Kinematic evidence of centering during vowel production

Benjamin Parrell[1], Mark Tiede[2], Vince Gracco[2], Doug Shiller[3]

[1]*University of Wisconsin–Madison*
[2]*Haskins Laboratories*
[3]*Université de Montréal*

bparrell@wisc.edu, tiede@haskins.yale.edu, vincent.gracco@yale.edu, douglas.shiller@umontreal.ca

Abstract

When producing a vowel, productions that begin relatively far from the center of a speaker's acoustic F1-F2 distribution for that vowel tend to move towards the center of the distribution by the vowel midpoint, a phenomenon known as centering. To date, centering has been only demonstrated in the acoustic signal. Here, we examine whether centering can also be seen at the level of speech kinematics. Using electromagnetic articulography to track the tongue, jaw and lips, we show that centering is observable in the kinematic signal, most clearly in the tongue dorsum. We also show that centering begins to occur in the kinematic signal prior to vowel onset, consistent with the idea that the phenomenon is not driven purely by auditory feedback.

Keywords: Speech production, vowel centering, kinematics

1. Introduction

The role of sensory systems in speech production has been a topic of interest since the early days of speech research (Fairbanks, 1954; Yates, 1963). Typically, the role of the auditory system in online control of speech has been probed by delaying auditory feedback (Yates, 1963) or by altering its spectral characteristics (Burnett et al., 1998; Purcell & Munhall, 2006). Many studies have shown through these methods that the neural control of speech movements is indeed sensitive to such external auditory perturbations.

However, the extent to which the speech articulatory system makes use of sensory feedback during normal, unaltered production is less clear. Blocking tactile sensation does lead to some changes in articulation, particularly for fricatives (Scott & Ringel, 1971). On the other hand, blocking auditory feedback with masking noise does not substantially affect articulation (Ladefoged, 1967; Ringel & Steer, 1963), though it does have a greater impact on voicing, pitch control, and timing. In general, the small effects of sensory masking on articulation suggest that online sensory feedback plays a minor, but non-negligible, role in typical speech production. This is particularly true for auditory feedback, where the delays associated with feedback corrections (~150 ms) are longer than the duration of many speech sounds (Tourville et al., 2008).

Recently, an alternative method to masking or external perturbations has been proposed to investigate the possible role of sensory feedback (Niziolek et al., 2013). Rather than imposing sensory perturbations, this method leverages the natural variability found in speech production to examine how speakers alter their productions online. These studies have shown that vowel productions which initially fall near the edge of the sound's distribution in F1/F2 space (for a given talker) exhibit movement towards the middle of the distribution over time, a phenomenon known as *centering*. While the sensorimotor mechanisms underlying this behavior are not clear, it has been suggested that auditory feedback may play a role. First, masking noise has been shown to attenuate the magnitude of the centering behavior (Niziolek et al., 2015). Second, trials which fall near the edge of the acoustic vowel distribution generate neural signals that are similar to those generated when auditory feedback is externally perturbed, suggesting that the auditory system is able to distinguish these peripheral trials from more typical productions (Niziolek et al., 2013). Moreover, the degree to which the peripheral trials differ from more central trials at a neural level is correlated with the magnitude of the centering behavior across participants.

Here, we test whether centering can be observed at the level of speech kinematics in addition to speech acoustics. The examination of speech kinematics also allows us to test whether centering behavior occurs closer to vowel onset, before auditory feedback would be available to the nervous system. If such centering occurs, it would suggest that this behavior may rely at least partially on sources other than auditory feedback, such as internal predictions (Parrell et al., 2019), somatosensation, or increasing restrictions on permitted variability at the planning level (Keating, 1990).

2. Methods

This pilot study involved two adult native speakers of American English (one female [S1], one male [S2]), with no reported speech, language or hearing deficits. All procedures were approved by the Yale University IRB.

2.1. Kinematic and Acoustic Recording

Electromagnetic articulography (*EMA*; Wave, Northern Digital Inc.) was used to measure the 3D position of sensors attached to the tongue (midsagittal dorsum [TD], blade [TB] and tip [TT]), jaw (upper and lower incisors, left premolar), and lips (upper/lower) relative to the head, sampled at 100 Hz. Head-correction was carried out on the basis of three reference sensors placed on the left and right mastoid and the nasion. Tongue, jaw and lip marker positions were aligned to each participant's occlusal plane, identified using a bite-plate containing three EMA sensors. Synchronized audio was collected at 44.1 kHz.

