SPEAKERS COMPENSATE FOR TEMPORAL PERTURBATIONS IN AUDITORY FEEDBACK REGARDLESS OF SYLLABLE STRUCTURE

Robin Karlin[1], Chris Naber[1], & Benjamin Parrell[1,2]

[1]Waisman Center, University of Wisconsin – Madison; [2]Communication Sciences and Disorders, University of Wisconsin – Madison
rkarlin@wisc.edu; cwnaber@wisc.edu; bparrell@wisc.edu

ABSTRACT

It is known that speakers monitor sensory feedback during speech to maintain accurate production. However, it is unclear the extent to which speakers use feedback for timing speech gestures, and how that may interact with phonological structure.

We present an altered auditory feedback study that accelerates and decelerates the formant transition in /aɪ/, testing the hypothesis that speakers use feedforward control to time gestures within a syllable, and auditory feedback to initiate the next syllable. Results suggest that speakers used directionally specific temporal information from auditory feedback: they shortened the perturbed vowel when transitions were accelerated, and slowed down when transitions were decelerated. Shortening was restricted to perturbed segments, while slowing occurred on both perturbed segments and following vocoids, suggesting that some of the slowing may be a non-specific response to feedback delay. Similar results between and within syllables suggest that temporal monitoring in speech has a flat lookahead structure.

Keywords: speech motor control, altered auditory feedback, timing

1. INTRODUCTION

A large body of literature has shown that speakers make use of sensory feedback to ensure accurate speech production. Evidence is particularly robust in the spectral domain, where speakers both compensate online [1]–[4] and adapt their future productions [5], [6] in response to perceived errors.

An outstanding question in the field is how speakers use temporal aspects of auditory feedback in speech production, particularly in the initiation of upcoming gestures. One hypothesis, selection-coordination theory [7], posits that speakers use a mix of feedforward and feedback mechanisms to time the initiation of speech gestures, with the balance of mechanisms based on phonological structures. Selection-coordination theory relies on the concept of "co-selection sets", which are sets of gestures that are selected together and executed using coordinative control (a type of feedforward control). Speakers activate the next co-selection set via sensory feedback that confirms the achievement of the goals of the ongoing co-selection set. Although the scope of gestures in a co-selection set has been suggested to vary based on the context or intent of any given utterance, syllable-sized units are suggested to be the prototypical example of coordinated gestures in a co-selection set, similar to what is used in the DIVA and GODIVA models of speech production [8], [9].

There is some evidence that speakers do attend to temporal aspects of auditory feedback. For example, a large body of literature has shown that delaying auditory feedback alters speech behavior, ranging from overall slowed production at delays from ~50-200 ms, to stuttering-like behavior at larger delays [10]–[13]. However, it is unclear if these responses are the result of compensating for delays (i.e., initiating gestures later because the previous goals were achieved later), or more general motor responses to perceived errors.

Cai et al. [14] explored this question by using bidirectional temporal perturbations, accelerating or decelerating the transition between "I" and "owe" in "I owe you a yoyo". Speakers significantly slowed their speech in the deceleration condition compared to both acceleration and baseline productions, but the acceleration condition did not significantly differ from baseline. Furthermore, the perturbation targeted a region of transition from one word to another, which may not be as specifically timed as gestures within a word. As such, this study did not provide direct evidence that speakers use auditory feedback to guide the timing of upcoming gestures. It also did not address how temporal control of speech interacts with phonological structure.

In the current study, we test the hypothesis suggested in selection-coordination theory that syllable-sized co-selection sets are timed using feedforward mechanisms, and subsequent syllables are timed by auditory feedback. Contrary to prediction, our results indicate that the timing of subsequent gestures responds to temporal changes in auditory feedback both within a single syllable and across syllables. This suggests that speakers use auditory feedback to guide the initiation of subsequent gestures regardless of syllabic affiliation.

2. METHODS

2.1. Participants

Six participants (3 F, 3 M) have participated in this study, of an anticipated 20. Participants ranged in age from 51-61 (median 53) and reported no history of neurological, hearing, or speech disorders. All participants were native speakers of General American English with a diphthongal production of /aɪ/. Participants were monetarily compensated for their time. All procedures were completed in accordance with the IRB at the University of Wisconsin – Madison and the University of California, San Francisco.

