THE CONTRIBUTION OF AUDITORY FEEDBACK TO CORRECTIVE MOVEMENTS IN VOWEL FORMANT TRAJECTORIES

Caroline A. Niziolek[1], Srikantan S. Nagarajan[2], John F. Houde[1]

Depts. of [1]Otolaryngology and [2]Radiology, University of California, San Francisco
cniziolek@ohns.ucsf.edu, sri@ucsf.edu, houde@phy.ucsf.edu

ABSTRACT

How much does auditory feedback shape the trajectory of a spoken utterance? When auditory feedback is altered experimentally, speakers make compensatory vocal adjustments that serve to correct for the alteration. However, it is unclear to what degree the sound of one's own voice is used to guide speech movements in more natural contexts. In this study, we compared the formant trajectories of monosyllabic words spoken in different levels of masking noise. Spoken vowels exhibited a "centering" effect in which formants that started out at the periphery moved to the center (median) with time. This effect occurred across all speakers and all noise conditions, although it was greatest in quiet and smallest in masking noise, when auditory feedback was not available. This finding suggests that auditory feedback substantially contributes to an ongoing corrective process in natural speech, although it is not the sole driver of vowel centering.

Keywords: speech production, auditory feedback, noise masking, variability

1. INTRODUCTION

Does auditory feedback play a role in shaping spoken acoustics in everyday speech? In feedback alteration studies, speakers unconsciously correct for imposed changes to pitch [2], [9], formant frequencies [3], [5], [13], and other acoustic cues [1], [4], [12], suggesting that auditory feedback is closely monitored and used to keep productions on target. However, it is unclear to what degree this monitoring and correction occurs in natural speech, when there is no artificial mismatch between what is spoken and what is heard. To characterize the contribution of auditory feedback without using artificial feedback alterations, our recent work [6] has taken advantage of the acoustic variability observed across multiple repetitions of the same syllable. By considering acoustic variance as a kind of natural feedback alteration, it is possible to examine how the most error-like utterances (that is, those that deviate greatly from the median utterance in a given context) are brought closer to the median over the course of a single syllable. This correction, termed *vowel centering*, consists of the inward movement of a given utterance—i.e., from the periphery of the formant distribution to the center—across the duration of that utterance.

The centering behavior may be driven largely by sensory feedback signals that rapidly alert the speaker to potential error-like productions; however, it may alternatively be explained by the passive dynamics of the motor system as a motor command transitions from onset to steady state. Here, we extend past work on vowel centering by examining whether it is affected by the availability of auditory feedback. If auditory feedback is not a driver of the centering phenomenon, then centering should be equally strong in the presence of feedback (speaking in quiet) and in its absence (speaking in noise).

2. METHODS

2.1. Procedure

Ten speakers (four female) with self-reported normal hearing and speech participated in the experiment. All experimental procedures were approved by the Institutional Review Board at the University of California, San Francisco. Participants were seated in a soundproof booth and wore headphones (Beyerdynamic DT 880 PRO) and a head-worn condenser microphone (AKG Pro Audio C520) positioned at a distance of 1 inch from the corner of the mouth. Participants were visually prompted with text strings to produce 150 randomized tokens each of the words "eat," "Ed," and "add," chosen to elicit the vowels /i/, /ɛ/, and /æ/. This paper presents data from only the central vowel /ɛ/. The 450 utterances were split between three noise conditions of 150 trials each, presented in a random order: quiet (no noise), low noise (white noise presented at 70 dB SPL), and high noise (white noise presented at 85 dB SPL). In addition to the text prompts, a VU (volume unit) meter was displayed on the screen, giving constant feedback about the volume of each utterance. Though we cannot rule out the possibility that not all auditory feedback was masked, participants were instructed to speak very quietly, keeping their volume to a level on the VU meter chosen to render their speech inaudible in the high noise condition.

2.2. Acoustic analysis

The first two vowel formant frequencies (F_1 and F_2) were tracked using the wave_viewer software package for Matlab [7]. The formants of one subject were excluded because of extremely poor tracking. Formant tracks were converted to mels, a perceptually-based logarithmic frequency scale [8], to better compare across formants and normalize across participants. Formants were averaged in each of two time windows: the first 50 ms of each production (*initial* time window, F_{1init} and F_{2init}) and the middle 50% of each production (*mid-utterance* time window, F_{1mid} and F_{2mid}). Subject-wise median formants for the vowel /ɛ/ were then calculated for each of these two time windows. For each trial, initial distance to the median was calculated as the Euclidean distance in 2D formant space, as in the following formula (median denoted by ~):

$$(1) \qquad d_{init} = \sqrt{\left(F_{1init} - \tilde{F}_{1init}\right)^2 + \left(F_{2init} - \tilde{F}_{2init}\right)^2}$$

A tercile split was used to define the trials closest to and farthest from the median, denoted as *central* and *peripheral* trials, respectively.