2.2. Procedure

Participants were instructed to produce individual words (visually presented on a computer display) drawn from the set *Ed*, *add*, *ebb*, *ab*, *shed*. These words were chosen to test whether centering depended on coarticulatory constraints between the vowel and coda consonant, which are higher for lingual codas (*Ed, add, shed*) than for bilabial codas (*ebb, ab*). 60 repetitions of each word were produced in pseudo-randomized order under four production conditions: 1) **Spoken** - words produced aloud with normal auditory feedback; 2) **Pantomimed** - silent articulation of each word without any phonation or frication; 3) **Whispered** - words whispered with normal auditory feedback;

4) **Whispered in masking noise** - words whispered while pink masking noise was presented over insert headphones.

Participants produced the target words in blocks of 150 productions (30x each of the 5 words) under a given speaking condition. Two blocks were produced under each speaking condition (cycling through the four speaking conditions two times), for a total of 60 repetitions per word in each condition. Only the Spoken condition is analyzed in this paper.

2.3. Acoustic data analysis

For each token produced under the different speaking conditions (all except Pantomime), the vowel onset and offset was manually segmented in Matlab on the basis of the acoustic waveform and spectrogram. Following (Niziolek & Kiran, 2018), F1/F2 traces spanning the vowel were estimated using LPC analysis in Praat (Boersma & Weenink, 2019), converted from Hz to mel units, and then averaged over a 50 ms window beginning at vowel onset (*Window 1*), as well as a 50 ms window centered in the middle of the vowel (*Window 2*). For each participant, using the F1/F2 values averaged over Window 1 and Window 2 and grouping the data within each word and speaking condition, a measure of *acoustic vowel centering* was obtained for each utterance as follows: A measure of *vowel distance* was computed within each of the two time windows as the Euclidean distance between the trial's F1/F2 values and the median F1/F2 values within that time window. Based on this distance metric, a set of *peripheral trials* was defined for further analysis as the 1/3 of trials furthest from the F1/F2 median in Window 1 (Niziolek et al., 2013). The measure of vowel centering for these peripheral trials was then computed as the signed difference in vowel distance between Window 1 and Window 2, such that positive values correspond to a reduction in variability for Window 2 associated with centering.

In order to ensure that the reduction in variability observed between time Window 1 and Window 2 was not due simply to regression to the mean over time, we additionally calculated the same centering metric but *reversing* the temporal order of the time windows in the analysis (i.e., treating the data from Window 2 as if it was the onset of the movement, and Window 1 as if it were the vowel midpoint, see Fig. 1). By examining whether the original centering measure exceeds this new measure of *reverse centering* (both of which include the same regression to the mean effects due to measurement noise or random physical variation), we can test the robustness of the centering effect as a phenomenon beyond simply a statistical artifact (Niziolek & Kiran, 2018).

2.4. Kinematic data analysis

Kinematic analysis was restricted to the tongue in the midsagittal plane (antero-posterior, x, and infero-superior, z). Using the same two time windows (Window 1 and Window 2) identified on the basis of the acoustic signal for each utterance, the mean x- and z-position was computed for each of the three tongue markers. A measure of *kinematic vowel centering* was computed using the same approach described above for acoustics, only in this case using the x and z positions of the EMA sensors in place of F1 and F2. The Euclidean distance relative to the median position was computed for each trial within each of the two time windows, and the difference in distance between Window 1 and Window 2 for the peripheral trials in Window 1 served as the measure of vowel centering (again, with positive values corresponding to a reduction in variability in Window 2 associated with centering).

A second variation of the centering analysis involved using a time Window 1 just prior to the vowel acoustic onset (-100 to -50 ms) and a time Window 2 soon after onset (+25 to +75 ms), in order to examine the possibility of centering effects at the earliest stages of movement (prior to the availability of auditory feedback).

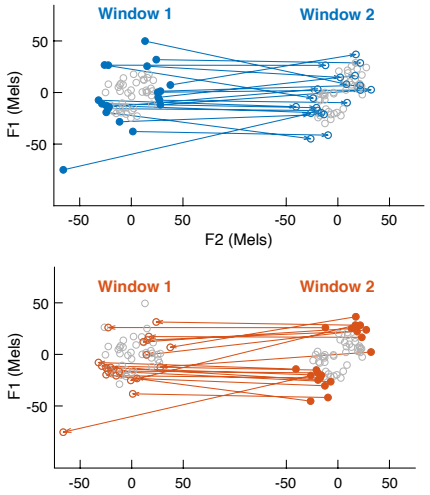

Figure 1: *Calculating centering (blue, change from Window 1 to Window 2) and reverse centering (orange, change from Window 2 to Window 1).*

2.5. Statistical analysis

Given the low number of participants, data from each participant was analyzed separately using ANOVAs. All models had *word* and *centering* (forward vs. reverse centering), as well as their interaction, as fixed factors.