2.2. Target phrases

The target vowel for this study was the diphthong /aɪ/. To test the hypothesis that 1) gestural timing across syllable boundaries relies on sensory feedback and 2) gestural timing within a syllable is controlled in a feedforward manner as a single coordinated unit, the vowel was embedded in two target contexts:

1. **buy donuts** [baɪ doʊnəts]: the next consonant follows a word boundary as an onset consonant;
2. **guide boaters [**gaɪd boʊɾəɪz**]:** the next consonant immediately follows the vowel as a coda.

As selection-coordination theory suggests that co-selection sets are formed from highly practiced and frequently co-activated gestures, in this study we use syllable boundaries that are also word boundaries to increase the likelihood that the following gesture is activated differently in each phrase.

Table 1: Summary of phrases used in the experiment.

Phrase	Target	Following gesture
BUY donuts	aɪ	# d
GUIDE boaters	aɪ	d #

2.3. Auditory feedback perturbation

Temporal perturbations were achieved using a novel formant clamp functionality of Audapter [15]. The formant clamp feeds in a predetermined formant trajectory when the onset of the target vowel is detected using online status tracking (OST), and returns to veridical feedback when OST detects the end of the vowel.

Clamped formant trajectories were based on average productions from the most recent 10 unperturbed trials for a specific target word. To change the timing of the diphthong transition, the average formants were warped offline via nonlinear resampling. As the General American production of /aɪ/ has a relatively late transition to [ɪ], the resampling functions for acceleration and deceleration were not symmetrical, but rather calibrated to produce a large temporal difference between the accelerated and decelerated conditions without making the decelerated condition sound monophthongal. The resulting perturbations were thus temporally larger for acceleration than for deceleration. Example perturbations are provided in Figure 1.

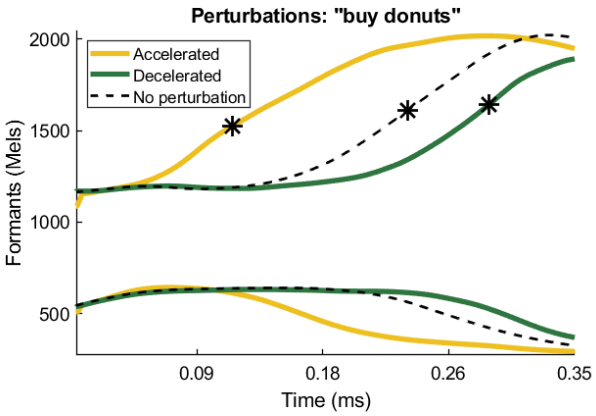

Figure 1: Mean F1 and F2 heard by one participant for "buy donuts", time-normalized to the overall mean duration of "buy". The asterisks on the F2 trajectory indicate point of maximum velocity in F2, where perturbation magnitude is measured.

2.4. Task

In each trial, participants saw the target phrase on the screen and read it out loud. The study consisted of 10 blocks. Participants heard veridical feedback for the first block, which served both as a baseline of production and to establish the mean duration and formant trajectories of each target phrase for formant clamping. In the remaining nine blocks, each target phrase was produced twice with each perturbation condition, and twice with no perturbation. In total, each unique combination of target phrase and perturbation condition was produced 18 times, with the exception of unperturbed trials, which were produced a total of 28 times (10 times during baseline block, 18 times during perturbation blocks). Trials were pseudorandomized with additional distractor phrases such that no two adjacent trials shared either target phrase or perturbation condition.

To facilitate the use of auditory feedback during the target vowel, participants were trained before the experiment to use a moderate speech rate and to place focus on the target word (e.g., "BUY donuts"). During the experiment, participants continued to receive feedback on their speech rate based on the duration of the target vowel as estimated by OST. The goal duration for the target vowel was 300 – 600 ms.

The achieved mean duration of the target diphthongs was 369 and 354 ms for "buy donuts" and "guide boaters", respectively (SD = 36, 33 ms).

2.5. Data processing and analysis

Trials were first segmented with Montreal Forced Aligner [16], and then hand-corrected by the first author, blind to perturbation condition. Transitions between /d/ and /b/ in "guide boaters" were not hand-corrected unless there was a visible release of /d/ or incorrect silent spans were inserted.

The dependent variable is the difference in segment duration between perturbed trials and unperturbed trials (including trials from the baseline block, as well as the unperturbed trials from experimental blocks). Reported estimated means are thus change from the mean of unperturbed trials.

