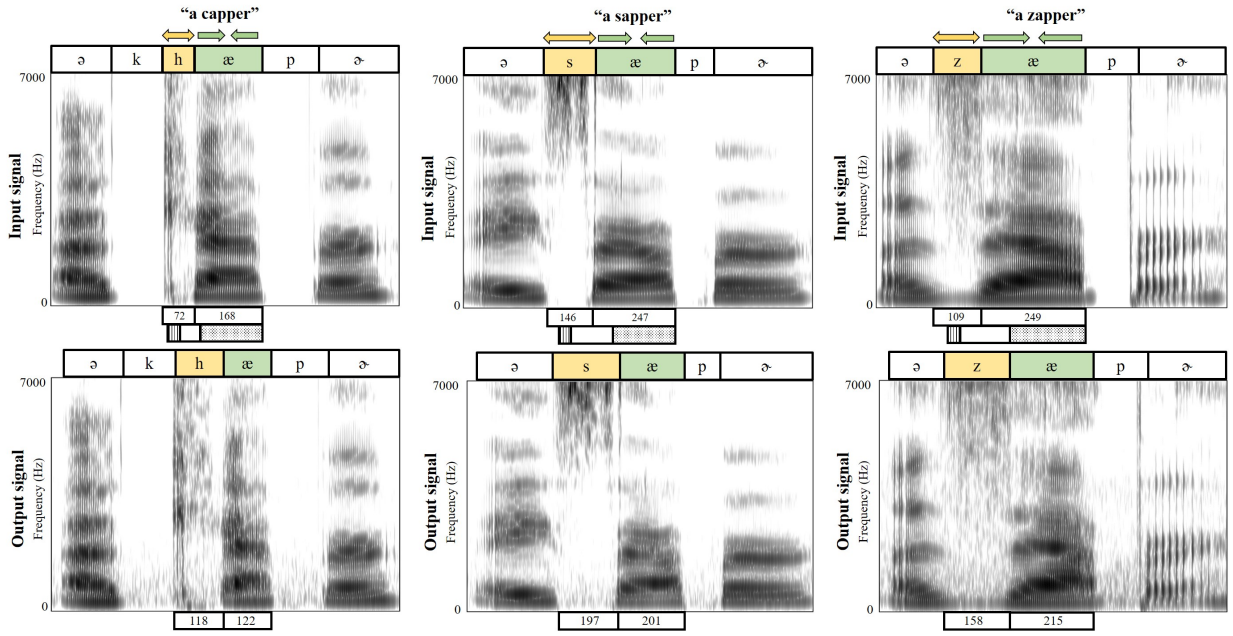

FIG. 1. Examples of the input (top) and output (bottom) signals from the hold phase, including the lag between signals. Left: "a capper". Center: "a sapper". Right: "a zapper". Target segment durations are given in ms below the spectrograms. Rectangles below the durations in the input signal indicate the time warp periods: stripes indicate the signal that underwent time dilation; unfilled indicates the hold period; dots indicate the catch up period. Noise in the output signal is due to the inclusion of noise in playback to mask participants from hearing their true, unaltered speech.

early, the lengthening would not apply exclusively to the consonant target. Variation in segment duration further complicated the issue, as the temporal parameters for perturbation had to be approximated using the pretest trials; if a participant substantially shifted their productions over the course of a session, those hard-coded parameters are no longer optimal for their speech. On occasion, within-participant variability resulted in the segments following the shortened vowel (/p/ and /ɚ/) sometimes receiving a small amount

	Cons. target		Vowel target		/p/		/ɚ/	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
capper	41.7	11.4	-33.7	16.1	-8.4	9.5	-2.5	9.1
sapper	43.5	3.2	-32.8	11.3	-10.9	11.0	0.6	5.0
zapper	40.2	9.9	-34.9	11.8	-6.6	8.0	0.1	3.9

TABLE I. Mean perturbation during the hold phase for each target word, by segment. Positive values indicate lengthening; negative values indicate shortening. Means are calculated over the dataset after the exclusion of trials due to insufficient perturbation. All units in ms.

of shortening. The mean perturbation achieved for each word and perturbation target are provided in Table I.

D. Statistical analysis

The audio from the participants' productions was hand-segmented to obtain the durations of each segment in the target utterance. Raters followed guidelines for segmentation, and the first author performed spot checks to ensure cross-experiment accuracy in segmentation. In almost all cases (58/60 words included in the analysis), any given block was segmented entirely by one person, minimizing potential spurious effects caused by interrater differences. The last 10 trials of the baseline phase served as a baseline of comparison for adaptation and aftereffects: adaptation was measured from the last 10 trials of the hold phase, in order to assess production at maximum learning (Daliri and Dittman, 2019; Lametti *et al.*, 2018;

	Cons. target		Vowel target	
	Mean	SD	Mean	SD
capper	68.7	15.7	-18.3	5.6
sapper	28.0	5.9	-16.4	2.7
zapper	40.3	7.9	-16.2	3.6

TABLE II. Mean perturbation during the hold phase for each target word, by segment, given as percent of the target segment duration. Positive values indicate lengthening; negative values indicate shortening. Means are calculated over the dataset after the exclusion of trials due to insufficient perturbation. All units in % of target segment duration.

Rochet-Capellan and Ostry, 2011); aftereffects were measured from the first 10 and last 10 trials of the washout phase (early and late washout, respectively). Reported estimated means are the change from the baseline, i.e. an increase or decrease in production duration. Positive values indicate that the segment is longer than baseline, and negative values indicate that the segment is shorter. In addition to the consonant and vowel target, analyses were also done on the vowel preceding the target word, the consonant closure of /k/, the /p/ following the target consonant, and the /schwar/ of the last syllable. As a proxy for speech rate, we analyzed utterance length from the onset of the article "a" to the end of the target word.

Although the target of time manipulation in this study was single consonants and vowels, altering the timing of a single movement alters its timing relationship with other movements

that it is coordinated with (Oschkinat and Hoole, 2020). In the acoustic signal, the consonant lengthening perturbation increases the proportion of the syllable that is taken up by that consonant; in addition to that, shortening the vowel further alters the proportions of the syllable such that it is far more heavily weighted to the consonant. For example, if a baseline production of sapper has an [s] duration of 150 ms, an [æ] duration of 200 ms, and a [p] duration of 50 ms, the proportion of [s] in the initial CVC syllable would be approximately 38%; that same token with a temporal perturbation of 40 ms (assuming no effect on [p]) would shift [s] to 190 ms and [æ] to 160 ms, increasing the proportion of [s] to 48%. In order to address the possibility that temporal control is implemented in a proportional, rather than absolute, manner, we also examine changes in the duration of the consonant target relative to the syllable (only the analysis for the consonant target proportions will be reported, as vowel proportions are effectively the complement of the consonant proportions).

We divide the target words into two syllables, where the /p/ is ambisyllabic and counted as both the coda of the first syllable and the onset of the second syllable (Elzinga and Eddington, 2014). Thus, the first syllable of sapper, for example, consists of [sæp]. There are three reasons for designating /p/ as ambisyllabic. First, this is the approach used in Mitsuya *et al.* 2014, with comparable words ("tipper" and "dipper"), and this allows for easier comparison between these studies. Second, all target words are nonce words composed of a real word CVC (cap, gap, sap, zap) with the addition of the agentive suffix –er. Third, there are differences in vowel duration associated with the voicing of intervocalic segments (e.g., "cabber" would have a longer vowel than "capper", see Lisker 1986), and as previously stated, the perception of voicing in English may be influenced by the proportion of vowel

duration to coda consonant duration. Syllable duration for each syllable is counted from the onset of the first segment to the offset of the last (i.e. onset of consonant closure or fricative duration for [k, s, z] to closure release for [p]). Proportions are calculated as the duration of the consonant target divided by the duration of that segment's syllable. As for individual segment durations, baseline proportions are calculated from the last 10 trials of the baseline phase, and changes in syllable proportions are measured as a change in percentage from that reference point. In order to ensure the uniformity of change in syllable proportion, only words that received both sufficient consonant and vowel perturbation are used in this analysis.

The data were analyzed with a linear-mixed effects model in R (R Core Team, 2019), using the package lme4 (Bates *et al.*, 2014). Onset consonants and vowels were analyzed separately. Models had fixed effects of word (representing different types of timing) and phase, as well as their interaction. Random intercepts were included for participant; random slopes were also tested for all models but the additional complexity in the random effects structure caused all models to either fail to converge or to have a singular fit. Models were built incrementally, and likelihood ratio tests used to compare models. Post-hoc tests were done using least means squared tests with a Bonferroni-Holm adjustment using the emmeans package (Lenth, 2019).

Models were run on both raw (ms) and normalized (change in production relative to received perturbation) data and produced the same overall results. Only the models for the raw data are presented; normalized results for vowel and consonant perturbation targets did not differ substantially from the raw values, and are provided in Appendix C.

