# The role of vowel perceptual cues in compensatory responses to perturbations of speech auditory feedback

Kevin J. Reilly[a] and Kathleen E. Dougherty
*Department of Speech-Language Pathology and Audiology, Northeastern University, 360 Huntington Avenue, Boston, Massachusetts 02115*

(Received 26 December 2012; revised 3 June 2013; accepted 12 June 2013)

The perturbation of acoustic features in a speaker's auditory feedback elicits rapid compensatory responses that demonstrate the importance of auditory feedback for control of speech output. The current study investigated whether responses to a perturbation of speech auditory feedback vary depending on the importance of the perturbed feature to perception of the vowel being produced. Auditory feedback of speakers' first formant frequency (F1) was shifted upward by 130 mels in randomly selected trials during the speakers' production of consonant-vowel-consonant words containing either the vowel /Λ/ or the vowel /ɝ/. Although these vowels exhibit comparable F1 frequencies, the contribution of F1 to perception of /Λ/ is greater than its contribution to perception of /ɝ/. Compensation to the F1 perturbation was observed during production of both vowels, but compensatory responses during /Λ/ occurred at significantly shorter latencies and exhibited significantly larger magnitudes than compensatory responses during /ɝ/. The finding that perturbation of vowel F1 during /Λ/ and /ɝ/ yielded compensatory differences that mirrored the contributions of F1 to perception of these vowels indicates that some portion of feedback control is weighted toward monitoring and preservation of acoustic cues for speech perception.

© *2013 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4812763]

PACS number(s): 43.70.Mn, 43.70.Bk, 43.70.Aj [BHS] Pages: 1314–1323

## I. INTRODUCTION

Perturbations of speech auditory feedback cause rapid compensations in speech output that highlight the importance of auditory feedback during speech production. In these investigations, a speaker's auditory feedback is perturbed on certain trials by either increasing or decreasing the value of a specific acoustic feature (e.g., fundamental frequency, F0). These perturbations result in rapid compensations by the speaker to correct the perceived error. A compensatory response consists of a change in the production of the perturbed feature that is opposite in direction to that of the perturbation. These compensatory responses have been observed during perturbation of speech parameters such as fundamental frequency (Jones and Munhall, 2000; Larson *et al.*, 2000; Xu *et al.*, 2004; Chen *et al.*, 2007), speech intensity level (Siegel and Pick, 1974; Heinks-Maldonaldo and Houde, 2005; Bauer *et al.*, 2006), and vowel formant frequencies (Houde and Jordan, 1998; Purcell and Munhall, 2006; Villacorta *et al.*, 2007; Tourville *et al.*, 2008). The finding that response latencies for compensatory responses are quite short, occurring between 100 and 250 ms after the perturbation, indicates that the central nervous system monitors the accuracy of speech output on an ongoing basis.

Recent findings have demonstrated that speech responses to a particular auditory feedback perturbation are not uniform but are influenced by a number of factors related to speech context. For example, compensatory responses to F0 perturbations during sustained vowel productions have smaller magnitudes and longer latencies than compensatory responses observed during production of English sentences (Chen *et al.*, 2007) and lexical tones sequences (Xu *et al.*, 2004). At the same time, responses to F0 perturbations during sentence production are smaller in magnitude than those observed during singing (Natke *et al.*, 2003). Compensatory responses to F0 perturbations are also influenced by a speakers' intended F0 trajectory (Chen *et al.*, 2007). Chen and colleagues (2007) reported that speakers compensated more to a perturbation that was in the opposite direction of their planned F0 trajectory than to one that was in the same direction. Together these findings indicate that the magnitude of compensation to perturbation of F0 reflects not just the magnitude of the perturbation but also the goals of the intended utterance.

Factors influencing responses to perturbations of suprasegmental parameters in auditory feedback, such as F0, have been well-investigated, but little is known about factors that modulate responses to segmental perturbations of auditory feedback. One possible influence on segmental speech responses concerns the effects of the perturbation on perceptual cues for either identifying or discriminating the phoneme being perturbed. A role for perceptual processing in auditory feedback control is indicated by speech production studies that document the tight coupling between detailed aspects of speech production and desired perceptual outcomes. For example, Wright (2004) examined whether speakers implicitly modulate their production of vowels in consonant vowel consonant (CVC) words to account for listeners' perceptual difficulties identifying words with dense phonologic neighborhoods and low lexical frequencies

[a] Author to whom correspondence should be addressed. Electronic mail: k.reilly@neu.edu

(Luce and Pisoni, 1998). Wright (2004), and later Munson and Solomon (2004), found that speakers expanded their vowel formant space significantly during production of vowels in words with a dense phonologic neighborhood and low lexical frequency compared to words with a sparse phonologic neighborhood and high lexical frequency. The finding that increases in vowel space also increase speech intelligibility (Picheny *et al.*, 1986; Moon and Lindblom, 1994; Bradlow *et al.*, 1996) indicates that speakers in the Wright (Wright) and Munson and Solomon (2004) studies implicitly modulated their vowel productions to account for the adverse perceptual effects of lexical and phonologic factors on word intelligibility. In a similar study, Aylett and Turk (2006) examined the effects of language redundancy on speakers' vowel productions. Language redundancy is a measure of a syllable's predictability based on context and its frequency of occurrence in a language. These investigators found that the size of speakers' vowel spaces tended to increase as the redundancy, or predictability, of a syllable decreased.

In summary, these studies demonstrate that speakers account for the perceptual effects of lexical, phonologic, and contextual factors by modulating their production of formant frequencies in ways that either increase or decrease a vowel's spectral distinctiveness. The close coupling between fine-grained aspects of vowel production and factors affecting speech perception is consistent with the idea that perceptual outcomes are explicitly parameterized in the motor commands for speech production (Lindblom, 1990, 1996; Schwartz *et al.*, 1997).

