Modulation of auditory-motor learning in response to formant perturbation as a function of delayed auditory feedback

Takashi Mitsuya[a)]
Department of Speech and Hearing Sciences, University of Washington, 1417 N.E. 42nd Street, Seattle, Washington 98105-6246, USA

Kevin G. Munhall
Department of Psychology, 62 Arch Street, Humphrey Hall, Room 232, Queen's University, Kingston, Ontario K7L 3N6, Canada

David W. Purcell
National Centre for Audiology, School of Communication Sciences and Disorders, Elborn College, Western University, London Ontario N6G 1H1, Canada

(Received 23 September 2016; revised 3 April 2017; accepted 5 April 2017; published online 19 April 2017)

The interaction of language production and perception has been substantiated by empirical studies where speakers compensate their speech articulation in response to the manipulated sound of their voice heard in real-time as auditory feedback. A recent study by Max and Maffett [(2015). Neurosci. Lett. **591**, 25–29] reported an absence of compensation (i.e., auditory-motor learning) for frequency-shifted formants when auditory feedback was delayed by 100 ms. In the present study, the effect of auditory feedback delay was studied when only the first formant was manipulated while delaying auditory feedback systematically. In experiment 1, a small yet significant compensation was observed even with 100 ms of auditory delay unlike the past report. This result suggests that the tolerance of feedback delay depends on different types of auditory errors being processed. In experiment 2, it was revealed that the amount of formant compensation had an inverse linear relationship with the amount of auditory delay. One of the speculated mechanisms to account for these results is that as auditory delay increases, undelayed (and unperturbed) somatosensory feedback is given more preference for accuracy control of vowel formants.
 [http://dx.doi.org/10.1121/1.4981139]

[ZZ] Pages: 2758–2767

I. INTRODUCTION

Language perception and language production have long been thought to be intricately related (e.g., Wernicke, 1874/1969), but the exact mechanism of the connection between input and output of speech and language is still a matter of research in many areas of psycholinguistics (Meyer *et al.*, 2016). One form of perception has a special place in this field: perception of one's own speech and language. Auditory feedback or monitoring mechanisms are a key feature of a number of models of speaking and listening (e.g., Wernicke, 1874/1969; Levelt, 1983), and a predictive mechanism that compares expected versus actual production is a part of other models (e.g., Garrod and Pickering, 2004; Pickering and Garrod, 2013; Dell and Chang, 2014). Studies of the control of speech articulation have long supported the importance of feedback and predictive or feedforward models. When specific characteristics of the auditory feedback of a speakers' speech are modified in real time, speakers change specific aspects of their speech production to compensate for the perceived error (i.e., both online correction and auditory-motor learning). This has been demonstrated for speaking amplitude (e.g., Bauer *et al.*, 2006), vocal frequency (e.g., Burnett *et al.*, 1998; Jones and Munhall, 2000; Larson *et al.*, 2007), fricative spectrum (Shiller *et al.*, 2009; Casserly, 2011), and vowel resonances/formants (e.g., Houde and Jordan, 1998; Purcell and Munhall, 2006; Villacorta *et al.*, 2007, Mitsuya *et al.*, 2015).

In the case of real time formant perturbation experiments, the first and/or the second formants ($F1$ and $F2$ hereafter, respectively) of a vowel are modified while speakers are producing a simple monosyllabic world. Speakers receive the modified signal in real time through the headphones they wear, and consequently, they hear a vowel slightly different from the one they intended to produce. In response, speakers spontaneously change their articulation to reduce the difference between the heard and intended formant values (i.e., error). Moreover, articulatory posture of the intended production is updated for the following productions.