E. Exclusions

Some data was excluded from analysis due to insufficient temporal perturbation. First, gapper has been excluded from analysis due to a high rate of perturbation failure: four out of 20 participants had the gapper block excluded prior to segmentation due to inconsistent burst detection, which led to lengthening segments other than the consonant target; a further three participants had gapper blocks excluded due to not reaching a minimum threshold of perturbation. The minimum threshold for temporal perturbation was set at three standard deviations below the mean perturbation for a word or 10 ms, whichever was higher. This threshold was computed separately for consonants and for vowels, thus analyses that involve only consonants use data from participants with adequate consonant lengthening, and analyses that involve only the vowel use data from participants with adequate vowel shortening. For the consonant analyses, this led to the exclusion of one zapper block from the dataset. For the vowel analyses, this led to the exclusion of four capper, one sapper, and two zapper blocks. No individual participant had any block excluded for both consonant and vowel analyses. A small number of individual trials (1.3% of tokens used for modeling) were excluded from analysis due to production errors (e.g., participant produced the wrong word, yawned during production, started production too late in the trial and was cut off).

III. RESULTS

A. Absolute duration of consonant target

Overall, participants did not adapt the absolute duration of consonant productions in response to lengthening (Figure 2). The addition of phase as a fixed effect significantly improves model fit ($\chi^2(3) = 8.59$, $p = 0.04$). However, the only two phases that are significantly different from each other are the hold (1.1 ± 1.4 ms) and late washout (-1.2 ± 1.4 ms) phases ($p = 0.02$). The main effect of phase is not indicative of adaptation, as neither the hold nor early washout (-0.1 ± 1.4 ms) phases differ from baseline ($p = 0.51$ and $p = 0.78$, respectively). The addition of word (as a proxy for timing type) as a second fixed effect does not significantly improve the model ($\chi^2(2) = 2.59$, $p = 0.27$); there is also no significant interaction between word and phase ($\chi^2(6) = 11.95$, $p = 0.06$).

As there was some variability between participants that resulted in differing magnitude of perturbation relative to segment duration, it is possible that speakers that received greater proportional perturbation were more likely to show adaptation. Proportional perturbation of the consonant targets did improve the model fit ($\chi^2(1) = 4.46$, $p = 0.03$), where participants that received greater proportional perturbation shortened their consonants more. However, it is likely that this is a spurious result, simply the consequence of attempting to account for a lack of group effect: participants that shortened their consonants would by definition have a larger proportional perturbation.

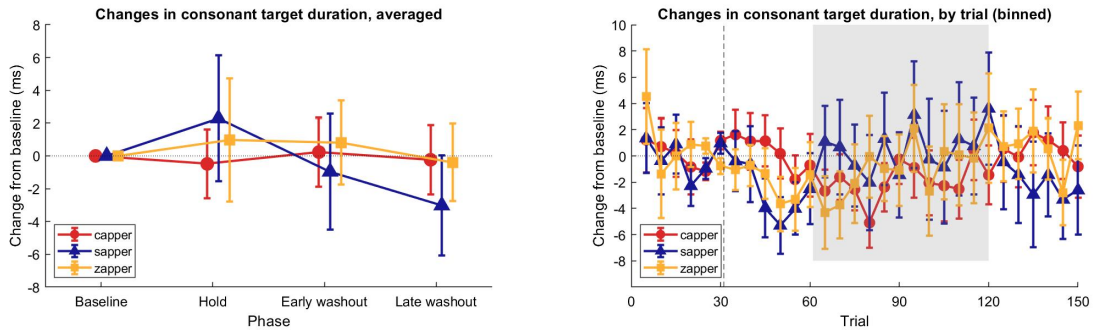

FIG. 2. Change from baseline in consonant target duration. Top left: by phase, averaged across participants (means ± standard error), including only data used in the model. Top right: behavior throughout the experiment, where each datapoint represents five trials, averaged across participants (means ± standard error). The dashed line indicates the beginning of the ramp phase and the shaded area indicates the hold phase.

B. Absolute duration of vowel target

In contrast with the response to perturbations of consonant duration, the shortened vowels in the perturbed auditory feedback led to systematic changes in production (Figure 3). Participants consistently lengthened their vowel productions during the hold phase and early washout phases and returned to baseline productions by the late washout phase. Phase as a fixed effect significantly improves the model fit ($\chi^2(3) = 801.11$, $p < 0.0001$); all phases are significantly different from each other ($p \leq 0.0001$ for all comparisons) except baseline and late washout ($p = 0.06$). Vowels are the longest in the hold phase (24.7 ± 1.2 ms). Vowels in the early washout phase (5.3 ± 1.2 ms) are shorter than those in the hold phase, but longer than in the baseline phase. Vowel duration returns to close to baseline values by the late washout phase (1.6 ± 1.2 ms).

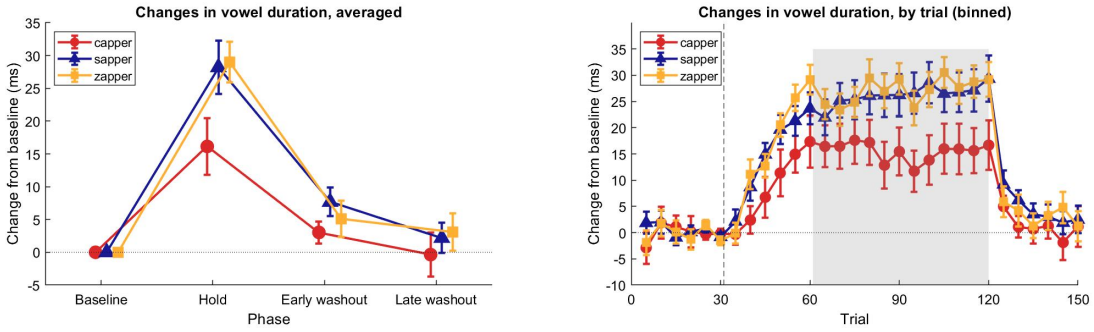

FIG. 3. Change from baseline in /æ/ duration. Top left: by phase, averaged across participants (means ± standard error). Top right: behavior throughout the experiment, where each datapoint represents five trials, averaged across participants (means ± standard error). Note the consistency in vowel lengthening across participant, compared to the highly variable consonant target behavior.

Adding word as a second fixed effect (as a proxy for timing type) also significantly improves model fit ($\chi^2(2) = 47.55$, $p < 0.0001$), as does the interaction between word and phase ($\chi^2(6) = 50.76$, $p < 0.0001$). The interaction is driven largely by differences in magnitude of adaptation in the hold phase across words. The vowel is lengthened less during the hold phase in capper (16.0 ± 1.6 ms) than either sapper (27.6 ± 1.5 ms, $p < 0.0001$) or zapper (29.2 ± 1.5 ms, $p < 0.0001$). During early washout, the vowel is closer to baseline levels in capper (2.8 ±1.6 ms) than sapper (7.3 ± 1.5 ms, $p = 0.03$), and in late washout also closer to baseline levels in capper (-0.6 ± 1.6 ms) than in zapper (3.4 ± 1.5 ms, $p = 0.03$). There are no other significant differences between words within phase (all $p \geq 0.21$).

Between words, the patterns of lengthening and returning to baseline are similar. For all words, vowels in the hold phase are significantly longer than in all other phases ($p <$ 0.0001 for all comparisons). For capper, this is the only phase that is significantly different

from any other phase. However, for both sapper and zapper, vowels are still longer in early washout than baseline (zapper $p = 0.02$; sapper $p < 0.0001$). For sapper, vowels are also significantly longer in early washout (7.3 ± 1.5 ms) than in late washout (1.8 ± 1.5 ms, $p = 0.0004$). Thus, while all words lengthened vowels during the hold phase, the rate of return to baseline levels varied between words. This could be simply due to the difference in magnitude of adaptive response; the vowel in capper did not increase during the hold phase as much as the vowel in either sapper or zapper. Unlike for the consonant data, the patterns of vowel lengthening are relatively consistent across participants—though there are differences in magnitude of change, for all blocks, at least half of participants lengthen the vowel in the hold phase, and the vast majority of participants lengthen the vowel in at least one of the blocks. Although there is some variation between individuals in perturbation magnitude relative to segment duration, mean proportional perturbation during the hold phase did not significantly predict mean adaptation ($\chi^2(1) = 0.44$, $p < 0.00010.51$).

C. Relative duration of consonant target as a proportion of the syllable

As may be expected given the results from the target consonants and vowel, the addition of phase significantly improves the model ($\chi^2(3) = 127.24$, $p < 0.0001$). The consonant takes up a lower proportion of the syllable duration during the hold phase (-1.6 ± 0.3%) than all other phases (all $p < 0.0001$), and the proportion of the syllable occupied by the consonant target is smaller in the early washout phase (-0.5 ± 0.3%) compared to baseline ($p = 0.02$). There was no difference between the late washout phase (-0.3 ± 0.3%) and baseline ($p = 0.12$). Results are illustrated in Figure 4.