These findings strongly suggest that perceptual outcomes are encoded in the feedforward control system for speech and raise the possibility that perceptual outcomes are similarly represented in the feedback control system such that feedback control is weighted toward the monitoring and preservation of acoustic cues for achieving a desired perceptual outcome. To address this question, the current study evaluated whether perturbation of a speech feature in contexts that affect vowel identification or discrimination elicits a larger compensatory response than the same perturbation in contexts that do not have an appreciable effect on vowel identification or discrimination. In the current study, a perturbation of vowel F1 was applied to speakers' auditory feedback during randomly selected trials in production of CVC words containing either the vowel, /Λ/, or the vowel, /ɝ/. These vowels were selected because their production is characterized by similar mid, central places of articulation and comparable F1 and F2 frequencies, but their perception is dependent on quite different acoustic features. Perception of the vowel /Λ/, like most English vowels, can be mainly characterized in terms of the spectral information carried by F1 and F2 (Peterson and Barney, 1952). However, the vowel /ɝ/ is unique among English vowels in that its perception derives largely from the low frequency location of its F3 and/or the proximity of its F3 and F2 frequencies (Lehiste and Peterson, 1959; Singh and Woods, 1971; Stevens, 1998; Heselwood and Plug, 2011). For example, Lehiste and Peterson (1959) found that listeners identified low-pass filtered productions of /ɝ/ with 0% accuracy when the filter eliminated energy at F2 and F3 and only passed energy at F1. In contrast, listeners identified high-pass filtered productions of the vowel /ɝ/ with 60% accuracy when the cutoff frequency was located between F1 and F2 and eliminated all or nearly all of the energy associated with F1. By comparison, listeners' recognition scores for the vowel /Λ/ were 0% at the same high-pass filter setting. Similar results have been observed in studies of vowel perceptual confusions and dissimilarity. Singh and Woods (1971) evaluated listeners' judgments of vowel dissimilarity using multidimensional scaling and found that a two dimensional articulatory configuration describing differences in F1 and F2 was sufficient to account for dissimilarity judgments in the set of American English vowels excluding /ɝ/. When the vowel set was expanded to include /ɝ/, an additional dimension was needed to account for the low F3 frequency associated with retroflexion. Similarly, Wilson and Bond (1977) found that an additional dimension was needed to account for vowel perception confusions when data for vowel /ɝ/ were included in the dataset. These findings indicate that, relative to other English vowels, perceptual identification and discrimination of /ɝ/ involves a greater contribution from F3 and F2 and a reduced contribution from F1.

In summary, the current study analyzed speech responses to F1 perturbations to determine whether compensation magnitudes and latencies were modulated in ways that reflect the importance of F1 as a vowel perceptual cue. It was predicted that because identification and discrimination of the vowel /ɝ/ is less dependent on the frequency location of F1 (Lehiste and Peterson, 1959; Singh and Woods, 1971; Wilson and Bond, 1977), compensatory responses during production of this vowel would have smaller magnitudes and longer latencies than compensatory responses during production /Λ/.

## II. MATERIALS AND METHODS

### A. Participants

Subjects for this study were 10 females (mean = 26.5 yr, SD = 7.3 yr) and 7 males (mean = 25.6 yr, SD = 8.6 yr). All subjects were native speakers of English with no reported history of speech, language, or hearing disorders.

### B. Speech stimuli

Speech stimuli consisted of four pairs of CVC words. In each pair, the words were identical except one of the words contained the vowel /Λ/, and the other, the vowel /ɝ/. The speech stimuli are listed in Table I. Participants were asked to prolong the vowel during production of each word to support the use of auditory feedback. Production was modeled for the participants by playing pre-recorded productions of the stimuli with the desired amount of vowel prolongation. Participants then practiced prolonging the vowel in each word before beginning the study.

### C. Experimental protocol

Subjects were seated in a sound treated booth (Acoustic Systems, Model RE-147 S) with visual access to a computer

TABLE I. The list of speech stimuli used in the current study. Pairs of stimuli were phonemically identical except that one word in each pair contained the vowel /ʌ/ (left column) and the other word contained the vowel /ɝ/ (right column)

| /ʌ/ words | /ɝ/ words |
|---|---|
| Puck | Perk |
| Bud | Bird |
| Cut | Kurt |
| Cub | Curb |

monitor that displayed the speech stimulus for each trial. A head-worn directional microphone (AKG model C520) was placed at a fixed distance of approximately 5.5 cm from the speaker's lips. Microphone signals were amplified using a Mackie VLZ3 mixer/preamp and sent to an external sound card (Delta 44, M-Audio; digital sampling rate = 12 000 Hz) and processed by PC-based digital signal processing software package (see Sec. II D). The output of the processed audio signal was sent from the external sound card to the speaker via calibrated, noise-isolating insert earphones (ER4 microPro earphones, Etymotic Research). The gain of the feedback signal relative to the microphone input level was approximately 20 dB sound pressure level (SPL). The total processing delay for this setup was ~15 ms. Following each trial, a copy of the speaker's microphone signal and the speech auditory feedback signal were saved to the computer's hard drive.

The experimental protocol consisted of four runs of 72 trials. A run contained nine presentations of each of the eight speech stimuli. The presentation order of the stimuli was permuted randomly within each run except that the same stimulus was not presented on consecutive trials and no stimuli that rhymed (e.g., "cut" and "putt") were presented on consecutive trials. On two of the nine productions of a speech stimulus in a run, a perturbation was applied to the speech auditory feedback signal. The magnitude of the applied perturbation was specified in mel units, not hertz, to control for the potential effects of between-vowel F1 differences on perception of perturbation magnitude during /ʌ/ and /ɝ/. A perturbation magnitude of 130 mels was used as this value corresponds closely to the magnitude of the hertz-based perturbation used in a previous study (Tourville *et al.*, 2008) that yielded reliable compensatory responses during production of /ɛ/. As mentioned in the preceding text, the perturbation was applied on random trials within a run but not on consecutive trials and not on consecutive presentations of a particular speech stimulus. As a result, each participant produced a total of 288 utterances. Each stimulus word was produced 36 times, and eight of those productions were perturbed.

After completing the study, a subset of participants ($n = 11$) was questioned to determine whether they perceived anything unusual about the auditory feedback they heard or if they perceived any changes in feedback of the course of the study. These questions were used to provide information about subjects' awareness of the perturbations to their auditory feedback.

## D. Apparatus

Formant tracking and formant perturbation were accomplished using a MATLAB Mex-based software package, AUDAPTER, developed by Cai and colleagues (Cai *et al.*, 2008; Cai *et al.*, 2010). This method uses linear predictive coding (LPC) analysis in conjunction with cepstral liftering and dynamic programming (Xia and Espy-Wilson, 2000) to track speakers' F1, F2, and F3 in 14 ms time intervals. The order of the LPC analysis in this study was set to 11 and 12 for female and male speakers, respectively. Perturbation of a speaker's F1 was accomplished via LPC filtering using pole-pair substitution in the $z$ plane. Formant tracking and, for perturbed trials, formant perturbations were initiated and terminated based on two short-term root mean square (rms) measures: A smoothed block-by-block trace of rms amplitudes derived from the unfiltered speech signal ($rms\_s$) and a smoothed block-by-block trace of rms amplitudes derived from the pre-emphasized speech signal ($rms\_p$). Tracking and perturbation of formants were carried out when instantaneous values for $rms\_s$ and $rms\_s/rms\_p$ exceeded pre-determined threshold values. The threshold values were set to correspond to the onset and offset of voicing for the target vowel. The threshold values for these measures were derived from a pilot study involving four participants (two male and two female) who produced 10 sets of the speech stimuli while receiving auditory feedback of their speech at the levels described.