Time sensitivity is an important constraint on real-time language processing (Christiansen and Chater, 2015) and the processing of feedback in speech has been shown to be sensitive to the timing of information. While some demonstrations of auditory feedback producing rapid immediate changes in speech characteristics have been reported (Sapir *et al.*, 1983), the timing of articulation and feedback processing is not conducive to a servomechanism. This does not imply that the timing of feedback is not crucial. Since the 1950s and early work on electronic communication systems it has been known that delays in auditory feedback can

[a)]Electronic mail: tmitsuya@uw.edu

 0001-4966/2017/141(4)/2758/10/$30.00

disrupt articulation (Lee, 1950). Later, however, it was demonstrated that short auditory feedback delays can induce fluency in persons who stutter (e.g., Kalinowski *et al.*, 1993).

These variations in feedback processing as a function of time might inform us about the units of planning and how real-time processes interact with auditory-motor learning. Recently, Max and Maffett (2015) reported that the temporal congruency of sensory feedback plays an important role in corrective vowel production. They manipulated the vocal tract resonance structure (i.e., formants) of the auditory feedback their speakers received with different levels of auditory delay while they were producing words, and examined how speakers changed their formant production in response to the manipulations of formant and feedback delay. Their data showed that with a delay as short as 100 ms, speakers' compensation for the formant manipulation, as an index for auditory-motor learning, was eliminated. Max and Maffett concluded that the speech motor control system gives no weight to auditory feedback delayed by 100 ms. This implies that the temporal window for incorporating auditory feedback is quite narrow (cf. Kalinowski *et al.*, 1993) or that the feedback control system might have different thresholds for some kinds of perturbations.

Here, we carried out similar experiments to Max and Maffett's (2015) study to examine the relationship between auditory feedback delay and auditory learning more closely using a different formant perturbation technique. The motivation of our study was to examine whether a difference in formant perturbation technique would elicit a different auditory motor learning in response to auditory delay. Max and Maffett (2015) increased all formants by 2.5 semitones, which introduces a larger perturbation for the higher formants in Hertz while keeping (1) the ratio of the formants (or formant dispersion) and (2) fundamental frequency ($F0$, hereafter) constant. Although the relationship between oral cavity configuration and its formant structure is complicated, $F1$ is strongly correlated with the phonetic value of vowel height, which is associated with openness of the jaw if $F0$ is kept constant (Traunmüller, 1981; Fahey and Diehl, 1996). However, changing all formants simultaneously while keeping formant dispersion and $F0$ constant may induce a perceptual manipulation other than a change in vowel quality, for example, the perception of the size of the vocal tract decreasing (see Fant, 1966, for a review). Therefore, Max and Maffett's (2015) results might have been influenced by a unique combination of change in the perception of the vowel and the size of the vocal tract.

In the current studies, we increased only $F1$ while speakers were producing the word "head." Increasing $F1$ while keeping all other acoustic cues constant (including $F0$ and higher formants), elicits perception of a more open vowel, which effectively sounds more like "had/hæd/" without inducing any other perceptual manipulations. In response, speakers typically lower their $F1$ to produce a vowel more like that of "hid/hɪd/." Auditory-motor learning (indexed by adaptive change in production) in response to $F1$ perturbation with the vowel /ɛ/ has been well studied and the data with an auditory delay would be easily comparable with the normative data that have been reported in the literature (e.g., Purcell and Munhall, 2006; MacDonald *et al.*, 2010, 2011; Mitsuya *et al.*, 2015). We would be able to better understand (1) the function relating auditory delay to auditory-motor learning for vowel formants, and (2) the nature of the control system's assessment of self-generated sensory consequence due to temporal (in)congruence of auditory feedback.

II. EXPERIMENT 1

The aim of this part of the study was to examine whether speakers changed their first formant production when they received perturbed feedback that only manipulated $F1$ in the presence and absence of 100 ms delay