2.6. Acoustic centering after vowel onset

The magnitude of centering was higher than reverse centering for both participants (Fig 2; S1: 8.0 vs 5.8 mels; S2: 17.2 vs 10.1 mels), although the effect of *centering* was not significant for either participant (S1: $F(1,140) = 2.34$, $p = 0.12$; S2: $F(1,140) = 3.5$, $p = 0.06$). The effect of *word* was also not significant for either participant ($p > 0.09$), However there was a significant interaction between *centering* and *word* for S2 ($F(4,140 = 6.5$, $p < 0.0001$) but not for S1 ($F(4,140) = 2.2$, $p = 0.06$).

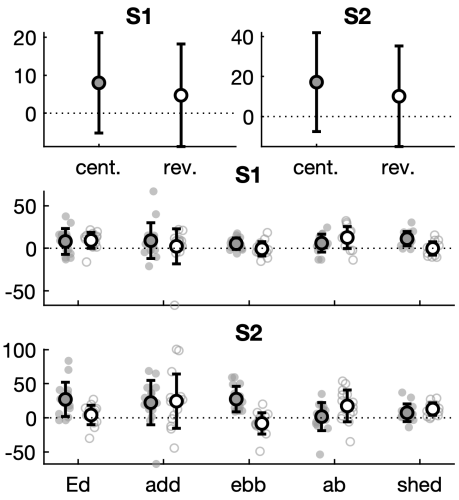

Figure 2: *Acoustic centering in mels for S1 (top left, middle) and S2 (top right, bottom) overall and by word. Shaded circles indicate centering and open circles indicate reverse centering.*

2.7. Kinematic centering after vowel onset

The magnitude of centering was significantly higher than reverse centering for sensors on the tongue dorsum (Fig 3; S1: 6.1 mm vs 1.6 mm; S2: 5.5 mm vs 1.2 mm) and jaw (Fig 4; S1: 3.2 mm vs 0.4 mm; S2: 12.5 mm vs 2.4 mm). The magnitude of centering was also higher than reverse centering for the tongue tip (13.3 vs 1.9 mm) and tongue body (8.7 vs 3.5 mm) for S2, while the magnitude of reverse centering was higher than centering for the tongue body (2.2 vs 6.3 mm) for S1. While there were a number of cases where the difference between forward and reverse centering varied across words (Figs 3,4; Table 1), there were no consistent patterns in this variation by coda place of articulation or the presence vs absence of an onset consonant.

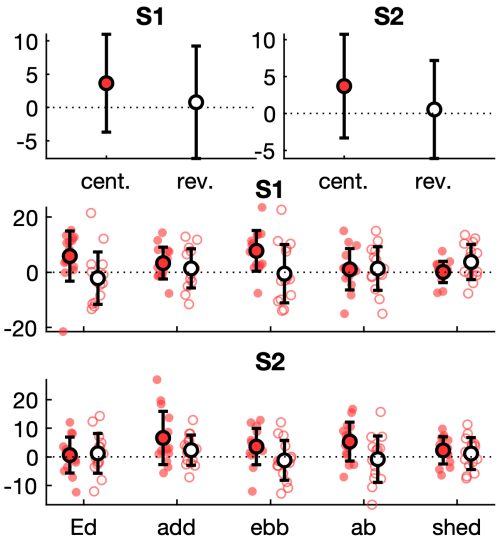

Figure 3: *Kinematic centering (mm) for tongue dorsum sensor for S1 (top left, middle) and S2 (top right, bottom) overall and by word. Shaded circles show centering, open circles show reverse centering.*