## E. Pre-processing of speech acoustic output

A graphical user interface (GUI) was developed in MATLAB (MATLAB, 2011) to process the speech acoustic signals for each trial. A plot was created that displayed the microphone signal (Fig. 1, top panel), formants 1–3 from the AUDAPTER software (Fig. 1, middle panel), and a spectrogram of the utterance overlaid with formants 1–3 from the AUDAPTER software (Fig. 1, bottom panel). The GUI allowed playback of both the microphone signal and the feedback signal to verify that the speaker produced the correct word on each trial. The information displayed in each panel was used to check the accuracy of the derived vowel onsets and offsets and corrections to these values were applied as needed. Vowel durations were derived by subtracting vowel offsets from the corresponding vowel onsets. Overt errors in the tracking of formants by the AUDAPTER software were identified by inspection of the spectrogram and overlaid formant contours in the bottom panel of the figure. This information was also used to identify trials when the formant tracking was either initiated or terminated during the production of the vowel. Errors in the accuracy or timing of formant tracking were present in 8% of the trials, and these trials were excluded from further analysis. Approximately 60% of these excluded trials were discarded due to errors in distinguishing F2 from F3 during production of /ɝ/.

## F. Across-speaker analysis of compensations for /ʌ/ and /ɝ/

Speakers' F1, F2, and F3 frequency contours were converted to mels and smoothed using a 41.3 ms Hamming

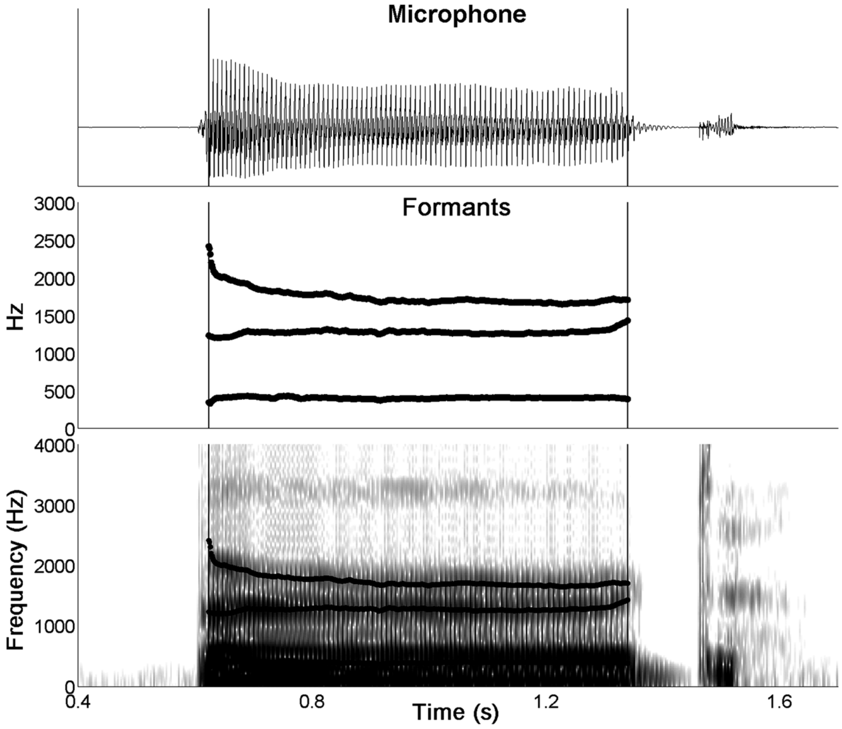

FIG. 1. An example of the graphical user interface for processing acoustic data streams for each trial. The display includes the microphone signal (top panel), formants 1–3 (middle panel) and the spectrogram with overlaid formants during a speaker's production of the word, *bird* (bottom panel).

window. Each speaker's formant contours during no-perturbation trials were averaged by formant number (F1, F2, and F3) and by vowel (/Λ/ vs /ɝ/). The mean no-perturbation formant contours were then subtracted from a speaker's corresponding F1, F2, and F3 contours during perturbed trials and averaged. The resulting formant *response contours* constituted a time-dependent measure of each speaker's average response to the perturbation by formant number and by vowel. As a result, six response contours were derived for each speaker; one response contour for each formant in the vowels /Λ/ and /ɝ/. Calculation of each response contour was restricted to time points that fell within the 80th percentile of all vowel lengths for a speaker to ensure that a sufficient number of samples were used to derive a response mean at each time point.

To determine the magnitude and latencies of compensatory responses, the response contours for each formant in each vowel were analyzed for significant non-zero differences in response to the F1 perturbation. One-sample *t*-tests were performed at each time point in the response contours for each vowel and formant. A compensation response to the F1 perturbation was defined as a significant non-zero difference ($p < 0.05$) at a particular time point and all subsequent time points. To prevent the spurious inclusion of very brief changes occurring at the end of the utterance, the time interval spanned by the non-zero differences was required to have a duration of at least 100 ms. The latency of a compensatory response corresponded to the time point associated with the first significant non-zero difference. The magnitude of a compensatory response was determined by averaging response contours and selecting the peak value occurring after the compensation latency.

## G. Within-speaker analysis of compensations for /Λ/ and /ɝ/

To evaluate whether speakers responded to the perturbation differently depending on the vowel being produced, each of a speaker's three response contours for /ɝ/ were subtracted from their corresponding response contours for /Λ/. The result was a set of three *difference contours* for each speaker. Each difference contour quantified that speaker's /Λ/ - /ɝ/ response contour difference for each formant. One-sample *t*-tests were performed at each time point in the difference contours to test for vowel-dependent differences in compensation magnitude and evaluate whether speakers exhibited greater compensation to the F1 perturbation during production of /Λ/ than /ɝ/. Within-speaker differences in the latencies of compensation for /Λ/ and /ɝ/ were assessed using the methods described by Tourville and colleagues (2008). Piecewise non-linear models were fit to individual speaker's F1 response contours for /Λ/ and /ɝ/. The non-linear model consisted of a constant segment describing no compensation that was followed by a logistic function describing compensation onset and magnitude. Parameter estimates for the latencies of compensation for /Λ/ and /ɝ/ were averaged across subjects. Confidence intervals of the latencies were estimated using a bootstrapping procedure that consisted of 1000 resamples of the F1 response contours with replacement for each vowel. Each pair of resampled data points represented an estimated average compensation latencies for /Λ/ and /ɝ/. Paired *t*-tests were used to evaluate statistical differences in the resulting 1000 × 2 array of latency estimates for each vowel.

## H. Average formant frequencies for /Λ/ and /ɝ/

An additional analysis was performed on non-perturbed trials to derive the mean formant frequencies for /Λ/ and /ɝ/

produced by each speaker. The vowel portion of each trial was extracted, pre-emphasized, and convolved with a Hamming window. LPC was then used to derive the vowel amplitude spectrum. The order of the LPC analysis was adjusted to provide the best fit for a given speaker and ranged from 12 to 14 for male speakers and from 11 to 12 for female speakers. A peak-picking method was used to identify the first three formant frequencies.