Changes in production are analyzed for the target vowel, the following consonants, and the following vowel. Data was analyzed using linear mixed effects models using the lme4 package in R [17], [18]. Models included random intercepts per participant. Models were built incrementally, starting with the null model, and adding fixed effects of perturbation (acceleration, deceleration, unperturbed), target phrase (buy donuts, guide boaters), and their interaction. Models were compared using the anova function from lmerTest [19]. Post-hoc tests were conducted with the emmeans package [20].

3. RESULTS

Overall results for each perturbation condition are illustrated in Figure 2.

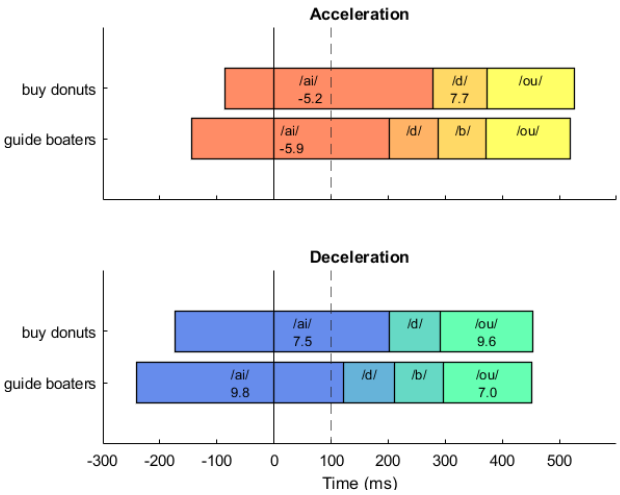

Figure 2: Changes in duration for segments in each phrase. Phrases are zero-aligned to the point of maximum perturbation (solid vertical line). The dashed line indicates the minimum latency necessary to compensate for auditory feedback (100 ms, [21]). All included values are significantly different from unperturbed.

3.1. Duration of /aɪ/

There is a significant effect of perturbation condition on the duration of /aɪ/ ($\chi^2(2) = 30.72$, $p < 0.0001$); compared to the unperturbed condition, /aɪ/ is shorter in the acceleration condition (-5.5 ± 2.1 ms, $p = 0.005$), and longer in the deceleration condition (8.7 ± 2.1 ms, $p = 0.007$). Adding target phrase to the model does not significantly improve model fit ($\chi^2(1) = 0.90$, $p = 0.34$); nor does the interaction between target phrase and perturbation condition ($\chi^2(2) = 0.69$, $p = 0.71$). This indicates that for both "buy donuts" and "guide boaters", speakers shortened the vowel under acceleration, and lengthened the vowel under deceleration. See Figure 3 for an illustration of these changes.

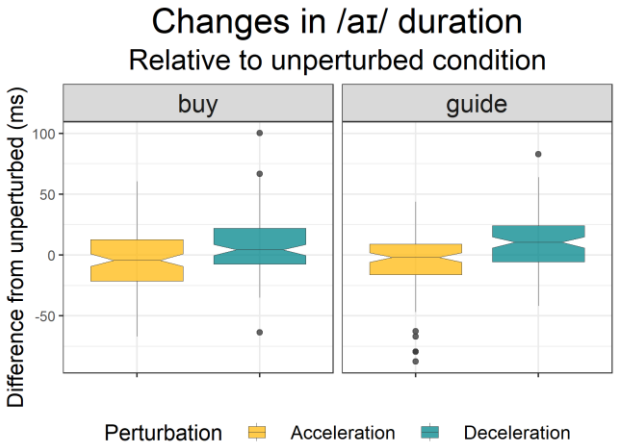

Figure 3: Changes in the produced duration of /aɪ/ under acceleration and deceleration for each phrase.

3.2. Duration of following consonants

Coda consonant of "guide": perturbation condition does not significantly improve model fit ($\chi^2(2) = 1.06$, $p = 0.59$), indicating that the coda /d/ in "guide" does not change in duration under either acceleration or deceleration.