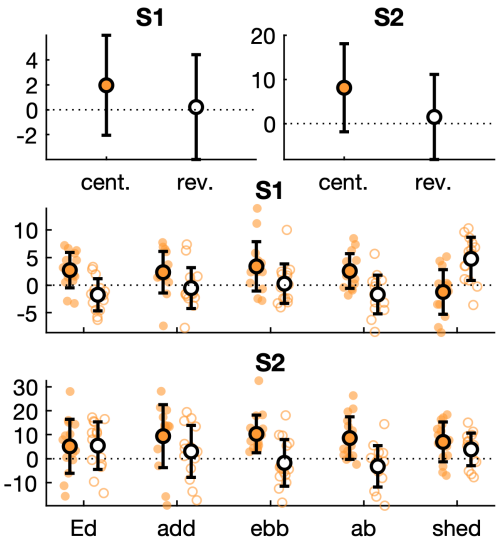

Figure 4: *Kinematic centering in mm in the jaw sensor for S1 (top left, middle) and S2 (top right, bottom) overall and by word. Shaded circles indicate centering and open circles indicate reverse centering.*

Table 1: *Participant-specific ANOVAs showing main effect of centering and centering*word interaction for tongue and jaw EMA sensors. Word was not significant for any comparison (all p >.0.08)*

		S1	S2
TT	centering	$F(1,140) = 2.5$, $p = 0.12$	**$F(1,140) = 22.8$, $p < 0.0001$**
	centering * word	$F(4,140) = 2.0$, $p = 0.10$	**$F(4,140) = 6.8$, $p < 0.0001$**
TB	centering	**$F(1,140) = 5.3$, $p = 0.02$**	**$F(1,140) = 6.5$, $p = 0.01$**
	centering * word	$F(4,140) = 1.8$, $p = 0.12$	**$F(4,140) = 3.5$, $p = 0.01$**
TD	centering	**$F(1,140) = 4.6$, $p = 0.03$**	**$F(1,140) = 4.5$, $p = 0.03$**
	centering * word	**$F(4,140) = 3.5$, $p = 0.01$**	$F(4,140) = 1.4$, $p = 0.24$
Jaw	centering	**$F(1,140) = 8.3$, $p = 0.005$**	**$F(1,140) = 17.9$, $p < 0.0001$**
	centering * word	**$F(4,140) = 11.1$, $p < 0.0001$**	$F(4,140) = 2.3$, $p = 0.06$

2.8. Kinematic centering prior to vowel onset

We found that centering magnitude, measured as the reduction in distance to the median from before acoustic vowel onset to immediately following vowel onset, was significantly greater than reverse centering in the same time windows for both participants in the tongue dorsum sensor (Fig 5; S1: 4.5 vs 1.1 mm ; S2: 6.1 vs 2.9 mm), in the tongue tip for S1 (2.6 vs 0.6 mm), and in the tongue body for S2 (7.0 vs 3.9 mm). See Table 2 for details.

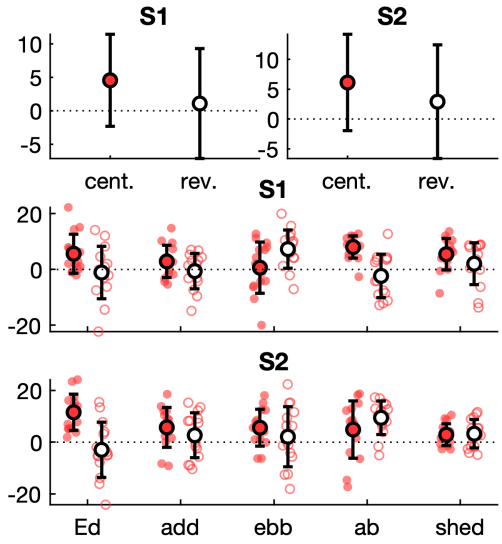

Figure 5: *Kinematic centering in mm in the tongue dorsum sensor (TD) for S1 (top left, middle) and S2 (top right, bottom) overall and by word. Shaded circles indicate centering and open circles indicate reverse centering.*

Table 2: *Participant-specific ANOVAs showing main effect of centering and centering*word interaction for tongue and jaw EMA sensors. Word was not significant for any comparison (all p >.08)*

		S1	S2
TT	centering	***F*(1,140) = 5.3, *p* = 0.03**	*F*(1,140) = 0.16, p = 0.69
	centering* word	***F*(4,140) = 5.9, *p* = 0.0003**	*F*(4,140) = 0.7, *p* = 0.61
TB	centering	*F*(1,140) = 1.2, *p* = 0.28	***F*(1,140) = 7.0, p = 0.01**
	centering* word	***F*(4,140) = 7.3, *p* < 0.0001**	***F*(4,140) = 2.9, *p* = 0.03**
TD	centering	***F*(1,140) = 8.9, *p* = 0.003**	***F*(1,140) = 5.6, *p* = 0.02**
	centering* word	***F*(4,140) = 6.0, *p* = 0.0002**	***F*(4,140) = 5.4, *p* = 0.0004**
Jaw	centering	*F*(1,140) = 1.8, *p* = 0.19	*F*(1,140) = 2.4, *p* = 0.12
	centering* word	*F*(4,140) = 0.6, *p* = 0.69	*F*(4,140) = 2.2, *p* = 0.07