Similarly, for each trial, the mid-utterance distance to the median was calculated:

$$(2) \qquad d_{mid} = \sqrt{\left(F_{1mid} - \tilde{F}_{1mid}\right)^2 + \left(F_{2mid} - \tilde{F}_{2mid}\right)^2}$$

To assess centering during the course of spoken utterances, we followed the procedure detailed in [6]. Briefly, the mid-utterance distance was subtracted from the initial distance to yield the *centering* for each trial, $C = d_{init} - d_{mid}$, such that positive values indicated movement towards the median over the course of the utterance. As in [6], the centering for peripheral trials was used as the dependent variable in a two-way ANOVA with factors of subject and noise level.

Importantly, because apparent centering behavior could result from mere regression to the mean, we used an ANOVA to compare absolute formant movement, that is, the Euclidean distance between the starting formants and mid-utterance formants, in center and peripheral trials. We additionally tested whether the average distance to the median over *all* trials (not just peripheral trials) decreased from the beginning to the middle of the trial.

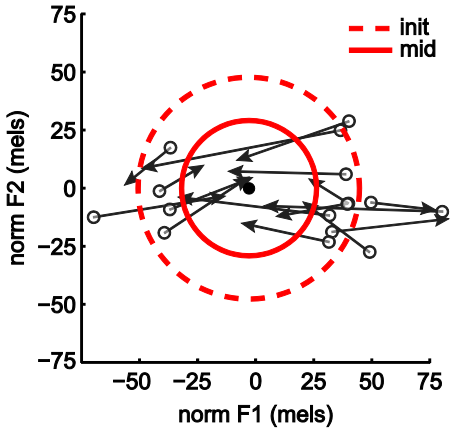
As shown in [6], speakers with larger variance (more formant spread across utterances) exhibit greater centering. To rule out the possibility that differences in centering across noise levels could be accounted for by noise-correlated differences in variance, we examined initial variance using a two-way ANOVA with factors of subject and noise level.

Finally, we looked at the effect of vowel duration on centering, as longer utterances might have more time for formant movement. The duration of the vowel portion of each trial was used as the dependent variable in a two-way ANOVA with factors of subject and noise level. Additionally, we tested whether average vowel duration was correlated with average centering on a per-subject basis.

3. RESULTS

All subjects exhibited centering behavior: overall, trials that started out far from the median moved inwards over time. An example of centering from a single subject (S01) in the quiet condition is shown in Fig. 1. Formant values are normalized to the median (black dot at (0,0)). Open circles denote the formants of each peripheral trial at utterance onset, relative to the median at utterance onset; connected arrowheads denote the formants of the same trials at mid-utterance, relative to the mid-utterance median. The radii of the red circles represent the average distance to the median in these two time windows.

Figure 1: A single-subject example of vowel centering over time (dashed ellipse = average d_{init}; solid ellipse = average d_{mid}).



The magnitude of centering differed across subjects (two-way ANOVA, main effect of subject, $F = 3.63$, $p = 0.0005$): population marginal means ranged from 4.2 to 33.6 mels. Importantly, this centering was not merely due to regression to the mean: peripheral trials traversed a larger average distance in 2D Euclidean space than center trials (three-way ANOVA, main effect of trial type, $F = 16.05$, $p = 0.0001$), with peripheral trials moving an average of 45.2 mels and center trials moving an average of 38.0 mels. Additionally, overall distance

to the median decreased from onset to mid-utterance across all trials.

Furthermore, the magnitude of centering differed across conditions (two-way ANOVA, main effect of noise level, $F = 4.6$, $p = 0.033$). Fig. 2 shows peripheral trials from all subjects, normalized as in Fig. 1. Across all subjects, centering was greatest in the quiet condition and smallest in the high-noise condition, with population marginal means of 18.8 and 11.1 mels, respectively (Fig 3.).

Figure 2: Peripheral trials for all subjects overlaid.