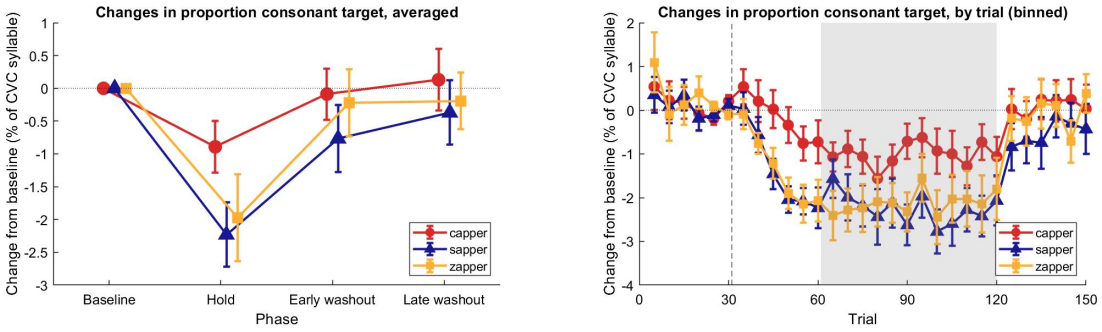

FIG. 4. Change from baseline of proportion consonant target in initial CVC syllable. Top left: averaged across participants (means ± standard error). Top right: behavior throughout the experiment, where each datapoint represents five trials, averaged across participants (means ± standard error). Bottom: individual data, by phase.

The addition of word as a second fixed effect also significantly improves the model ($\chi^2(2) = 17.82$, $p < 0.0001$). In this case, it is sapper that is significantly different than the other two words, where sapper overall shows more change across phases compared to baseline (-0.9 ± 0.3%) than capper (-0.4 ± 0.3%, $p = 0.0001$) and zapper (-0.5 ± 0.3%, $p = 0.006$). The addition of the interaction term between word and phase does not significantly improve the model ($\chi^2(6) = 11.64$, $p = 0.07$). Overall, the proportion of consonant target in the CVC syllable decreases during the hold phase, and the effect lingers through early washout. If temporal control in speech relies on relative rather than absolute durations, these effects suggest an overall adaptation of syllable timing in response to the perturbation.

D. Rate of speech

412 It is also conceivable that changes in vowel target duration reflect a broader change in
413 the timing dynamics of the utterance. However, although rate of speech was not explicitly
414 controlled in this study, participants overall maintained a consistent rate of speech across
415 the phases. The addition of phase as a fixed effect significantly improves model fit ($\chi^2(3) =$
416 109.17, $p < 0.0001$). Utterances produced during the hold phase are longer than all other
417 phases ($p < 0.0001$ for all comparisons); there are no significant differences between any
418 of the other phases ($p \geq 0.10$ for all comparisons). The magnitude of difference between
419 the hold phase and baseline utterance duration is roughly equivalent to the difference in
420 vowel target duration. Utterances are 23.7 ms $\pm$ 3.3 ms longer during the hold phase than
421 during the baseline phase; compare 24.7 ms $\pm$ 1.2 ms for the vowel alone. This indicates
422 that utterances are longer only because the vowel duration is increased; thus, differences in
423 vowel duration are not due to global changes in speech rate but instead targeted control of
424 timing in the first syllable.

425 ## E. Non-targeted segments

426 ### *1. Initial article "a"*

427 The article preceding the target word was not targeted for perturbation, but speakers
428 may have adjusted this vowel as part of the strategy of adjusting overall timing relationships.
429 The addition of word as a fixed effect does not significantly improve the model fit ($\chi^2(2) =$
430 3.03, $p = 0.22$). The addition of phase as a second fixed effect does significantly improve
431 the model ($\chi^2(3) = 7.57$, $p = 0.04$); however, the only phases that are significantly different

from each other are baseline and early washout (-6.2 ± 3.7 ms, p = 0.03). No other phases are statistically different from each other (all p > 0.17).

2. Stop closure of /k/

Using stop closure for "capper" instead of VOT duration changes the models slightly, but does not indicate that speakers adapted consonant closure in order to adjust overall timing of the consonant. The addition of phase as a fixed effect significantly improves model fit ($\chi^2(3) = 27.29$, $p < 0.0001$). However, there is no indication of adaptation; the hold phase (1.5 ± 1.5 ms) is significantly longer than both the early washout (-1.5 ± 1.5 ms, p = 0.002) and late washout (-2.6 ± 1.5 ms, $p < 0.0001$) phases, but not significantly different from baseline (p = 0.20). Closure duration is also longer in the baseline phase than late washout (p = 0.005). No other phases differ significantly from each other (all p > 0.16). This indicates that closure duration shortened slightly during the washout phase compared to the rest of the experiment. The addition of word as a second fixed effect also significantly improves the model ($\chi^2(2) = 11.29$, p = 0.004). Only capper differs from zapper (p = 0.003); no other words are significantly different from each other (all p > 0.12). There is no significant interaction between word and phase ($\chi^2(6) = 9.08$, p = 0.17).

3. Post-vocalic /p/

Although the /p/ after the shortened vowel was not deliberately targeted for time manipulation, due to the variable nature of speech, the shortening intended for the vowel occasionally continued into the closure for /p/. For these analyses, we are including the

same set of participants that had adequate shortening for the vowel target, as participants that did not receive adequate shortening of the vowel tended to have shortened /p/ instead (i.e., shortening did not start until /p/ had started, leaving /æ/ the original duration). Overall, however, the magnitude of perturbation for the /p/ was small compared to the vowel: capper mean = -8.4 ms (SD = 9.5 ms); sapper mean = -10.9 ms (SD = 11.0 ms); zapper mean = -6.6 ms (SD = 8.0 ms).

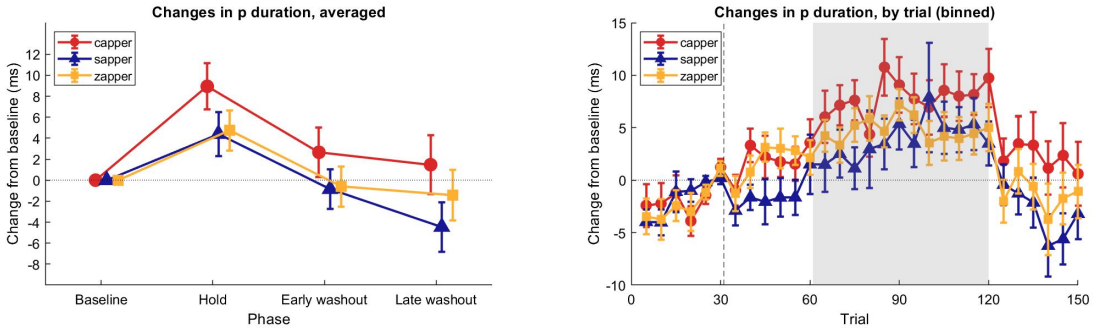

FIG. 5. Change from baseline in /p/ duration. Left: by phase, averaged across participants (means ± standard error). Right: behavior throughout the experiment, where each datapoint represents five trials, averaged across participants (means ± standard error).

Changes in /p/ closure duration are shown in Figure 5. Phase as a fixed effect significantly improves the model ($\chi^2(3) = 131.25$, $p < 0.0001$). The duration of /p/ in the hold phase (6.2 ± 1.0 ms) is significantly longer than all other phases ($p < 0.0001$ for all comparisons). The early washout phase (0.6 ± 1.0 ms) is also significantly longer than the late washout phase (-1.1 ± 1.0 ms, $p = 0.03$), but neither washout phase is longer than baseline (early washout: $p = 0.64$, late washout: $p = 0.07$).

The addition of word as a fixed effect also significantly improves the model ($\chi^2(2) = 44.30$, $p < 0.0001$), as does the addition of the interaction between phase and word ($\chi^2(6) = 16.44$, $p = 0.01$). For all words, the /p/ is significantly longer in the hold phase than in all other phases (all $p < 0.0005$). For capper and zapper, this is the only phase that is different, and /p/ duration returns to baseline during early washout (no significant difference between either washout phase and baseline, all $p > 0.1$). However, for sapper, the duration of /p/ continues to decrease through washout: late washout (-4.3 ± 1.2 ms) is significantly shorter than early washout (-0.7 ± 1.2 ms, $p = 0.004$) and baseline (0.1 ± 1.2 ms, $p = 0.0004$). The /p/ lengthens significantly more during the hold phase in capper (9.5 ± 1.3 ms) than either sapper (4.5 ± 1.2 ms, $p = 0.0001$) or zapper (5.2 ± 1.2 ms, $p = 0.0006$), though again the estimated change from baseline to hold is small. The apparent lack of after-effects is consistent with this change being purely compensatory rather than reflecting adaptation of feedforward/predictive control. Alternatively, this could potentially be due to the small magnitude of both perturbation and compensatory behavior (cf. the difference of 19.4 ms between hold and early washout for the vowel target).