## III. RESULTS

### A. Vowel formant frequencies and durations

The mean F1, F2, and F3 frequencies produced by each speaker during no-perturbation trials were derived for /Λ/ and /ɝ/. Figure 2 displays the group average and 95% confidence intervals for each formant by vowel. The formant frequencies observed for /Λ/ and /ɝ/ in this study were comparable to those reported in previous studies (Peterson and Barney, 1952; Hillenbrand *et al.*, 1995); paired *t*-tests were used to evaluate between-vowel differences for each of the three formants. To control for the use of multiple *t*-tests in this analysis, the *p* value for determining statistical significance was set to $0.05/3 = 0.0167$. This analysis identified significant within-speaker differences in F1 [$t(16) = 8.06$, $p < 0.001$] with speakers producing higher F1 frequencies during /Λ/ by an average of 131 Hz. Speakers' F3 during /Λ/ was also significantly higher than their F3 during /ɝ/ [$t(16) = 11.35$, $p < 0.001$] and the average F3 difference was 767 Hz. Last, speakers' average F2 for /Λ/ was significantly lower their average F2 for /ɝ/ [$t(16) = -3.07$, $p = 0.007$] by an average of approximately 100 Hz.

The vowel formant differences observed in this study are generally consistent with those of previous studies (Peterson and Barney, 1952; Hillenbrand *et al.*, 1995). As expected, speakers' production of /ɝ/ was characterized by a much lower F3 than their production of /Λ/. The finding that F1 was lower for /ɝ/ than /Λ/ has also been reported previously (Peterson and Barney, 1952; Hillenbrand *et al.*, 1995). The magnitude of the F1 and F2 differences were considerably smaller than that observed for F3; the implications of these F1 and F2 differences on the current findings are addressed in Sec. IV.

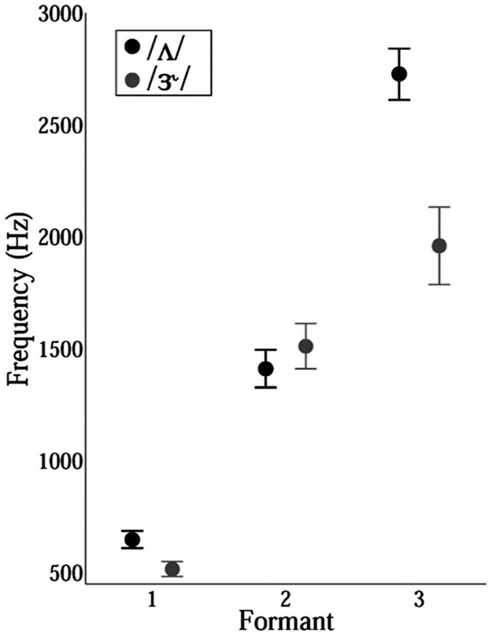

FIG. 2. Mean and 95% confidence intervals of speakers' average frequencies for formants 1–3 during no-perturbation productions of /Λ/ (black) and /ɝ/ (gray).

Speakers' average vowel durations for /Λ/ was 0.675 s (SD = 0.255 s) and 0.729 (SD = 0.248 s) for /ɝ/. A paired *t*-test of speakers' vowel durations revealed significant within-speaker differences in vowel durations between /Λ/ and /ɝ/ [$t(16) = 5.29$, $p < 0.001$]. On average, speakers' /ɝ/ durations were significantly longer than their /Λ/ durations by 54 ms (SD = 42 ms).

### B. Formant response contours

F1, F2, and F3 response contours were analyzed for each vowel to assess the effects of the F1 perturbation on speakers' production those formants during /Λ/ and /ɝ/. Figure 3 displays the mean and 95% confidence intervals of speakers' response contours by increasing formant number for /Λ/ (left panels) and /ɝ/ (right panels). One-sample *t*-tests were performed on the set of responses at each time point for a given formant. In this analysis, non-zero changes in the response contours of a particular formant indicated that speakers altered their production of that formant during perturbed trials compared to no-perturbation trials. This analysis revealed a decrease in speakers' F1 average response contours for /Λ/ that reached significance starting 173 ms (shown by the dotted line) after the onset of voicing (Fig. 3, top left panel). The magnitude of this decrease was 18.6 mels. Significant non-zero differences were not detected in either the F2 or F3 response contours for /Λ/, which indicates that the F1 perturbation did not lead to altered production of these formants.

Analysis of the formant response contours for /ɝ/ yielded similar results. A significant decrease was observed in the response contour for F1 (Fig. 3, top right panel). The latency of this decrease was 254 ms, and the magnitude of this decrease was 7.1 mels. No significant changes were detected in either the F2 or F3 response contours for /ɝ/. The finding that significant decreases in F1 were observed for both vowels during an upward perturbation of their F1 is consistent with the findings of a number studies demonstrating that speakers compensate for a perturbation of auditory feedback by altering their production of the perturbed feature in the direction opposite to the perturbation. In addition, the finding that compensation was present for F1, and was not observed for either F2 or F3, is consistent with previous findings that indicate compensation is specific to the acoustic feature being perturbed (Houde and Jordan, 2002; Villacorta *et al.*, 2007; Tourville *et al.*, 2008; Cai *et al.*, 2010) but see MacDonald and colleagues (2011) for an exception.

### C. Formant difference contours

Within-speaker analyses were carried out using speakers' difference contours to compare the magnitude and

