Auditory feedback is used for adaptive control of timing in speech

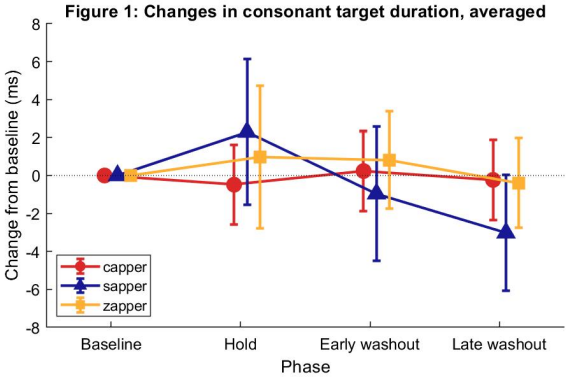
Robin Karlin, Chris Naber, and Benjamin Parrell
Department of Communication Sciences & Disorders, University of Wisconsin–Madison

Introduction: Real-time altered auditory feedback has demonstrated a key role for auditory feedback in both online feedback control and in updating feedforward/predictive control for future utterances. To date, much of this research has examined control in the spectral domain, e.g. vowel formants (Houde & Jordan, 1998; Houde & Jordan 2002; Purcell & Munhall 2006), f0 (Jones & Munhall 2000), intensity (Patel et al. 2015), and fricative spectral center of gravity (Shiller et al. 2009). However, relatively little is known about how speakers respond to similar perturbations in the time domain. One study (Mitsuya, MacDonald, and Munhall, 2014) found that speakers compensated for alterations to voice onset time (VOT, i.e. "tipper" vs. "dipper"). However, in their study the tokens for each participant were recorded in advance and not manipulated online. Thus, any changes produced by participants had no effect on their perceived productions. Here, we extend the temporal perturbation paradigm to test the hypothesis that auditory feedback is used to regulate timing in speech. We introduce a real-time perturbation of speech timing, contingent on ongoing production, to examine effects of temporal perturbations on the control of both relative timing between two distinct actions (VOT) and inherent timing of a single action (fricative and vowel duration).

Methods: 20 speakers (18F, 2M) participated in this study. No participant reported any history of speech, hearing, or neurological disorders. Both voiced and voiceless consonants were tested, with four target consonants: /g, k, z, s/; the experiment also targeted the vowel /æ/. Each consonant was tested in a separate session, and the order of the sessions was counterbalanced across participants. Each session 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 an intended maximum perturbation of 60 ms; and a 30-trial *washout* phase with veridical feedback. On each trial, the participant produced the phrase "a *TARGET*", where the target was one of the words "gapper", "capper", "zapper", and "sapper". For /g, k/, VOT was lengthened, while for /s, z/ the fricative was lengthened. The vowel /æ/ immediately following the consonant was shortened by the same amount so that the overall syllable duration remained unchanged. The experiment was presented in Matlab; time perturbation was achieved with Audapter (Cai et al 2008). The achieved maximum perturbation for /k, s, z/ was 41.7 ± 11.4, 43.5 ± 3.2, and 40.2 ± 9.9 ms, respectively; due to insufficient perturbation, data from /g/ has been excluded from analysis. 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.

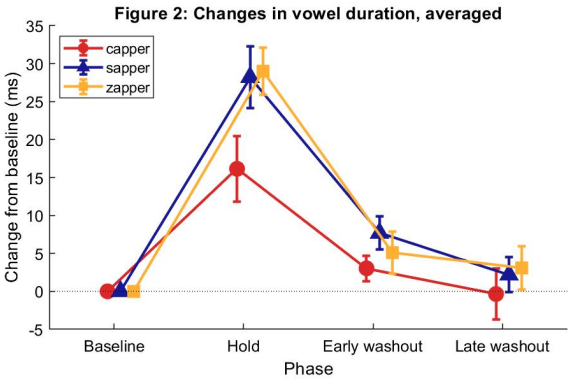
The resulting audio was hand-segmented to obtain consonant (VOT or fricative) and vowel durations. The last 10 trials of the baseline phase served as a baseline of comparison for adaptation and aftereffects: adaptation was measured from the last 10 trials of the hold phase; aftereffects were measured from the first 10 and last 10 trials of the washout phase (early and late washout, respectively). The resulting data were analyzed with a linear-mixed effects model, with fixed effects of phase, word, and their interaction. Random intercepts were included for participant. Models were built incrementally and compared with likelihood ratio tests. Post-hoc tests were done using least means squared with a Bonferroni-Holm adjustment. Estimates are reported as distance from the baseline mean; positive values indicate increased duration and negative values indicate decreased duration.