4. Final syllable nucleus /ɚ/

The final /ɚ/was also not intentionally perturbed. As it was more distant from the targeted segments, it also was only infrequently affected by the shortening portion. The mean perturbation was under 1 ms for all words except capper, which had a mean shortening of 2.5 ms. These analyses include all participants, as insufficient perturbation for either the

consonant target or the vowel target does not consistently affect the perturbation of /ɚ/. Results for /ɚ/ are shown in Figure 6.

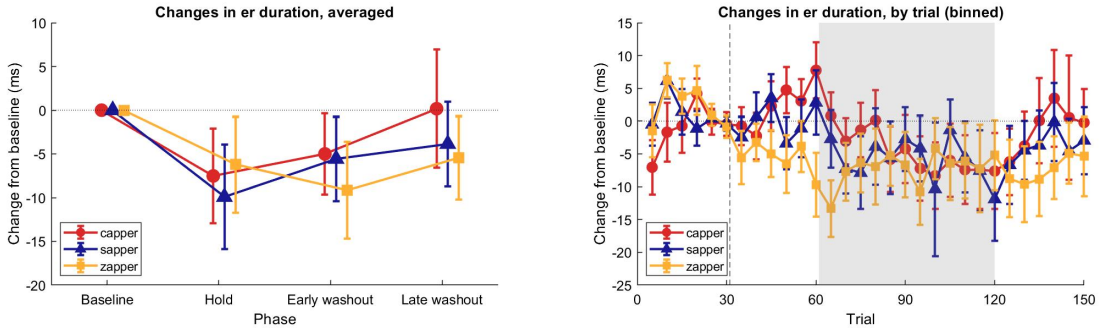

FIG. 6. Change from baseline in /ɚ/ duration. Left: by phase, averaged across participants (means ± standard error). Right: behavior throughout the experiment, where each datapoint represents five trials, averaged across participants (means ± standard error).

Adding phase as a fixed effect significantly improves the fit of the model ($\chi^2(3) = 29.26$, $p < 0.0001$). Participants shorten their productions of /ɚ/ during the hold phase compared to baseline ($p < 0.0001$); the magnitude of the difference is small, similar to the differences observed for /p/ (-7.5 ± 2.5 ms). The shortening persists through the early washout phase (-6.3 ± 2.5 ms), which is not significantly different from the hold phase ($p = 0.43$); the early washout phase is significantly different from the baseline ($p = 0.0002$). However, the shortening disappears by the late washout phase: there is a significant difference between the hold and late washout (-2.8 ± 2.5 ms) phases ($p = 0.01$), and the late washout phase is not significantly different from the baseline phase ($p = 0.12$). There is no significant difference between the early and late washout phases ($p = 0.08$).

Adding word as an additional fixed effect does not improve the fit of the model ($\chi^2(2) = 2.88$, $p = 0.23$), nor does the interaction between phase and word ($\chi^2(6) = 6.56$, $p = 0.36$). Thus, for all words there is some general shortening of /əˠ/ during the hold phase, and a return to baseline by the end of the washout phase. Given that there was minimal perturbation of this segment, the shortened duration in the hold phase does not seem indicative of direct adaptation for the applied perturbation. However, the lingering aftereffects in early washout indicate that it may be part of a larger adaptive strategy for the word overall.

F. Participant awareness

Overall, participants did not realize that the timing of their speech was being perturbed. Participants first indicated via multiple choice if they thought they were in a group that received true feedback or a group that received manipulated feedback. If they thought they received true feedback, they were informed that everybody was actually in the manipulated feedback group. If they thought they received manipulated feedback, they then indicated when they had realized (early in the experiment, late in the experiment, only now that the experimenter is bringing it up). Participants then described what they thought the manipulation was.

Nine participants reported that they thought that they had received true feedback. One additional participant only thought something was different when they were asked if they thought they had received true feedback. Of the 10 people that thought they had received

manipulated feedback, five reported noticing early in the experiment and five reported noticing late in the experiment.

Five participants referred to time when guessing what the manipulation was: one thought their timing felt cut off shorter than normal; one thought that the feedback was slightly slower than how they had actually said it; one thought they sometimes heard what they were saying before they said it; one thought either volume or rate may have been different; and one thought that some of the vowels were held longer and more emphasized than they had said them. The remaining participants had varied impressions: four thought their pitch was being manipulated; three felt that their speech sounded generally different than expected; four indicated some mechanical issue or roboticness; and four could not guess what was manipulated at all.

To check for effects of participant awareness on compensation magnitude, a binary variable for awareness was added to the maximal models for both consonant and vowel target. The five participants that mentioned some temporal variable were coded as aware, and the remaining 15 were coded as not aware. The addition of awareness as a fixed effect does not significantly improve the fit of the models examining absolute change in duration for either the consonant target ($\chi^2(1) = 0.14$, $p = 0.71$) or the vowel target ($\chi^2(1) = 0.07$, $p = 0.79$). The interaction between awareness and phase also does not significantly improve the model for vowel target ($\chi^2(0.82) = 3$, $p = 0.85$), but does for consonant target ($\chi^2(3) = 13.48$, $p = 0.004$). This interaction is driven by participants that were aware of temporal perturbation producing longer consonant targets in the hold phase than during the baseline phase (5.3 ± 1.8 ms, $p = 0.02$) or the late washout phase (8.3 ± 1.8 ms, $p < 0.0001$), indicating

increased following responses. Participants that were not aware of temporal perturbations produced no significant differences between any phases (all $p = 1.00$). Furthermore, there is no difference between aware and unaware participants within each phase (all $p \geq 0.16$). This indicates that participant awareness did not increase the likelihood or magnitude of adaptive responses to oppose the perturbation.

IV. DISCUSSION

In this study, we examined how speakers adapt to auditory feedback perturbations in the temporal domain. Contrary to Mitsuya *et al.* (2014), but consistent with Oschkinat and Hoole (2020), we found that participants did not adapt to lengthened consonant durations in syllable onset position by shortening their productions in absolute time; rather, there was variable shortening and lengthening both across and within participants, and across and within segments. However, we did find that participants consistently responded to a shortened vowel by lengthening their vowel productions. This is consistent with Mitsuya *et al.* (2014) but was not observed in Oschkinat and Hoole (2020). However, we did observe adaptation in the relative duration of both VOT and fricative duration when measured as a proportion of overall syllable duration.

We observed no effects of timing type or phonemic boundaries. We did not observe any consistent differences in adaptation between VOT, reflecting relative timing between two motor events, and fricative duration, reflecting the duration of a single motor action. This was true for measurements in both absolute and relative time. This suggests that both types of timing may be monitored and controlled in a similar manner. Similarly, we found

few consistent differences between /s/ and /z/, despite the fact that lengthening /s/ and no impact on its phonemic status while lengthening /z/ makes it more like /s/ (Baum and Blumstein, 1987; Bjorndahl, 2018; Jongman, 1989). However, these negative results are tempered by the lack of an overall change in absolute duration of consonant targets.

It is unclear why the perturbations introduced in this study led to adaptation in absolute duration in the vowels but not the consonant targets. We see several potential explanations for this. One possibility is that the lengthening in the vowels was not in fact adaptation at all, but rather solely an effect of delayed auditory feedback (DAF). Auditory feedback is delayed in multiple ways in this study. The measured latency for our experimental setup with the applied temporal perturbation is 35 ms (cf. Kim *et al.* 2020). In addition, when the consonant is lengthened, the vowel onset is also delayed relative to production by the amount of consonant lengthening—about 40 ms in the three words used in the analysis. A similar delay under DAF leads to prolongation of vowels (Kalveram 1984 as cited in Kalveram and Jäncke 1989). Kalveram and Jäncke 1989 report a similar but slightly smaller magnitude of lengthening in healthy controls, 13.82 ms of lengthening for stressed vowels with 40 ms of feedback delay (compare to 16.0 ms for capper, 27.6 ms for sapper and 29.2 ms for zapper, with approximately 75 ms of total delay—software, hardware, and perturbation-related—in the current study). In addition, they report no lengthening in an unstressed vowel preceding the stressed syllable (1.80 ms), and a smaller lengthening effect in an unstressed syllable following the stressed syllable (7.59 ms).