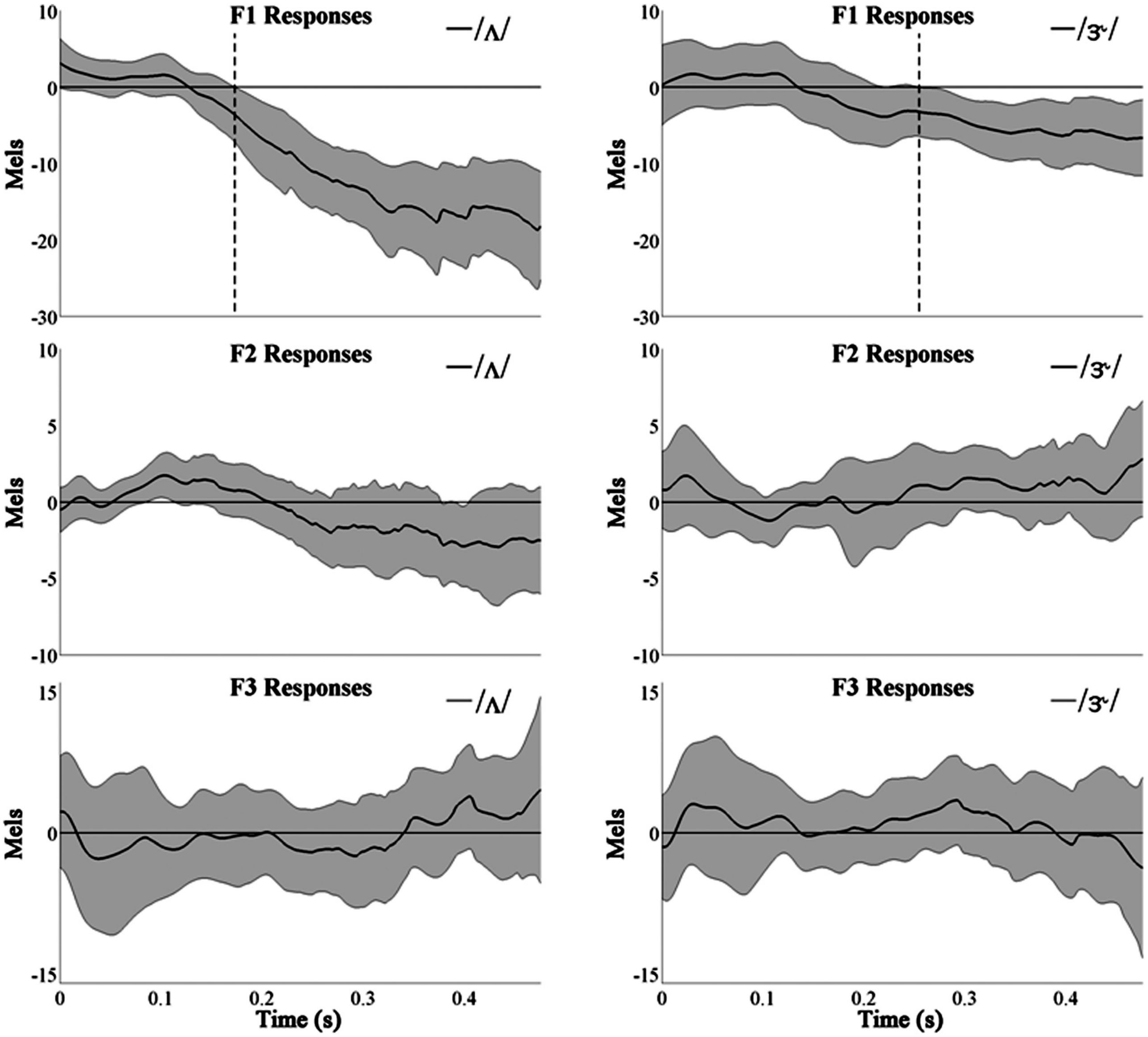

FIG. 3. Mean and 95% confidence intervals of speakers' responses to the F1 perturbation. Response contours for each formant are plotted as a function of time. The panels in the left column show response contours during production of /Λ/, and the panels in the right column show response contours during production of /ɝ/. Within each column, the response contours are displayed in order of increasing formant number. The onsets of significant responses to the F1 perturbation are depicted with dashed vertical lines when present.

latency of speakers' compensation during /Λ/ vs /ɝ/. One-sample *t*-tests were performed on the set of speakers' difference contours at each formant. In this analysis, nonzero difference contours for a particular formant indicated differences in the magnitude of compensation between /Λ/ vs /ɝ/. Specifically, a significant negative deflection in the difference contours indicated that compensation was greater during perturbation of /Λ/ and a significant positive deflection indicated that compensation was greater during perturbation of /ɝ/. Figure 4 displays the mean and 95% confidence intervals of speakers' response differences contours for F1 (top panel), F2 (middle panel), and F3 (bottom panel). As indicated in this figure (top panel), a significant decrease was present in the difference contours for F1 indicating that compensation to the F1 perturbation was greater during /Λ/ than during /ɝ/. The average latency at which the compensation for /Λ/ deviated significantly from the compensation for /ɝ/ was 220 ms and the average magnitude of the deviation was 12.1 mels.

### D. F1 compensation latencies for /Λ/ and /ɝ/

Using the bootstrapping procedure described in Sec. II, a $1000 \times 2$ array of estimated average compensation latencies was derived; one column for each vowel. A paired *t*-test was performed comparing the array of estimated latencies for each vowel to test for significant difference in the compensation latencies of speakers' F1 compensatory responses for /Λ/ vs /ɝ/. This analysis revealed that the onset of F1 compensation for /Λ/ occurred at a significantly shorter latency than the onset of F1 compensation for /ɝ/, $t(999) = -27.23$, $p < 0.05$. The average latency difference was 28 ms.

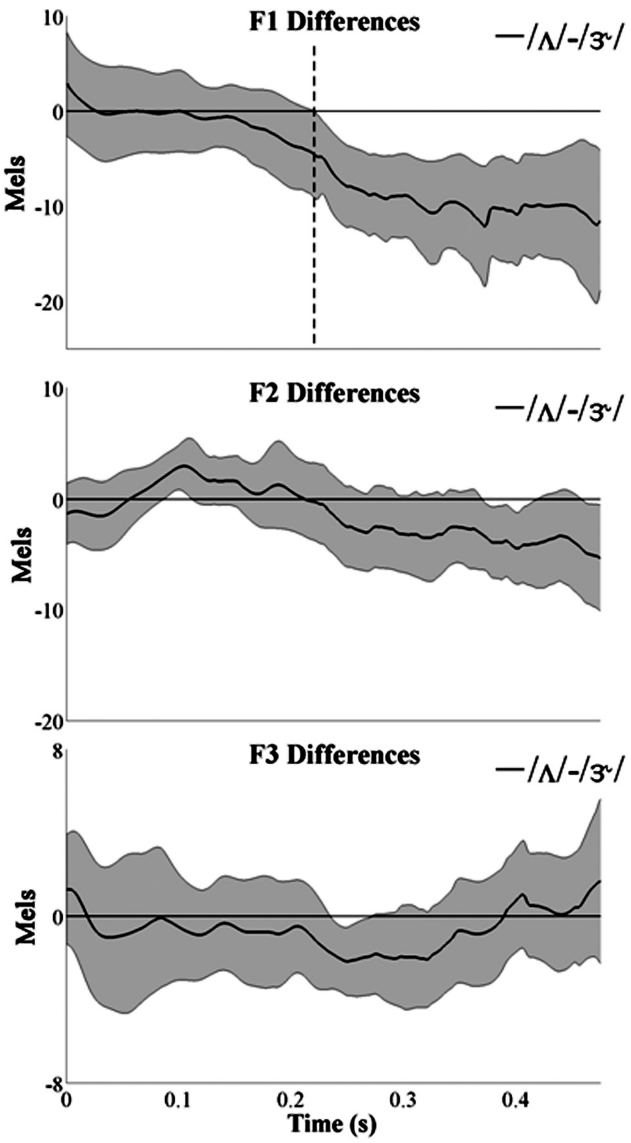

FIG. 4. Mean and 95% confidence intervals of the contours showing the response difference (/Λ/ – /ɝ/) for each formant. Difference contours for F1 are displayed in the top panel, difference contours for F2 are shown in the middle panel, and difference contours for F3 are shown in the bottom panel.

### E. Perceived changes to auditory feedback

Following the study, a subset of speakers was questioned about their perception of the auditory feedback. These speakers did not report awareness of anything unusual about the auditory feedback presented by the earphones. In addition, the speakers did not report changes in the auditory feedback over the course of the study.