*Onset consonants of following words (**d**onuts, **b**oaters):* Perturbation condition does not improve model fit ($\chi^2(2) = 2.86$, $p = 0.24$), nor does the addition of target word ($\chi^2(1) = 0.52$, $p = 0.47$). However, the interaction between target word and perturbation does improve model fit ($\chi^2(2) = 7.92$, $p = 0.02$). The /d/ in "buy donuts" is significantly longer in the acceleration condition compared to no perturbation (7.7 ± 2.3 ms, $p = 0.03$); no other segments show any significant differences.

3.3. Duration of following vowel /oʊ/

Adding perturbation condition significantly improves model fit ($\chi^2(2) = 43.94$, $p < 0.0001$). Vowels are significantly longer in the deceleration condition (7.8 ± 1.8 ms) compared to both acceleration (-0.5 ± 1.8

ms) and no perturbation (1.0 ± 1.7 ms; both $p < 0.0001$), but acceleration is not significantly different from no perturbation ($p = 0.43$). Target phrase does not significantly improve model fit ($\chi^2(1) = 1.57$, $p = 0.21$), nor does the interaction between target phrase and perturbation ($\chi^2(2) = 0.30$, $p = 0.86$). For both "buy donuts" and "guide boaters", the following stressed vowel is significantly longer in the deceleration condition (buy: 9.6 ± 1.8 ms; guide: 7.0 ± 1.8 ms) compared to both acceleration and no perturbation ($p \leq 0.001$ for all comparisons).

4. DISCUSSION

These preliminary results suggest that speakers use auditory feedback of their own speech to guide the timing of upcoming articulatory gestures, regardless of whether they belong to a different word. In the acceleration condition, the vowels in both "buy donuts" and "guide boaters" were significantly shorter than in unperturbed trials. In addition, the /d/ in "buy donuts" was significantly longer; no other segments showed significant changes. Similarly, in the deceleration condition the /aɪ/ in both "buy donuts" and "guide boaters" was significantly longer than in unperturbed trials. However, increased durations were also found in the /oʊ/ in the following words for both phrases.

Critically, in both acceleration and deceleration conditions, the /aɪ/ in "buy donuts" and "guide boaters" changed similarly. This suggests that speakers advanced the /d/ closure in time regardless of if it was in the same word or in the next word, indicating that the role of auditory feedback in gestural initiation is not affected by phonological boundaries, at least at the word level.

One possible interpretation of this pattern of data is that speech gestures are planned with specific intergestural timing goals that are based in motor predictions. When speakers receive feedback that some part of the plan (e.g., the lingual gesture) has achieved a state at a different time than they predicted, they adjust other gestures so that relative timing remains intact. In the case of "buy donuts" and "guide boaters", the tongue tip gesture adjusts to maintain relative timing with the tongue body: under acceleration, the tongue tip gesture is accelerated to achieve its goal earlier; under deceleration, it is delayed. Because our results using speech acoustics could examine only the achievement of oral closure, it is unclear whether these temporal changes are accomplished by advancing and delaying the initiation of the tongue tip gesture associated with the consonant or by altering the trajectory of an ongoing gesture. Thus, it is unclear if speakers do in fact use auditory feedback to time gestural initiation, as suggested by selection-coordination theory, or if temporal compensation is limited to adjusting ongoing movements (e.g., adjusting the velocity of a movement). More detailed information on this process could shed light on the planning horizon for gestures within and across syllable and/or word boundaries. This question could be resolved in the future through examination of speech kinematics.

Finally, in both the acceleration and deceleration perturbations, speakers lengthened unperturbed segments: in acceleration, the /d/ in "donuts" was lengthened, and in deceleration, the stressed vowel in the following word was lengthened in both phrases. One possibility is that there are two components to the reaction to temporal perturbations: first, a compensatory reaction, which appears in the perturbed vowel, and second, a non-specific response to a large temporal error, which appears later. Slowing down could facilitate the use of sensory feedback for state estimation after an error in temporal prediction is perceived by the speaker. An attempt at motor recovery may be reflected in effects induced by delayed auditory feedback, which at small magnitudes induces lengthened segments and at large delay magnitudes results in disfluencies and articulatory breakdown [11], [12]. It is unclear why only "buy donuts" showed additional lengthening in the acceleration condition, while neither /d/ nor /b/ (or both combined) showed any changes in "guide boaters". As the coda and onset gestures are frequently overlapped both in this dataset (as evidenced by the lack of release burst on /d/) and in English generally [22], one possibility is that there was some gestural elongation that resulted in increased gestural overlap, thus not producing acoustic evidence of elongation. Kinematic data could potentially provide insight on this issue.