However, adjustment for DAF is not a fully satisfactory explanation for the results in this study. First, DAF studies delay large portions of the signal, rather than delaying and

shortening the stressed vowel, as was the case in this study—thus, it is unclear if these vowel prolongation effects would even occur in this kind of "selective" auditory feedback delay. Second, participants in this study overall shortened the unstressed /ɚ/, rather than lengthening it as reported by Kalveram and Jäncke 1989. Furthermore, there are mixed results in the speed at which participants return to baseline—Yates 1963 notes that some studies report immediate return to baseline speech, while others report a more gradual return. In this study, the behavior of the vowel is different from the behavior of /p/, which did not show aftereffects despite increasing in duration during the hold phase. For sapper and zapper, vowels in the early washout phase were still significantly longer than baseline, and only returned to baseline levels during the late washout phase. These aftereffects in the washout phase are consistent with previous studies of sensorimotor learning in both temporal and spectral aspects of speech (Houde and Jordan, 1998; Mitsuya *et al.*, 2014; Oschkinat and Hoole, 2020; Villacorta *et al.*, 2007), and suggest that participants did adapt their temporal control of vowel duration. However, the rapid decrease in vowel duration from the hold phase to the early washout suggests that a large portion of the duration increase in the hold phase was likely caused by more general online compensation for delayed vowel onset, or, alternatively, as compensation after perceiving a lengthened consonant.

A second possibility is that shortening is more salient than lengthening and provokes a stronger adaptive response. This may have been compounded by shortening a stressed vowel, as duration is a strong cue for stress in English. For sapper and zapper, participants lengthened their vowels by nearly 100% of the perturbation (27.9 ms compensation for sapper in response to a perturbation of -32.8 ms, and 29.0 ms compensation in response to a

perturbation of -34.9 ms for zapper), which could suggest that there was a duration target that participants were attempting to reach. While previous work did show VOT shortening in response to artificially lengthened VOT (Mitsuya et al. 2014), the size of this effect was quite small (3.6 ms, compared to mean 34.9 ms perturbation) in comparison with the lengthening effects observed in that study, both in response to artificially shortened VOT (10.3 ms) and artificially shortened vowels (22.1 ms, compared to mean 16.7 ms perturbation), as well as in comparison with the lengthening effects observed in the current study. Mitsuya *et al.* (2014) also report that lengthened vowels did not lead to compensatory changes in vowel length. Similarly, in the coda condition in Oschkinat and Hoole (2020) where adaptation was seen, participants showed more opposing adaptation to coda shortening (34 ms / 17.2% of the perturbation) than for vowel lengthening (9 ms / 10.3% of the perturbation). These results suggest the sensorimotor system be driven to maintain a certain minimal duration rather than a specific overall duration.

It is also possible that speakers do not attend to precise durations of individual segments, but rather attend to segments' durations relative to other segments in some higher planning unit (Fowler, 1981; Munhall *et al.*, 1992, 1994). This is consistent with our results analyzing consonant duration as a proportion of the syllable, which revealed significant adaptation which was not apparent in absolute time. For example, although some participants increased the duration of their consonants, the overall proportional duration of their consonant still shifted downward to oppose the perturbation, likely through even greater increases in vowel duration. If speakers attend to proportional timing, lengthening the vowel target (or the entire rime) would have the same general compensatory effect as shortening the consonant

target—with both strategies, the proportion of syllable occupied by the consonant will decrease. As for why speakers might prefer to lengthen the vowel rather than shorten the consonant target, there is some evidence that consonants in onset position temporally less flexible than vowels. First, vowels change more in duration under different speech rate conditions than consonants (Gay, 1978, 1981; Guenther, 1995; Volaitis and Miller, 1992); it thus may be easier to maintain desired proportional durations by altering the vowel target rather than the consonant target. In addition, (Oschkinat and Hoole, 2020) found that that the segments in syllable rimes are more responsive to timing adjustments than the segments in syllable onsets. Although the present study does not provide a thorough test of this hypothesis, this analysis would also account for the change in /p/ duration reported in this study.

Although the adaption observed in this proportional durational analysis may appear very small (-1.6 ± 0.3%), it is important to note that even if a participant shortened their consonant by the full perturbation amount and also lengthened their vowel by the full perturbation amount, the difference in proportion would still be small. As an example, consider our original sapper example. If the baseline production had an [s] duration of 150 ms, an [æ] duration of 200 ms, and a [p] duration of 50 ms, the proportion of [s] is 38%; with full adaptation in both consonant (-42 ms) and vowel (+34 ms), the [s] proportion would only be 28%, or a 10 percentage point difference. Adaptation for auditory perturbations of speech is never complete, and typically ranges around 20-50% of the perturbation (e.g., Cai *et al.* 2010; Houde and Jordan 2002; MacDonald *et al.* 2011; Mitsuya *et al.* 2013; Munhall *et al.* 2009; Villacorta *et al.* 2007). The adaptation seen in our study (1.6%) is roughly 16%

of the maximal possible value (10%), only slightly lower than the amount of adaptation typically seen for spectral perturbations.

An additional point in favor of the relevance of proportional duration is that participants produced slightly shorter /ɚ/ vowels during the hold phase, despite minimal to no perturbation of this segment. In this case, the relevant proportion is not within a single syllable, but rather across syllables, where vowels in stressed syllables are longer in English than unstressed vowels. During the hold phase, the stressed vowel /æ/ was shortened, while the unstressed vowel /ɚ/ remained unaltered. This would shift the duration ratio between the stressed and unstressed syllables. In this case, shortening the vowel in the unstressed syllable would help to preserve the baseline duration ratio between the two syllables.

The final possibility we see for the differences in adaptation for the consonant and vowel perturbations is that changes in vowel durations cause changes in the durations of other segments in the syllable. That is, it is possible that there was some attempt to shorten consonants, but this was overpowered by the vowel lengthening. Since the target consonants and the target vowel were in the same syllable, it is possible that it was simply difficult to entirely re-time the syllable such that the consonant onset was shorter and the vowel nucleus was longer. This issue could also affect the degree to which participants are able to decrease the proportion of consonant: one strategy to lengthen vowels could be to slow down the entire syllable, which would increase the duration of both the consonant and the vowel. This overall lengthening would then attenuate any decrease in consonant duration.

Although there have been some attempts to explicitly model time in speech, how duration is monitored, controlled, and altered is highly underspecified in most models. For example,

speech sounds (syllables) in the DIVA model (Guenther, 1995; Tourville and Guenther, 2011) have a fixed duration, with time-varying trajectories that produce the desired formants and articulatory positions; this model explicitly incorporates the monitoring of spatial information for both compensation and adaptation, but not of temporal information. Alternatively, the Task Dynamics model (Saltzman, 1986; Saltzman and Byrd, 2000; Saltzman and Munhall, 1989) expresses time through a system of planning oscillators that activate and deactivate articulatory gestures. Although π gestures (Byrd and Saltzman, 2003) have been invoked to lengthen articulatory gestures, typically at prosodic boundaries, through local changes in speech rate (the time course of gestural evolution) there has been little work to address whether and how segment durations can be flexibly controlled. In particular, neither the DIVA nor the Task Dynamics model account for altering the dynamics of a syllable with the explicit goal of a particular segment occupying a greater or lesser proportion of the syllable. The changes in proportional duration observed in this study indicate that current models should be revisited to address temporal control more explicitly, potentially incorporating domain-general, phonology-extrinsic timing mechanisms, as suggested by Turk and Shattuck-Hufnagel (2020).

V. CONCLUSION

Overall, this study provides support for the hypothesis that temporal information in the auditory feedback signal is actively monitored and used to update future speech production. We observed increases in vowel duration for perturbations that shortened vowels. These increases were likely the result of both online compensation for feedback delays and durational

adaptation, as the duration remained elevated after the removal of the perturbation in the early washout phase. Although we did not observe changes in absolute consonant duration when VOT or fricative duration was lengthened, there was a reduction in the duration of these segments as a proportion of the overall syllable. As for vowels, this change was visible in both the hold and early washout phases, consistent with adaptive learning. This result suggests that relative duration between segments may be more important for temporal control than absolute duration, consistent with previous theoretical suggestions (Boucher, 2002; Kessinger and Blumstein, 1998). While we observed changes for perturbations of both vowel and consonant duration, the effects were much larger (up to 100% of the perturbation) for perturbations of vowel duration. We have suggested possible causes for this differential effect, and future studies that control for the potential confounding variables discussed above are necessary to resolve these issues (e.g. a study that contrasts shortening vs. lengthening within consonants or vowels). Lastly, we observed no difference in adaption between speech events hypothesized to reflect relative (VOT) and absolute (fricative duration) timing or between perturbations that pushed a segment towards (lengthening for /z/) or away from (lengthening for /s/) a phonemic category boundary. However, these results are tempered by the relatively small amount of compensation seen for consonants, as well as the finding of compensation only for proportional, and not absolute, duration. Again, future work that drives a more robust adaptive response could address this issue more thoroughly.

ACKNOWLEDGMENTS

This work supported by NIH grant R01 DC017091.

Bates, D., Maechler, M., Bolker, B., Walker, S. *et al.* (**2014**). "lme4: Linear mixed-effects models using Eigen and s4," R package version **1**(7), 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," The Journal of the Acoustical Society of America **82**(3), 1073–1077.