3. Discussion and conclusion

Overall, the amount of acoustic centering observed in these data is smaller than that reported in previous studies and, indeed, the main effect of acoustic centering was not significant for either participant. This suggests that the two speakers in the current study may be on the lower range of centering behavior observed in the larger population. Whether this is due to individual variation or, potentially, to changes in speech caused by speaking with the EMA sensors is unknown at this point, but worth further investigation. The articulatory component also deserves further study, as the extent to which speakers control (unobserved) parasagittal tongue configuration to effect acoustic centralization is also unclear.

Despite the low magnitude of acoustic centering in our data, we have shown that centering is indeed visible in speech kinematics. Centering is most consistently seen in the tongue dorsum, though it also variably seen in other articulators such as the tongue tip and jaw. There were no clear and consistent differences between the stimulus words in our data. Although significant interactions between word and centering direction were found in many analyses, it was not the case that words with bilabial codas patterned differently from words with coronal codas. This may suggest that centering is not driven purely by coarticulatory effects, as the coarticulatory constraints between the vowel and adjacent consonants for *Ed, add,* and *shed* are substantially higher than for *ebb* and *ab*.

Importantly, centering is present not only after the acoustic onset of the vowel, but also from before to soon after vowel onset. This suggests that centering is driven, at least partly, by factors other than auditory feedback. These potential influences include somatosensory feedback, internal predictions (of auditory feedback, somatosensory feedback, and/or articulator positions), and increasing restrictions on the permitted variability at vowel midpoint compared to vowel onset, consistent with the window model of coarticulation (Keating, 1990).

4. References

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer*. (Version 6.0.47) [Computer software]. ttp://www.praat.org/

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

Fairbanks, G. (1954). Systematic Research In Experimental Phonetics:* 1. A Theory Of The Speech Mechanism As A Servosystem. *Journal of Speech and Hearing Disorders*, *19*(2), 133–139. https://doi.org/10.1044/jshd.1902.133

Keating, P. A. (1990). The window model of coarticulation: Articulatory evidence. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology I* (pp. 451–470). Cambridge University Press.

Ladefoged, P. (1967). *Three areas of experimental phonetics*. Oxford University Press.

Niziolek, C. A., & Kiran, S. (2018). Assessing speech correction abilities with acoustic analyses: Evidence of preserved online correction in persons with aphasia. *International Journal of Speech-Language Pathology*, *0*(0), 1–11. https://doi.org/10.1080/17549507.2018.1498920

Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2013). What does motor efference copy represent? Evidence from speech production. *Journal of Neuroscience*, *33*(41), 16110–16116. https://doi.org/10.1523/JNEUROSCI.2137-13.2013

Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2015). The contribution of auditory feedback to corrective movements in vowel formant trajectories. In T. S. C. for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. The University of Glasgow.

Parrell, B., Ramanarayanan, V., Nagarajan, S., & Houde, J. (2019). The FACTS model of speech motor control: Fusing state estimation and task-based control. *PLOS Computational Biology*, *15*(9), e1007321. https://doi.org/10.1371/journal.pcbi.1007321

Purcell, D. W., & Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *Journal of Acoustical Society of America*, *119*(4), 2288–2297.

Ringel, R. L., & Steer, M. D. (1963). Some Effects of Tactile and Auditory Alterations on Speech Output. *Journal of Speech, Language, and Hearing Research*, *6*(4), 369–378. https://doi.org/10.1044/jshr.0604.369

Scott, C. M., & Ringel, R. L. (1971). Articulation without oral sensory control. *Journal of Speech, Language, and Hearing Research*, *14*(4), 804–818.

Tourville, J. A., Reilly, K. J., & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage*, *39*(3), 1429–1443. https://doi.org/10.1016/j.neuroimage.2007.09.054

Yates, A. J. (1963). Delayed auditory feedback. *Psychological Bulletin*, *60*(3), 213–232. https://doi.org/10.1037/h0044155