## IV. DISCUSSION

The current study investigated auditory feedback control of vowel production and the extent to which this control is organized around the realization of perceptually important features in speech acoustic output. Speakers' F1 was randomly perturbed upward by 130 mels during vowels produced in CVC words containing either the vowel /Λ/, for which F1 is an important perceptual cue, or /ɝ/, for which F1 is a less important perceptual cue. The perturbation resulted in compensatory decreases in F1 during both vowels. However, compensatory responses for /Λ/ exhibited significantly larger magnitudes and significantly shorter latencies than compensatory responses for /ɝ/. These findings suggest that compensatory responses were modulated by vowel perceptual cues such that larger compensations were produced when F1 contributed more to the perception of the perturbed vowel than when F1 contributed less to the perception of the perturbed vowel.

The magnitude of the mel-based perturbation in the current study was similar to the F1 perturbation used by Tourville and colleagues (2008) on random trials of speakers' productions of the vowel /ɛ/ in different CVC words. In this latter study, the perturbation of /ɛ/ was achieved by either increasing or decreasing instantaneous F1 values in hertz by 30%. The results of the Tourville *et al.* study (2008) revealed average percentage compensations of 13.6% and 13.0% to upward and downward F1 shifts, respectively. In the current study, speakers compensated an average of 18.6 mels (14% of the applied perturbation) during /Λ/ and 7.1 mels (4.7% of the applied perturbation) during /ɝ/. A comparison of the findings from the two studies indicates that the compensation differences between /Λ/ and /ɝ/ in the current study were not due to unusually high compensation magnitudes for /Λ/ but rather were due to low compensation magnitudes for /ɝ/. This observation is consistent with the idea that the reduced compensation during /ɝ/ reflected the reduced contribution of F1 to perception of /ɝ/.

Context-dependent modulation of responses to perturbations of sensory feedback is a hallmark feature of physiologic control systems (Prochazka *et al.*, 2000). Modulation of speech responses to perturbations of sensory feedback has been well-documented during auditory feedback perturbations of f0 (Natke *et al.*, 2003; Xu *et al.*, 2004; Chen *et al.*, 2007) and was reported even earlier in the somatosensory modality (Kelso *et al.*, 1984). In one study, Kelso and colleagues (1984) observed that unexpected force loads applied to the jaw during the closing movement for /b/ elicited rapid compensation responses that included downward movement of upper lip to achieve labial closure. However, upper lip compensations were not observed when the same perturbation was unexpectedly applied during the closing movement for /z/, which does not require labial closure. These findings indicated that upper lip responses to the unexpected force loads were modulated depending on the importance of the upper lip to production of the sound segment being perturbed. The current findings are similar to those of Kelso *et al.* (1984) in that F1 responses to the auditory feedback perturbation were modulated depending on the importance of the perturbed feature to perception of the vowel being produced.

Of relevance to the issue of auditory and somatosensory perturbations is an adaptation study by Feng and colleagues (2011), who evaluated responses to auditory and somatosensory perturbations delivered alone and simultaneously. The investigators observed that speakers demonstrated the expected adaptation to the auditory alone perturbation, the somatosensory alone perturbation, and simultaneous perturbations of auditory and somatosensory that were compatible (i.e., adaptation to the somatosensory perturbation decreased the auditory error). However, when simultaneous perturbations were incompatible (i.e., adaptation to the somatosensory perturbation increased the auditory error), the investigators observed that speakers only adapted to the auditory perturbation and did not adapt to the somatosensory perturbation. The authors suggested that auditory feedback is prioritized over somatosensory feedback; this is consistent with the findings of the current study that highlight the importance of auditory perceptual outcomes in feedback control for speech.

Contemporary accounts of auditory feedback control for speech posit that the magnitude and direction of compensation responses derive from a central error signal that reflects the difference between actual auditory feedback and an efference copy of motor commands that describes expected auditory feedback (Guenther *et al.*, 2006; Tourville *et al.*, 2008; Ventura *et al.*, 2009; Hickok *et al.*, 2011; Houde and Nagarajan, 2011; Guenther and Vladusich, 2012). The modulation of compensation magnitudes and latencies during /Λ/ vs /ɝ/suggests that a portion of this error signal reflects auditory feedback monitoring of perceptually relevant features. These features could relate to either vowel identification or vowel discrimination. For example, the magnitude of the error signal may increase when the discrepancy between predicted and actual feedback includes features that are important for identification of a target vowel. As F1 makes a smaller contribution to /ɝ/ identification (Lehiste and Peterson, 1959), perturbation of F1 during this vowel would elicit a smaller error signal and, as a result, a smaller compensatory response. Alternatively, the magnitude of the error signal may be sensitive to discrepancies between predicted and actual feedback that affect vowel discrimination. In the current study, the F1 perturbation would have reduced the spectral contrast distance between /Λ/ and /a/ but would have had less of an effect on the contrast distance between /ɝ/ and other vowels because F1 is less important for discriminating /ɝ/ from neighboring vowels (Singh and Woods, 1971). The effects of the perturbation on vowel discrimination could account for compensation differences between /Λ/ and /ɝ/ if the error signal is modulated by the perturbation's influence on features that contrast a target vowel from neighboring vowels. As the design of the current study did not allow for the separate evaluation of perceptual effects related to identification vs discrimination, it is not possible to speculate whether this specific aspect of speech perception was prioritized by speakers in the current study.

A potentially confounding factor in the interpretation of these findings concerns the between-vowel F1 differences observed in this study. Although F1 frequencies for both /Λ/ and /ɝ/ lie in the midrange of F1 frequencies for English vowels, speakers' F1 frequencies for /Λ/ were significantly higher than their F1 frequencies for /ɝ/. To control for the effects of between-vowel differences in F1 on perturbation magnitude, the perturbations in this experiment were expressed in mel units, which scale nonlinearly with hertz and more closely correspond to the psychophysical properties of the human auditory system (Stevens and Volkmann, 1940). In the current study, the average F1 frequencies for /Λ/ and /ɝ/ were 647 and 516 Hz, respectively. Given these values, a perturbation of 130 mels produced an average F1 increase in Hz of 125% during /Λ/ and an average F1 increase of 129% during /ɝ/. The finding that the perturbation percentage differences between /Λ/ and /ɝ/ were so small suggests that between-vowel differences in F1 did not contribute to the compensation differences observed in the current study. It is also unlikely that the small F2 differences contributed to the current findings.

Similarly, the significantly longer vowel durations observed for /ɝ/ may have affected the measurement of responses for this vowel because formant data near the end of longer productions of /ɝ/ would have been excluded from analysis. However, several aspects of the data suggest that this was not the case. First, the length of the analysis window for evaluating speakers' responses was 500 ms, which is a comparatively long time window for evaluating compensatory responses [e.g., Tourville and colleagues (2008) used a window size of 250 ms]. As a result, there was more than sufficient time for speakers' to demonstrate larger compensatory responses for /ɝ/. In addition, the slope of the mean response contour for /ɝ/ was quite shallow and does not suggest that extending the window by tens of milliseconds would have yielded a substantially different finding for this vowel. Last, formant contours at the end of vowels tended to reflect the transition into the final consonant, and it is difficult to imagine how these formant transitions would have increased speaker's compensation during for /ɝ/.