Behroozmand, R., and Larson, C. R. (**2011**). "Error-dependent modulation of speech-induced auditory suppression for pitch-shifted voice feedback," BMC neuroscience **12**(1), 54.

Bjorndahl, C. (**2018**). "(manuscript) A story of /v/: voiced spirants in the obstruent-sonorant divide," Ph.D. thesis, Cornell University.

Boucher, V. J. (**2002**). "Timing relations in speech and the identification of voice-onset times: A stable perceptual boundary for voicing categories across speaking rates," Perception & Psychophysics **64**(1), 121–130.

Browman, C. P., and Goldstein, L. M. (**1986**). "Towards an articulatory phonology," Phonology **3**, 219–252.

Burnett, T. A., Freedland, M. B., Larson, C. R., and Hain, T. C. (**1998**). "Voice F0 responses to manipulations in pitch feedback," The Journal of the Acoustical Society of America **103**(6), 3153–3161.

Byrd, D., and Saltzman, E. (**2003**). "The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening," Journal of Phonetics **31**(2), 149–180.

Cai, S. (**2014**). *A manual of Audapter*, Speech Laboratory, Department of Speech, Language, and Hearing Sciences, Boston University, 2.1.012 ed.

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 /iau/," Proceedings of the 8th ISSP 65–68.

Cai, S., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (**2010**). "Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong /iau/and its pattern of generalization," The Journal of the Acoustical Society of America **128**(4), 2033–2048.

Cai, S., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (**2011**). "Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing," Journal of Neuroscience **31**(45), 16483–16490.

Casserly, E. D. (**2011**). "Speaker compensation for local perturbation of fricative acoustic feedback," The Journal of the Acoustical Society of America **129**(4), 2181–2190.

Daliri, A., and Dittman, J. (**2019**). "Successful auditory motor adaptation requires task-relevant auditory errors," Journal of Neurophysiology **122**(2), 552–562.

Denes, P. (**1955**). "Effect of duration on the perception of voicing," The Journal of the Acoustical Society of America **27**(4), 761–764.

Elman, J. L. (**1981**). "Effects of frequency-shifted feedback on the pitch of vocal productions," The Journal of the Acoustical Society of America **70**(1), 45–50.

753 Elzinga, D., and Eddington, D. (**2014**). "An experimental approach to ambisyllabicity in
754 English," Topics in Linguistics **14**(1), 34–47.
755 Floegel, M., Fuchs, S., and Kell, C. A. (**2020**). "Differential contributions of the two cerebral
756 hemispheres to temporal and spectral speech feedback control," Nature Communications
757 **11**(1), 1–12.
758 Fowler, C. A. (**1981**). "A relationship between coarticulation and compensatory shortening,"
759 Phonetica **38**(1-3), 35–50.
760 Gafos, A. I. (**2002**). "A grammar of gestural coordination," Natural Language & Linguistic
761 Theory **20**(2), 269–337.
762 Gay, T. (**1978**). "Effect of speaking rate on vowel formant movements," The Journal of the
763 Acoustical Society of America **63**(1), 223–230.
764 Gay, T. (**1981**). "Mechanisms in the control of speech rate," Phonetica **38**(1-3), 148–158.
765 Grube, M., Cooper, F. E., Chinnery, P. F., and Griffiths, T. D. (**2010**). "Dissociation of
766 duration-based and beat-based auditory timing in cerebellar degeneration," Proceedings
767 of the National Academy of Sciences **107**(25), 11597–11601.
768 Guenther, F. H. (**1995**). "Speech sound acquisition, coarticulation, and rate effects in a
769 neural network model of speech production.," Psychological review **102**(3), 594.
770 Houde, J. F., and Jordan, M. I. (**1998**). "Sensorimotor adaptation in speech production,"
771 Science **279**(5354), 1213–1216.
772 Houde, J. F., and Jordan, M. I. (**2002**). "Sensorimotor adaptation of speech I: Compensation
773 and adaptation," Journal of Speech, Language, and Hearing Research **45**, 295–310.

774 Ivry, R. B., Spencer, R. M., Zelaznik, H. N., and Diedrichsen, J. (**2002**). "The cerebellum
775 and event timing," Annals of the new York Academy of Sciences **978**(1), 302–317.
776 Jones, J. A., and Munhall, K. G. (**2000**). "Perceptual calibration of F0 production: Evidence
777 from feedback perturbation," The Journal of the Acoustical Society of America **108**(3),
778 1246–1251.
779 Jongman, A. (**1989**). "Duration of frication noise required for identification of English frica-
780 tives," The Journal of the Acoustical Society of America **85**(4), 1718–1725.
781 Kalveram, K.-T. (**1984**). "Geschlechterunterschiede bei der audio-motorischen Kontrolle
782 der Phonation (sex differences in the audio-motor control of phonations)," Zeitschrift für
783 experimentelle und angewandte Psychologie **31**(1), 39–47.
784 Kalveram, K. T., and Jäncke, L. (**1989**). "Vowel duration and voice onset time for stressed
785 and nonstressed syllables in stutterers under delayed auditory feedback condition," Folia
786 Phoniatrica **41**(1), 30–42.
787 Kessinger, R. H., and Blumstein, S. E. (**1998**). "Effects of speaking rate on voice-onset time
788 and vowel production: Some implications for perception studies," Journal of Phonetics
789 **26**(2), 117–128.
790 Kim, K. S., Wang, H., and Max, L. (**2020**). "It's about time: minimizing hardware and
791 software latencies in speech research with real-time auditory feedback," Journal of Speech,
792 Language, and Hearing Research **63**(8), 2522–2534.
793 Klein, E., Brunner, J., and Hoole, P. (**2019**). "The relevance of auditory feedback for con-
794 sonant production: The case of fricatives," Journal of Phonetics **77**, 100931.

795 Korzyukov, O., Bronder, A., Lee, Y., Patel, S., and Larson, C. R. (**2017**). "Bioelectrical
796 brain effects of one's own voice identification in pitch of voice auditory feedback," Neu-
797 ropsychologia **101**, 106–114.

798 Lametti, D. R., Smith, H. J., Watkins, K. E., and Shiller, D. M. (**2018**). "Robust sensori-
799 motor learning during variable sentence-level speech," Current Biology **28**(19), 3106–3113.

800 Lenth, R. (**2019**). *emmeans: Estimated Marginal Means, aka Least-Squares Means*, https:
801 //CRAN.R-project.org/package=emmeans, r package version 1.3.5.1.

802 Lisker, L. (**1986**). ""Voicing" in English: A catalogue of acoustic features signaling /b/
803 versus /p/ in trochees," Language and speech **29**(1), 3–11.

804 Liu, H., Behroozmand, R., and Larson, C. R. (**2010**). "Enhanced neural responses to self-
805 triggered voice pitch feedback perturbations," Neuroreport **21**(7), 527.

806 MacDonald, E. N., Purcell, D. W., and Munhall, K. G. (**2011**). "Probing the independence
807 of formant control using altered auditory feedback," The Journal of the Acoustical Society
808 of America **129**(2), 955–965.

809 Max, L., and Maffett, D. G. (**2015**). "Feedback delays eliminate auditory-motor learning in
810 speech production," Neuroscience letters **591**, 25–29.

811 Mitsuya, T., MacDonald, E. N., and Munhall, K. G. (**2014**). "Temporal control and compen-
812 sation for perturbed voicing feedback," The Journal of the Acoustical Society of America
813 **135**(5), 2986–2994.

814 Mitsuya, T., Munhall, K. G., and Purcell, D. W. (**2017**). "Modulation of auditory-motor
815 learning in response to formant perturbation as a function of delayed auditory feedback,"
816 The Journal of the Acoustical Society of America **141**(4), 2758–2767.

817 Mitsuya, T., Samson, F., Ménard, L., and Munhall, K. G. (**2013**). "Language dependent
818 vowel representation in speech production," The Journal of the Acoustical Society of Amer-
819 ica **133**(5), 2993–3003.
820 Mochida, T., Gomi, H., and Kashino, M. (**2010**). "Rapid change in articulatory lip move-
821 ment induced by preceding auditory feedback during production of bilabial plosives," PLoS
822 One **5**(11), e13866.
823 Munhall, K., Fowler, C. H., Hawkins, S., and Saltzman, E. (**1992**). ""compensatory short-
824 ening" in monosyllables of spoken English.," Journal of Phonetics **20**, 225–239.
825 Munhall, K. G., Löfqvist, A., and Kelso, J. S. (**1994**). "Lip-larynx coordination in speech:
826 Effects of mechanical perturbations to the lower lip," The Journal of the Acoustical Society
827 of America **95**(6), 3605–3616.
828 Munhall, K. G., MacDonald, E. N., Byrne, S. K., and Johnsrude, I. (**2009**). "Talkers alter
829 vowel production in response to real-time formant perturbation even when instructed not
830 to compensate," The Journal of the Acoustical Society of America **125**(1), 384–390.
831 Niziolek, C. A., and Guenther, F. H. (**2013**). "Vowel category boundaries enhance cortical
832 and behavioral responses to speech feedback alterations," Journal of Neuroscience **33**(29),
833 12090–12098.
834 Ogane, R., and Honda, M. (**2014**). "Speech compensation for time-scale-modified auditory
835 feedback," Journal of Speech, Language, and Hearing Research **57**(2), S616–S625.
836 Oschkinat, M., and Hoole, P. (**2020**). "Compensation to real-time temporal auditory feed-
837 back perturbation depends on syllable position," The Journal of the Acoustical Society of
838 America **148**(3), 1478–1495.