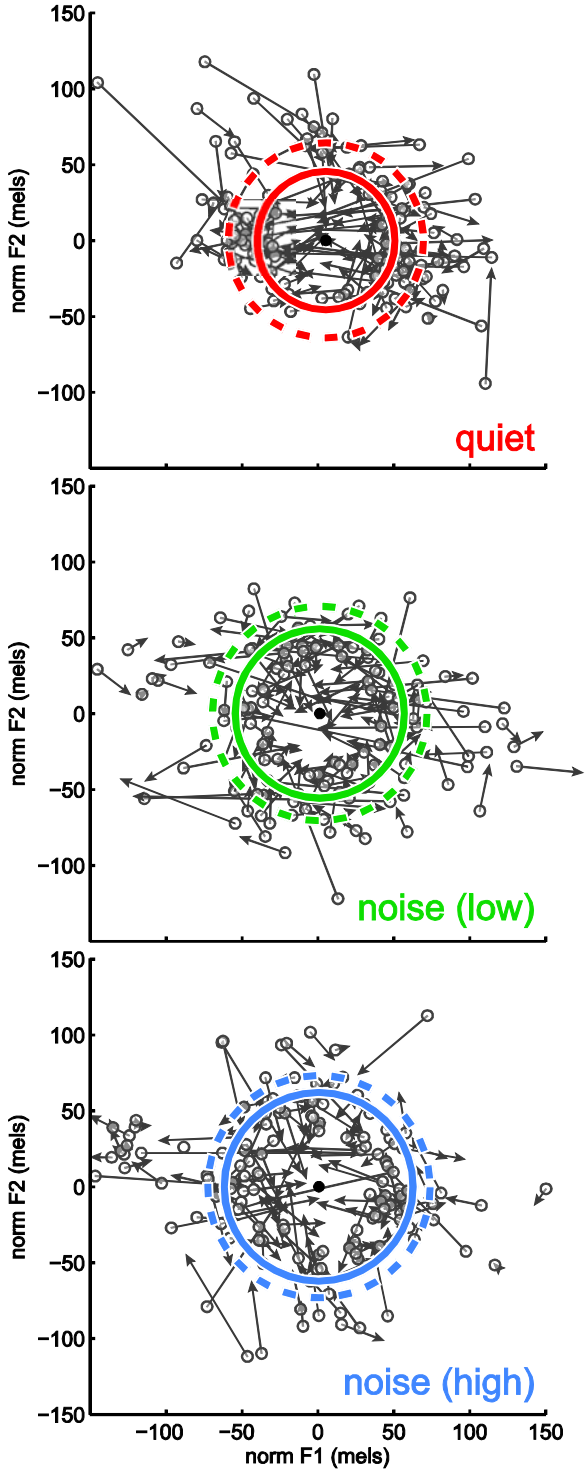
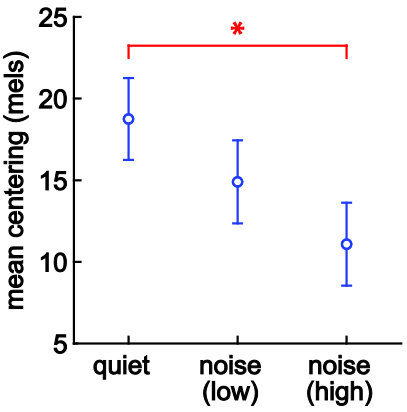
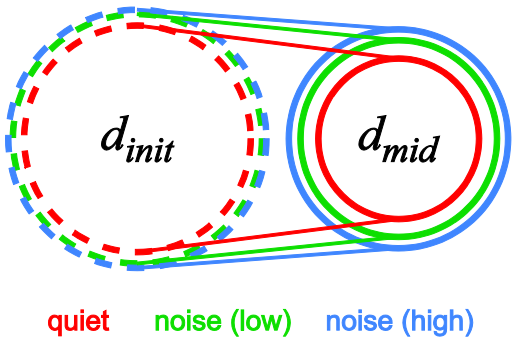


Figure 3: Magnitude of centering in different noise conditions (population marginal means in mels).



Interestingly, the variance at utterance onset, measured as the average initial distance to the median (d_{init}) and denoted by the dashed ellipses in Fig. 2, also differed across conditions (two-way ANOVA, main effect of noise level, $F = 5.5$, $p = 0.004$). As shown in Fig. 4, the initial variance was greater in the high noise condition (blue dashed ellipse) than in the quiet condition (red dashed ellipse). Rather than the increased variance leading to an increase in centering, however, the opposite was true: centering was smallest in the high noise condition despite the large potential to decrease this variance over time.

Figure 4: Schematic of initial and mid-utterance variance (average distance to median over all peripheral trials), condensed from the three plots in Fig. 2.



Finally, we examined vowel duration to ensure that differences in centering were not attributable to duration effects. Average vowel duration varied two-fold across speakers, ranging from 144 ms to 341 ms (mean = 219 ms). We predicted that longer vowels would afford a speaker more time to correct deviant formants and bring them closer to median values. Vowel duration differed across conditions (two-way ANOVA, main effect of noise level, $F = 21.01$, $p < 0.0001$), with both noise conditions having

significantly greater durations than the quiet condition. This is perhaps unsurprising given that an increased vowel duration is an oft-replicated finding in studies of the Lombard effect [10], [14]. However, as with the increase in variance noted above, the increase in vowel duration in the noise conditions did not lead to more observed centering, but the reverse. Duration also failed to significantly correlate with centering on a per-subject basis (Pearson's $r = 0.427$, $p = 0.251$).