Another possible explanation for the reduced F1 compensation during /ɝ/ concerns the vocal tract configurations for this vowel and their sensitivity to modulation of F1. The finding that significant F1decreases were observed during perturbation of /ɝ/ indicates that modulation of F1 was feasible, but it still remains possible that vocal tract adjustments for modulating F1 were constrained during /ɝ/ compared to /Λ/. In their analysis of MRI images, Espy-Wilson and colleagues (Espy-Wilson *et al.*, 2000; Zhang *et al.*, 2005; Zhou *et al.*, 2008) reported that the low F3 frequency associated with rhotic phonemes such as /ɝ/ and /r/ reflects a front cavity resonance that is bounded by the lips and teeth anteriorly, a palatal constriction posteriorly, and includes a large sublingual volume. By contrast, the frequency of F1 for /ɝ/ arises from the geometry of the palatal constriction and/or the cavity posterior to this constriction. Based on these data, it would seem that there was at least one mechanism for lowering F1, and this would have involved reducing tongue height to increase the cross-sectional area at the palatal constriction. Adjustments of tongue height constitute the primary means of modulating F1 across the vowel space (Stevens, 1989, 1998) and tongue lowering would have likely contributed to some portion of the compensatory decreases in F1 observed during /Λ/. As a result, lowering of the tongue during /ɝ/ perturbation seems capable of producing decreases in F1 that would be comparable to those observed during /Λ/ perturbation. At the same time, lowering of the tongue in the oral cavity would reduce the volume of the sublingual cavity and increase F3 as well as the difference between F3 and F2, which are the primary cues for perception of /ɝ/ (Lehiste and Peterson, 1959; Singh and Woods, 1971; Stevens, 1998;

Heselwood and Plug, 2011). In summary, it is likely that at least one solution for producing larger F1 compensations during /ɝ/ was available but this solution was not used because it might have compromised the perception of this vowel.

The 11 participants who completed the post-interview questionnaire did not report any awareness of the F1 perturbation to either vowel. This finding is consistent with speaker reports in other auditory perturbation studies (Houde and Jordan, 2002; Tourville *et al.*, 2008) and suggests that compensation differences observed in the current study arose from perceptual processing of the F1 perturbation that was largely implicit in nature.

The role of perceptual processes in speech responses to auditory feedback perturbations has been addressed previously. For example, Villacorta and colleagues (2007) observed a significant correlation between speakers' compensations to persistent auditory feedback perturbations of F1 and their auditory acuity in a vowel F1 auditory discrimination task. The findings of the study by Villacorta and colleagues (2007) support a role of perceptual processes in speech compensation responses. However, the speaker-specific differences in auditory acuity described by these investigators are different from the perceptual effects described in the current study that relate to the perceptual structure of the shared vowel system across speakers and listeners.

At least one other study has investigated the influence of a vowel system's perceptual structure on speech responses to vowel auditory feedback perturbations. Mitsuya and colleagues (2011) investigated cross-language differences in compensation to formant perturbations of auditory feedback. These investigators observed differences in compensation magnitudes that reflected differences in the distribution of vowels in English and Japanese. Specifically, an upward perturbation of F1 that "pushed" vowel formants toward an adjacent vowel in English, but not in Japanese, yielded greater compensation by English speakers than Japanese speakers. In contrast, speakers from the two language groups produced comparable compensations to a downward perturbation of F1 that "pushed" vowel formants toward an adjacent vowel in both languages. These findings are consistent with the idea that compensation responses are modulated by the importance of the perturbed feature to perception of the vowel being perturbed.

The notion that auditory feedback control for speech is weighted toward perceptually relevant features is also indicated by the findings of Perkell and colleagues (2007). These investigators studied speakers' productions of vowels in quiet and in the presence of different levels of background noise and found that at low and moderate noise to signal ratios, speakers increased the distinctiveness of spoken vowels by increasing vowel F1/F2 contrast distances. The association between vowel contrast distance and speech intelligibility (Picheny *et al.*, 1986; Moon and Lindblom, 1994; Bradlow *et al.*, 1996) indicates that speakers in the Perkell *et al.* study (Perkell *et al.*, 2007) increased the spectral distinctiveness of spoken vowels to offset the effects of background noise on the perception of speech output.

In summary, the differences in compensation magnitude and latencies observed in the current study mirror the different contributions of F1 to perception of /ʌ/ and /ɝ/. This finding indicates that that some portion of feedback control is weighted toward the monitoring and preservation of acoustic cues relevant for speech perception. F1 makes a non-zero contribution to perception of /ɝ/, but this contribution is smaller than its contribution to perception of /ʌ/. Similarly, perturbation of F1 during /ɝ/ elicited significant non-zero compensatory responses but the magnitudes of these responses were smaller and their latencies were longer when compared to the compensatory responses during /ʌ/. The question of whether the current findings involved perceptual processing related to vowel identification, vowel discrimination, or both is not known, and a follow-up investigation is underway to address this issue.

## ACKNOWLEDGMENTS

This research was supported by Grant No. DC011159 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

Aylett, M., and Turk, A. (**2006**). "Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei," J. Acoust. Soc. Am. **119**, 3048–3058.

Bauer, J. J., Mittal, J., Larson, C. R., and 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**, 2363–2371.

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech. I: Global and fine-grained acoustic-phonetic talker characteristics," Speech Commun. **20**, 255–272.

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (**2008**). "A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/," in *8th International Seminar on Speech Production*, edited by R. Sock, S. Fuchs, and Y. Laprie (INRIA, Strasbourg, France), pp. 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," J. Acoust. Soc. Am. **128**, 2033–2048.

Chen, S. H., Liu, H. J., Xu, Y., and Larson, C. R. (**2007**). "Voice F-0 responses to pitch-shifted voice feedback during English speech," J. Acoust. Soc. Am. **121**, 1157–1163.

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., and Alwan, A. (**2000**). "Acoustic modeling of American English /r/," J. Acoust. Soc. Am. **108**, 343–356.

Feng, Y. Q., Gracco, V. L., and Max, L. (**2011**). "Integration of auditory and somatosensory error signals in the neural control of speech movements," J. Neurophysiol. **106**, 667–679.

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (**2006**). "Neural modeling and imaging of the cortical interactions underlying syllable production," Brain Lang. **96**, 280–301.

Guenther, F. H., and Vladusich, T. (**2012**). "A neural theory of speech acquisition and production," J. Neurolinguist. **25**, 408–422.

Heinks-Maldonaldo, T. H., and Houde, J. F. (**2005**). "Compensatory responses to brief perturbations of speech amplitude," ARLO **6**, 131–137.