839 Patel, R., Reilly, K. J., Archibald, E., Cai, S., and Guenther, F. H. (**2015**). "Responses to
840 intensity-shifted auditory feedback during running speech," Journal of Speech, Language,
841 and Hearing Research **58**(6), 1687–1694.

842 Port, R. F., and Dalby, J. (**1982**). "Consonant/vowel ratio as a cue for voicing in English,"
843 Perception & Psychophysics **32**(2), 141–152.

844 Purcell, D. W., and Munhall, K. G. (**2006**a). "Adaptive control of vowel formant frequency:
845 Evidence from real-time formant manipulation," The Journal of the Acoustical Society of
846 America **120**(2), 966–977.

847 Purcell, D. W., and Munhall, K. G. (**2006**b). "Compensation following real-time manipu-
848 lation of formants in isolated vowels," The Journal of the Acoustical Society of America
849 **119**(4), 2288–2297.

850 R Core Team (**2019**). *R: A Language and Environment for Statistical Computing*, R Foun-
851 dation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

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

854 Saltzman, E. (**1986**). "Task dynamic coordination of the speech articulators: A preliminary
855 model," Status Report on Speech Research 9.

856 Saltzman, E., and Byrd, D. (**2000**). "Task-dynamics of gestural timing: Phase windows and
857 multifrequency rhythms," Human Movement Science **19**(4), 499–526.

858 Saltzman, E. L., and Munhall, K. G. (**1989**). "A dynamical approach to gestural patterning
859 in speech production," Ecological psychology **1**(4), 333–382.

860 Shiller, D., Mitsuya, T., and Max, L. (**2020**). "Prior short-term habituation to auditory
861 feedback delays does not mitigate their disruptive effect on speech auditory-motor adap-
862 tation," Neuroscience **446**, 213–224.

863 Shiller, D. M., Sato, M., Gracco, V. L., and Baum, S. R. (**2009**). "Perceptual recalibration
864 of speech sounds following speech motor learning," The Journal of the Acoustical Society
865 of America **125**(2), 1103–1113.

866 Stuart, A., and Kalinowski, J. (**2015**). "Effect of delayed auditory feedback, speech rate,
867 and sex on speech production," Perceptual and motor skills **120**(3), 747–765.

868 Teki, S., Grube, M., Kumar, S., and Griffiths, T. D. (**2011**). "Distinct neural substrates of
869 duration-based and beat-based auditory timing," Journal of Neuroscience **31**(10), 3805–
870 3812.

871 Tourville, J. A., Cai, S., and Guenther, F. (**2013**). "Exploring auditory-motor interactions in
872 normal and disordered speech," in *Proceedings of Meetings on Acoustics ICA2013*, Acous-
873 tical Society of America, Vol. 19, p. 060180.

874 Tourville, J. A., and Guenther, F. H. (**2011**). "The diva model: A neural theory of speech
875 acquisition and production," Language and cognitive processes **26**(7), 952–981.

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

878 Turk, A., and Shattuck-Hufnagel, S. (**2020**). "Timing evidence for symbolic phonological
879 representations and phonology-extrinsic timing in speech production," Frontiers in Psy-
880 chology **10**, 2952.

881 Villacorta, V. M., Perkell, J. S., and Guenther, F. H. (**2007**). "Sensorimotor adaptation to
882 feedback perturbations of vowel acoustics and its relation to perception," The Journal of
883 the Acoustical Society of America **122**(4), 2306–2319.
884 Volaitis, L. E., and Miller, J. L. (**1992**). "Phonetic prototypes: Influence of place of artic-
885 ulation and speaking rate on the internal structure of voicing categories," The Journal of
886 the Acoustical Society of America **92**(2), 723–735.
887 Wei, K., and Körding, K. (**2009**). "Relevance of error: what drives motor adaptation?,"
888 Journal of neurophysiology **101**(2), 655–664.
889 Yates, A. J. (**1963**). "Delayed auditory feedback.," Psychological bulletin **60**(3), 213.
890 Zelaznik, H. N., Spencer, R. M., Ivry, R. B., Baria, A., Bloom, M., Dolansky, L., Justice, S.,
891 Patterson, K., and Whetter, E. (**2005**). "Timing variability in circle drawing and tapping:
892 probing the relationship between event and emergent timing," Journal of motor behavior
893 **37**(5), 395–403.

894 APPENDIX A: FIGURES SHOWING INDIVIDUAL PARTICIPANT DATA

895 Figures displaying data by individual participant. Asterisks mark participants that indi-
896 cated that they thought there was a temporal perturbation of any sort.

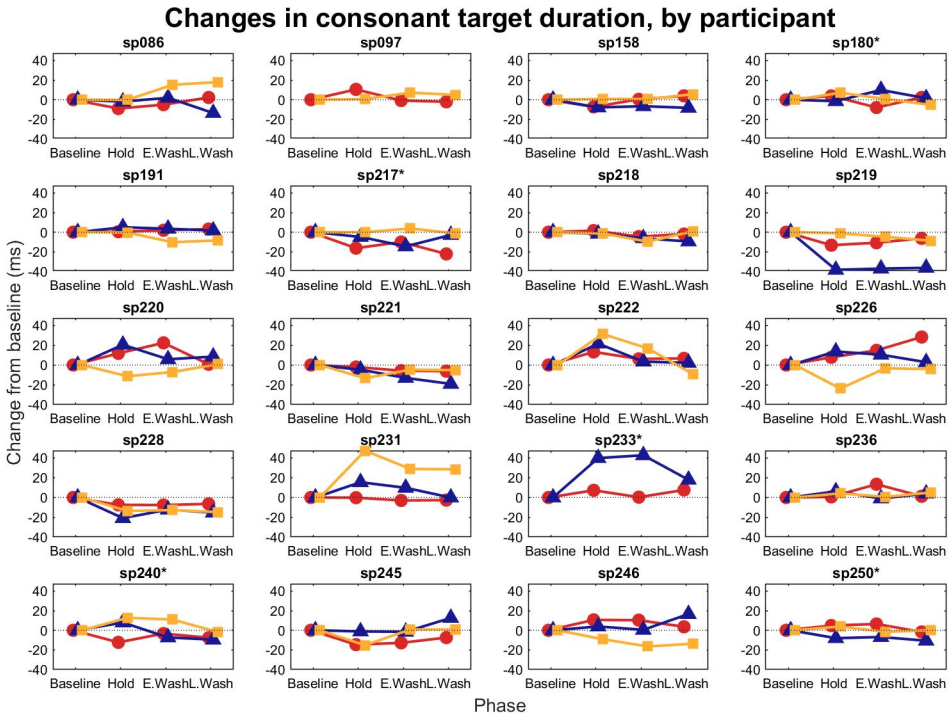
Changes in consonant target duration, by participant
sp086
sp097
sp158
sp180*
sp191
sp217*
sp218
sp219
sp220
sp221
sp222
sp226
sp228
sp231
sp233*
sp236
sp240*
sp245
sp246
sp250*
Baseline Hold E.Wash L.Wash
40 20 0 -20 -40
Change from baseline (ms)
Phase

FIG. 7.