4. DISCUSSION

We have shown here that vowel formant centering is a robust phenomenon in which speakers' most peripheral productions are adjusted to become more central, even in utterances as short as 150 ms. Additionally, centering was found to be greater in quiet than in noisy conditions, giving support to the notion that auditory feedback monitoring contributes to these corrective vocal movements on an utterance-to-utterance basis. In this framework, feedback need not be externally manipulated to induce an error-like response; vowel centering, or at least a component of it, appears to act as natural "compensation" for internally-produced variation. The feedback monitoring process may then keep utterances on track, detecting and correcting nascent errors before they are made. Our recent work [6] provides further evidence that the auditory system plays a role in the error detection and correction process, demonstrating that centering is correlated with the magnitude of cortical error signals (auditory M100 evoked potential) that encode how much a production deviates from a vowel target. As these auditory cortical signals occur at a latency of less than 100 ms, preceding the centering in time, it is plausible that they may reflect a process that drives the corrective behavior.

Of course, centering was also present (albeit reduced) during noisy conditions in which little to no voice feedback was available. This finding is evidence that the centering is not purely due to auditory feedback, but also to non-auditory processes. For example, somatosensory feedback may contribute to the assessment of error when the tongue's position deviates from that expected to produce a target vowel sound. A third cause may involve the dynamics of the vocal motor system and its neural control signals [11]. Further investigation of centering and other dynamic acoustic changes in ongoing speech production will help to delineate these contributions.

5. ACKNOWLEDGEMENTS

We thank Chiara Bertolini and Laura Visentin for assistance in running subjects. This work was supported by NIH grants F32-DC011249 (C.A.N.) and R01-DC010145 (J.F.H.) and NSF grant BCS-0926196 (J.F.H).

6. REFERENCES

[1] Bauer, J. J., Mittal, J., Larson, C. R., Hain, T. C. 2006. Vocal responses to unanticipated perturbations in voice loudness feedback: An automatic mechanism for stabilizing voice amplitude. J. Acoust. Soc. Am. 119(4), 2363–2371.

[2] Burnett, T. A., Freedland, M. B., Larson, C. R., Hain, T. C. 1998. Voice F0 responses to manipulations in pitch feedback. *J. Acoust. Soc. Am.* 103(6), 3153–3161.

[3] Cai, S., Ghosh, S. S., Guenther, F. H., 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. *J. Neurosci.* 31(45), 16483 –16490.

[4] Heinks-Maldonado, T. H., Houde, J. F. 2005. Compensatory responses to brief perturbations of speech amplitude. *Acoust. Res. Lett. Online* 6(3), 131–137.

[5] Niziolek, C. A., Guenther, F. H. 2013. Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. *J. Neurosci.* 33(29), 12090–12098.

[6] Niziolek, C. A., Nagarajan, S. S., Houde, J. F. 2013. What does motor efference copy represent? Evidence from speech production. *J. Neurosci.* 33(41), 16110–16116.

[7] Niziolek, C. A., Houde, J. F. 2015. wave_viewer: First release. doi:10.5281/zenodo.13839

[8] O'Shaughnessy, D. 1987. *Speech communication: human and machine*. New York: Addison-Wesley Pub. Co.

[9] Patel, R., Niziolek, C., Reilly, K., Guenther, F. H. 2011. Prosodic adaptations to pitch perturbation in running Speech. J. Speech Lang. Hear. Res. 54(4), 1051–1059.

[10] Patel, R., Schell, K. W. 2008. The influence of linguistic content on the Lombard effect. *J. Speech Lang. Hear. Res*. 51(1), 209–220.

[11] Perrier, P., Ostry, D. J.1996. The equilibrium point hypothesis and its application to speech motor control. *J. Speech Hear. Res.* 39(2), 365–378.

[12] Shiller, D. M., Sato, M., Gracco, V. L., Baum, S. R. 2009. Perceptual recalibration of speech sounds following speech motor learning. J. Acoust. Soc. Am. 125(2), 1103–1113.

[13] Tourville, J. A., Reilly, K. J., Guenther, F. H. 2008. Neural mechanisms underlying auditory feedback control of speech. NeuroImage 39(3), 1429–1443.

[14] Van Summers, W., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., Stokes, M. A. 1988. Effects of noise on speech production: Acoustic and perceptual analyses. *J. Acoust. Soc. Am.* 84(3), 917–928.