Heselwood, B., and Plug, L. (**2011**). "The role of F2 and F3 in the perception of rhoticity: Evidence from listening experiments," in *Proceedings of the 17th International Congress of Phonetic Sciences*, August 17-21, Hong Kong, pp. 867–870.

Hickok, G., Houde, J., and Rong, F. (**2011**). "Sensorimotor integration in speech processing: Computational basis and neural organization," Neuron **69**, 407–422.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Houde, J. F., and Jordan, M. I. (**1998**). "Sensorimotor adaptation in speech production," Science **279**, 1213–1216.

Houde, J. F., and Jordan, M. I. (**2002**). "Sensorimotor adaptation of speech I: Compensation and adaptation," J. Speech Lang. Hear. Res. **45**, 295–310.

Houde, J. F., and Nagarajan, S. S. (**2011**). "Speech production as state feedback control," Front. Hum. Neurosci. **5**, 82.

Jones, J. A., and Munhall, K. G. (**2000**). "Perceptual calibration of F0 production: Evidence from feedback perturbation," J. Acoust. Soc. Am. **108**, 1246–1251.

Kelso, J. A. S., Vatikiotis-Bateson, E., Tuller, B., and Fowler, C. A. (**1984**). "Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures," J. Exp. Psychol.: Hum. Percept. Perform. **10**, 812–832.

Larson, C. R., Burnett, T. A., Kiran, S., and Hain, T. C. (**2000**). "Effects of pitch-shift velocity on voice Fo responses," J. Acoust. Soc. Am. **107**, 559–564.

Lehiste, I., and Peterson, G. E. (**1959**). "The identifiability of filtered vowels," Phonetica **4** 161–177.

Lindblom, B. (**1990**). "Explaining phonetic variation: A sketch of the H&H theory," in *Speech Production and Speech Modelling*, edited by W. Hardcastle and A. Marchal (Kluwer Academic Publishers, Dordrecht), pp. 403–439.

Lindblom, B. (**1996**). "Role of articulation in speech perception: Clues from production," J. Acoust. Soc. Am. **99**, 1683–1692.

Luce, P., and Pisoni, D. B. (**1998**). "Recognizing spoken words: The neighborhood activation model," Ear Hear. **19**, 1–36.

MacDonald, E. N., Purcell, D. W., and Munhall, K. G. (**2011**). "Probing the independence of formant control using altered auditory feedback," J. Acoust. Soc. Am. **129**, 955.

MATLAB (**2011**). Version 7.10.0 (R2011b) (The MathWorks Inc., Natick, MA).

Mitsuya, T., MacDonald, E. N., Purcell, D. W., and Munhall, K. G. (**2011**). "A cross-language study of compensation in response to real-time formant perturbation," J. Acoust. Soc. Am. **130**, 2978–2986.

Moon, S. J., and Lindblom, B. (**1994**). "Interaction between duration, context, and speaking style in English stressed vowels," J. Acoust. Soc. Am. **96**, 40–55.

Munson, B., and Solomon, N. P. (**2004**). "The effect of phonological neighborhood density on vowel articulation," J. Speech Lang. Hear. Res. **47**, 1048–1058.

Natke, U., Donath, T. M., and Kalveram, K. T. (**2003**). "Control of voice fundamental frequency in speaking versus singing," J. Acoust. Soc. Am. **113**, 1587–1593.

Perkell, J. S., Denny, M., Lane, H., Guenther, F., Matthies, M. L., Tiede, M., Vick, J., Zandipour, M., and Burton, E. (**2007**). "Effects of masking noise on vowel and sibilant contrasts in normal-hearing speakers and postlingually deafened cochlear implant users," J. Acoust. Soc. Am. **121**, 505–518.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of vowels," J. Acoust. Soc. Am. **24**, 174–184.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1986**). "Speaking clearly for the hard of hearing. II: Acoustic characteristics of clear and conversational speech," J. Speech Hear. Res. **29**, 434–446.

Prochazka, A., Clarac, F., Loeb, G. E., Rothwell, J. C., and Wolpaw, J. R. (**2000**). "What do reflex and voluntary mean? Modern views on an ancient debate," Exp. Brain Res. **130**, 417–432.

Purcell, D. W., and Munhall, K. G. (**2006**). "Compensation following real-time manipulation of formants in isolated vowels," J. Acoust. Soc. Am. **119**, 2288–2297.

Schwartz, J. L., Boe, L. J., Vallee, N., and Abry, C. (**1997**). "The dispersion-focalization theory of vowel systems," J. Phonetics **25**, 255–286.

Siegel, G. M., and Pick, H. L. (**1974**). "Auditory feedback in the regulation of voice," J. Acoust. Soc. Am. **56**, 1618–1624.

Singh, S., and Woods, D. R. (**1971**). "Perceptual structure of 12 American English vowels," J. Acoust. Soc. Am. **49**, 1861–1866.

Stevens, K. N. (**1989**). "On the quantal nature of speech," J. Phonetics **17**, 3–45.

Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT Press, Cambridge, MA).

Stevens, S. S., and Volkmann, J. (**1940**). "The quantum of sensory discrimination," Science **92**, 583–585.

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

Ventura, M. I., Nagarajan, S. S., and Houde, J. F. (**2009**). "Speech target modulates speaking induced suppression in auditory cortex," BMC Neurosci. **10**, 58.

Villacorta, V. M., Perkell, J. S., and Guenther, F. H. (**2007**). "Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception," J. Acoust. Soc. Am. **122**, 2306–2319.

Wilson, H. F., and Bond, Z. S. (**1977**). "An INDSCAL analysis of vowels excerpted from four phonetic contexts," J. Phonetics **5**, 361–367.

Wright, R. (**2004**). "Factors of lexical competition in vowel articulation," in *Papers in Laboratory Phonology VI*, edited by J. Local, R. Ogden, and R. Temple (Cambridge University Press, Cambridge, UK), pp. 26–50.

Xia, K., and Espy-Wilson, C. (**2000**). "A new strategy of formant tracking based on dynamic programming," in *Proceedings of the Sixth International Conference on Spoken Language Processing*, Beijing, China, pp. 55–58.

Xu, Y., Larson, C. R., Bauer, J. J., and Hain, T. C. (**2004**). "Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences," J. Acoust. Soc. Am. **116**, 1168–1178.

Zhang, Z. Y., Espy-Wilson, C., Boyce, S., and Tiede, M. (**2005**). "Modeling of the front cavity and sublingual space in American English rhotic sounds," in *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, Speech Processing*, Vols. 1–5, pp. 893–896.

Zhou, X. H., Espy-Wilson, C. Y., Boyce, S., Tiede, M., Holland, C., and Choe, A. (**2008**). "A magnetic resonance imaging-based articulatory and acoustic study of 'retroflex' and 'bunched' American English /r/," J. Acoust. Soc. Am. **123**, 4466–4481.