5. CONCLUSION

Overall, this study provides evidence that speakers use auditory feedback to guide the timing of upcoming gestures. Crucially, this study indicates that speech gestures within a syllable are not timed exclusively via feedforward mechanisms, as proposed by selection-coordination theory. Instead, the data suggests that speakers always use feedback to maintain desired timing relationships between sequential speech gestures regardless of the syllabic or even word level affiliation of those gestures.

6. ACKNOWLEDGEMENTS

This work was supported in part by NIH grants F32DC019535 (RK), R01 DC017091 (BP), P50HD105353 (Waisman Center).

7. REFERENCES

[1] J. F. Houde and M. I. Jordan, "Sensorimotor adaptation of speech I: Compensation and adaptation," *Journal of Speech, Language, and Hearing Research*, vol. 45, pp. 295–310, 2002.

[2] K. G. Munhall, E. N. MacDonald, S. K. Byrne, and I. Johnsrude, "Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 384–390, 2009.

[3] J. A. Jones and K. G. Munhall, "Perceptual calibration of F0 production: Evidence from feedback perturbation," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1246–1251, 2000.

[4] V. M. Villacorta, J. S. Perkell, and F. H. Guenther, "Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception," *The Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. 2306–2319, 2007.

[5] J. F. Houde and M. I. Jordan, "Sensorimotor adaptation in speech production," *Science*, vol. 279, no. 5354, pp. 1213–1216, 1998.

[6] D. W. Purcell and K. G. Munhall, "Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation," *The Journal of the Acoustical Society of America*, vol. 120, no. 2, pp. 966–977, 2006.

[7] S. Tilsen, "Selection and coordination: The articulatory basis for the emergence of phonological structure," *Journal of Phonetics*, vol. 55, pp. 53–77, 2016.

[8] J. A. Tourville and F. H. Guenther, "The DIVA model: A neural theory of speech acquisition and production," *Language and cognitive processes*, vol. 26, no. 7, pp. 952–981, 2011.

[9] J. W. Bohland, D. Bullock, and F. H. Guenther, "Neural representations and mechanisms for the performance of simple speech sequences," *Journal of cognitive neuroscience*, vol. 22, no. 7, pp. 1504–1529, 2009.

[10] P. Howell and D. J. Powell, "Delayed auditory feedback with delayed sounds varying in duration," *Perception & psychophysics*, vol. 42, no. 2, pp. 166–172, 1987.

[11] K. T. Kalveram and L. Jäncke, "Vowel duration and voice onset time for stressed and nonstressed syllables in stutterers under delayed auditory feedback condition," *Folia Phoniatrica*, vol. 41, no. 1, pp. 30–42, 1989.

[12] J. R. Malloy, D. Nistal, M. Heyne, M. C. Tardif, and J. W. Bohland, "Delayed Auditory Feedback Elicits Specific Patterns of Serial Order Errors in a Paced Syllable Sequence Production Task," *Journal of Speech, Language, and Hearing Research*, vol. 65, no. 5, pp. 1800–1821, 2022.

[13] A. J. Yates, "Delayed auditory feedback.," *Psychological bulletin*, vol. 60, no. 3, p. 213, 1963.

[14] S. Cai, S. S. Ghosh, F. H. Guenther, and J. S. Perkell, "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*, vol. 31, no. 45, pp. 16483–16490, 2011.

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

[16] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner." Jan. 18, 2017. [Online]. Available: http://montrealcorpustools.github.io/Montreal-Forced-Aligner/

[17] D. Bates, M. Maechler, B. Bolker, S. Walker, and others, "lme4: Linear mixed-effects models using Eigen and S4," *R package version*, vol. 1, no. 7, pp. 1–23, 2014.

[18] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2019. [Online]. Available: https://www.R-project.org/

[19] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "Package 'lmerTest,'" *R package version*, vol. 2, no. 0, 2015.

[20] R. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*. 2019. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[21] U. Natke and K. T. Kalveram, "Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables," *Journal of Speech, Language, and Hearing Research*, vol. 44, pp. 1–8, 2001.

[22] D. Byrd, "Influences on articulatory timing in consonant sequences," *Journal of phonetics*, vol. 24, no. 2, pp. 209–244, 1996.