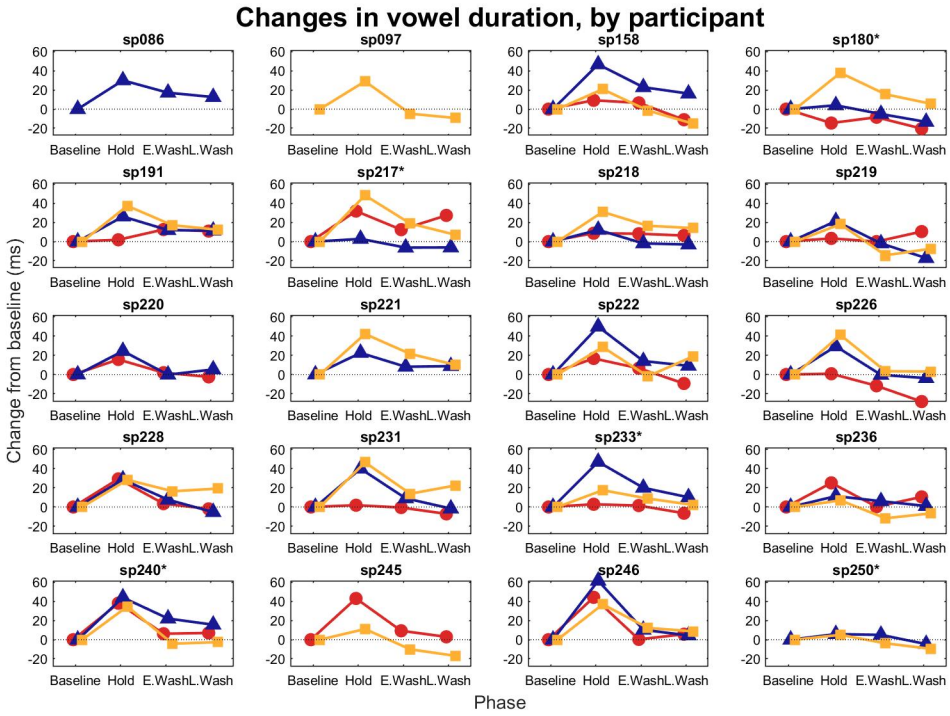
Changes in vowel duration, by participant
sp086
sp097
sp158
sp180*
sp191
sp217*
sp218
sp219
sp220
sp221
sp222
sp226
sp228
sp231
sp233*
sp236
sp240*
sp245
sp246
sp250*
60
40
20
0
-20
Baseline Hold E.Wash L.Wash
Change from baseline (ms)
Phase

FIG. 8.

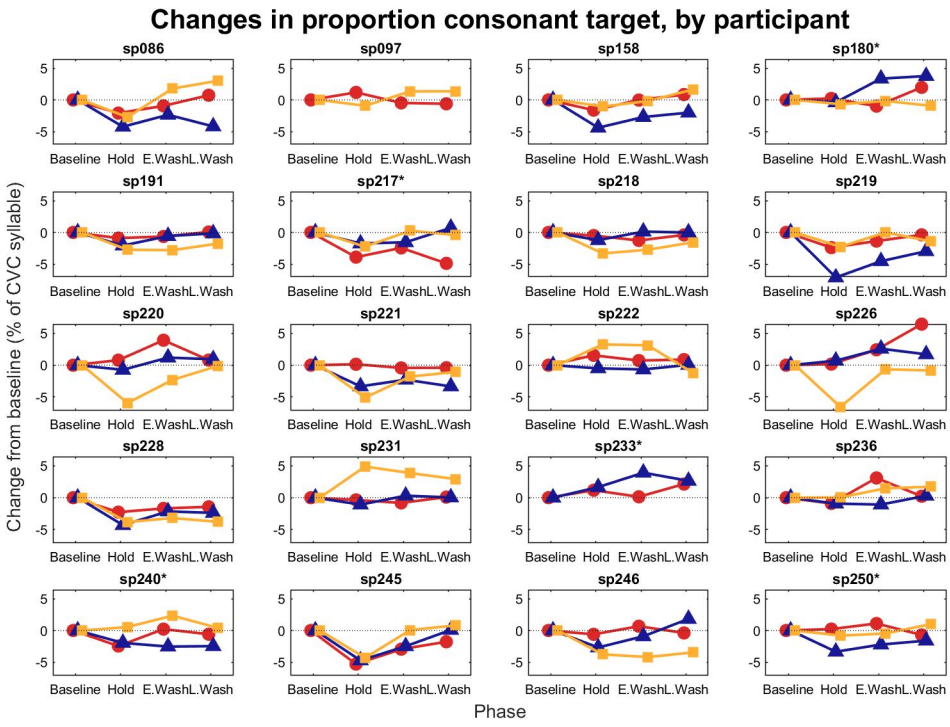
Changes in proportion consonant target, by participant
sp086
sp097
sp158
sp180*
sp191
sp217*
sp218
sp219
sp220
sp221
sp222
sp226
sp228
sp231
sp233*
sp236
sp240*
sp245
sp246
sp250*
Baseline Hold E.Wash L.Wash
5
0
-5
Change from baseline (% of CVC syllable)
Phase

FIG. 9.

APPENDIX B: OST AND PCF DETAILS

Default OST (Online Status Tracking) and PCF (Perturbation ConFiguration) settings for "a sapper":

1. OST

The lines in OST files are composed of:

1. The initial status. All trials start at 0 and advance according to the heuristics used.
2. The type of heuristic that will be used to advance to the next status. E.g. a heuristic that looks for the RMS intensity to increase to a certain threshold and stay there for a given amount of time is "INTENSITY_RISE_HOLD".
3. The first parameter for that heuristic. For INTENSITY_RISE_HOLD it is the threshold to cross
4. The second parameter for that heuristic. For INTENSITY_RISE_HOLD it is the duration that it needs to stay above the first parameter (in seconds).
5. Optional third parameter, currently blank for all heuristics.

0 INTENSITY_RISE_HOLD 0.012 0.010 {}

Start at status 0. To move to status 2 (onset of V1), looks for RMS intensity that remains above the threshold for 10 ms.

2 INTENSITY_RATIO_RISE 0.250 0.002 {}

Achieved status 2. To move to status 4 (onset of /s/), looks for loud high frequency noise

that lasts 2 ms.

4 INTENSITY_RATIO_FALL_HOLD 0.400 0.010 {}

Achieved status 4. To move to status 6 (onset of /æ/), looks for decrease in high frequency noise from /s/, needs to be below threshold for 10 ms.

6 NEG_INTENSITY_SLOPE_STRETCH_SPAN 5.000 -1.000 {}

Achieved status 6. To move to status 8 (onset of /p/ closure), looks for stretch of decreasing intensity that lasts at least 5 frames and where the sum of the decreases is at least -1.

8 INTENSITY_RISE_HOLD_POS_SLOPE 0.015 0.010 {}

Achieved status 8. To move to status 10 (onset of /ɚ/), looks for increase in RMS intensity that crosses a threshold of 0.015 and stays above for 10 ms.

10 INTENSITY_FALL 0.005 0.010 {}

Achieved status 10. To move to status 12 (end of /ɚ/), looks for a fall in RMS intensity that crosses 0.005 and remains below that threshold for 10 ms.

12 OST_END NaN NaN {}

Achieved status 12. No further statuses to be tracked.

2. PCF

The time warping in this study was accomplished with a single time warping event. The functionality used in the PCF file was a single line:

4, 0.00, 0.25, **0.080**, 0.100, 1.50

The six components are:

1. ostStat_initial: the OST status that triggers time warping. Note that status 4 in the PCF corresponds with status 4 in the OST, which was the detection of the high frequency noise for /s/.
2. tBegin: an amount of time (in s) to wait after achieving the OST status in item 1 before initiating slow-down
3. rate1: the time dilation component. I.e. 0.25 is one quarter as fast as original time. For gapper only, this was the component changed to lengthen VOT; in order to produce more lengthening, rate1 decreased. This was because there was not reliably enough positive VOT to produce sufficient lengthening at a rate of 0.25.
4. dur1: the overall duration of the slowed down portion. Maximum perturbation was set at 0.080 s, where at 0.25 for rate1 there would be 60 ms of lengthening (20 ms slowed to 80 ms = 60 ms difference). For capper, sapper, and zapper, this was the component that was changed to produce the different amounts of perturbation, while rate1 remained constant.
5. durHold: how long to wait (in s) before speeding up playback to catch up with real time.
6. rate2: time compression component. E.g. 1.5 is 1.5 times as fast as original time. Playback goes at this rate until warped time matches real time.

APPENDIX C: NORMALIZED DATA

Model tables for normalized data. Models are compared to the model above using LRTs.

1. Consonant target models

Model	**df**	χ^2	**p**
1 + (1|Part)			
Phase + (1|Part)	3	10.24	0.02*
Phase + Word + (1|Part)	2	3.28	0.19
Phase + Word + Phase:Word + (1|Part)	6	10.93	0.09

Phase	Est. Mean	SE
Baseline	0.3%	3.6%
Hold	3.7%	3.6%
Early washout	-0.1%	3.6%
Late washout	-2.9%	3.6%

All comparisons n.s. $p \geq 0.32$ except hold and late washout, $p = 0.009$.

965 **2. Vowel target models**

Model	**df**	χ^2	**p**
1 + (1|Part)			
966 Phase + (1|Part)	3	739.65	< 0.0001***
Phase + Word + (1|Part)	2	331.40	< 0.0001***
Phase + Word + Phase:Word + (1|Part)	6	51.78	< 0.0001***

Phase	Est. Mean	SE
Baseline	0.9%	4.1%
967 Hold	74.8%	4.1%
Early washout	16.6%	4.1%
Late washout	5.3%	4.1%

968 All comparisons significant at $p \leq 0.0001$ except baseline and late washout, n.s. $p = 0.12$.