Results: *Consonants* Overall, the expected shortening for lengthened consonant durations was not found (Fig. 1). There is a main effect of phase ($\chi^2(3) = 8.59$, $p = 0.04$); however, the only two phases that differ significantly from each other are hold (1.1 ± 1.4 ms) and late washout (-1.2 ± 1.4 ms, $p = 0.02$). Adding word to the model does not significantly improve the fit ($\chi^2(2) = 2.59$, $p = 0.27$), nor does the addition of the interaction between word and phase ($\chi^2(6) = 11.95$, $p = 0.06$).



Vowel In contrast, durational adaptation was observed in the

vowel (Fig. 2), shown by a main effect of phase ($\chi^2(3) = 801.11$, $p < 0.0001$). All phases are significantly different from each other (all $p \leq 0.0001$) except baseline and late washout ($p = 0.06$). Vowels are the longest in the hold phase (24.7 ± 1.2 ms) and are longer in early washout (5.28 ± 1.2 ms) than in baseline. Vowel duration returns to baseline values by the late washout phase (1.6 ± 1.2 ms). The aftereffects seen in the washout phases indicate changes to feed-forward temporal control. Word as a fixed effect significantly improves the fit of the model ($\chi^2(2) = 47.55$, $p < 0.0001$), as does the interaction between word and phase ($\chi^2(6) = 50.76$, $p < 0.0001$). The interaction is driven by the difference in magnitude of adaptation across words; the vowel is lengthened less during the hold phase in *capper* (16.0 ± 1.6 ms) than in either *sapper* (27.6 ± 1.5 ms) or *zapper* (29.2 ± 1.5 ms).



Proportional analysis Although consonants did not decrease in absolute duration, the increase in vowel duration decreases the proportion of the initial CVC syllable occupied by the onset consonant, thus effectively "shortening" it. There is a significant effect of phase on consonant proportion ($\chi^2(3) = 127.24$, $p < 0.0001$); the consonant takes up a lower proportion of the syllable during the hold phase (-1.5 ± 0.3%) than all other phases (all $p < 0.0001$).

Conclusions: This study shows that speakers incorporate auditory feedback in predictive control of speech timing: speakers compensated for shortened vowels by lengthening during the hold phase, and some lengthening was still present when veridical feedback was restored, indicating that motor plans were adjusted for future utterances. A similar effect was not found for lengthened consonants, contra previous findings; however, in combination with vowel lengthening, overall participants reduced the proportion of the syllable occupied by the consonant. This suggests that speakers may attend to proportional timing within a syllable rather than absolute millisecond timing of individual segments, which aligns with previous findings in both production (Boucher 2002) and perception (Port and Dalby, 1982). It may also be the case that the timing of syllable rimes is more flexible than onsets (Oschkinat and Hoole, 2019). Adapting vowel duration would both satisfy proportional timing goals and utilize the more flexible strategy to maximize compensatory effect.

References

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.

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

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. Science, 279(5354), 1213-1216.

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

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

Mitsuya, T., MacDonald, E. N., & Munhall, K. G. (2014). Temporal control and compensation for perturbed voicing feedback. The Journal of the Acoustical Society of America, 135(5), 2986-2994.

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

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

Purcell, D. W., & Munhall, K. G. (2006). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. The Journal of the Acoustical Society of America, 120(2), 966-977.

Shiller, D. M., Sato, M., Gracco, V. L., & Baum, S. R. (2009). Perceptual recalibration of speech sounds following speech motor learning. The Journal of the Acoustical Society of America, 125(2), 1103-1113.