A. Methods

1. Participants

Twenty female native Canadian English speakers with no hearing or speech impairments participated in this study with ages ranging from 18 to 31 yrs ($\bar{X} = 24.1$ yrs, standard deviation = 2.7 yrs). Although the sample size in the current study is much larger than that of Max and Maffett (2015), the number of participants per condition in the current experiment was comparable to that of previous studies that used the same formant perturbation technique (MacDonald *et al.*, 2010, 2011; Munhall *et al.*, 2009; Mitsuya *et al.*, 2011, 2013, 2015; Mitsuya and Purcell, 2016). Because there is a large difference in formant values across sexes, only female speakers were included in order (1) to keep the variability of formants small, and (2) to have the perceptual consequence of formant perturbation similar across participants. All had normal audiometric hearing thresholds within the range of 500–4000 Hz ($\leq$20 dB hearing level) tested using a Madsen Itera audiometer and a Telephonics Audiometry Transducers TDH-39 P headset (Otometrics/Audiology Systems, 296D000-9). Informed consent was obtained from the participants.

2. Equipment

The experiment took place in a sound attenuated room (Eckel Industries of Canada, model C26). Participants sat in front of a computer monitor with a portable microphone (Shure WH20), and headphones (Sennheiser HD 265). They were instructed to say the word "head" when the word was presented on the screen. As in Mitsuya *et al.* (2015), their microphone signal was amplified (Tucker-Davis Technologies MA3 microphone amplifier), low-pass filtered with a cutoff frequency of 4500 Hz (Frequency Devices type 901), digitized at 10 kHz, and filtered in real time to produce formant feedback manipulation (National Instruments PXI-8106 embedded controller). The participants heard this processed signal at approximately 80 dBA sound pressure level (SPL) with speech shaped noise (Madsen Itera) of 50 dBA SPL.

3. Acoustic processing

Voicing was detected using a statistical amplitude threshold and real-time formant manipulation was performed with an infinite impulse response filter (Purcell and Munhall, 2006).

Formants were estimated every 900 μs, using an iterative Burg algorithm (Orfanidis, 1988). Based on these estimates, filter coefficients were calculated such that a pair of spectral zeros was placed at the existing formant frequency and a pair of spectral poles was placed for the new formant to de-emphasize and emphasize existing voice harmonics, respectively. Our signal processing introduces up to approximately 6 ms delay from microphone to headphone.

A parameter that determines the number of coefficients used in the autoregressive analysis was estimated by collecting six tokens of each English vowel /i, ɪ, e, ɛ, æ, ɔ, o, u, ʊ, ɚ/ in the /hVd/ context ("heed," "hid," "hayed," "head," "had," "hawed," "hoed," "who'd," "hood," and "heard," respectively). A visual prompt of these words was presented on a computer screen for 2.5 s with an inter-trial interval of approximately 1.5 s. Speakers were instructed to say the prompted word without gliding their pitch. The best model order for the target vowel was chosen, based on minimum variance in formant frequencies $F1$ and $F2$ over the middle portion of the vowel. Model order estimation was done for each of the headphone conditions.

Offline formant analysis was done with the same method reported in Munhall *et al.* (2009). Each utterance's vowel boundaries were estimated based on harmonicity of the power spectrum. These bounds were then inspected and corrected if necessary. The middle 40%–80% of a vowel's duration was used to estimate the first three formants, with a 25 ms window that was shifted in 1 ms increments until the end of the middle portion of the vowel segment. From these sliding window estimates an average value was calculated. Formant estimates were inspected and were relabeled if mislabeled (e.g., $F1$ being labeled as $F2$) or removed if in error.

4. Design

Participants produced 120 productions of the word "head" with a visual prompt for 2.5 s with an inter-trial interval of approximately 1.5 s. They performed this task twice, once with no delay (0DL, hereafter) and once with 100 ms delay (100DL, hereafter). The order of the delay conditions was counterbalanced across subjects. The 120 trials were broken into four experimental phases (see Fig. 1). The first 20 trials (trial 1–20: Baseline) had no formant perturbation applied. In the second phase (trial 21–70: Ramp), an incremental increase of 4 Hz was applied to participants' $F1$ of auditory feedback for each of 50 successive trials. At the end of this phase, the maximum perturbation of 200 Hz was applied. In the third phase (trial 71–90: Hold), the 200 Hz maximum perturbation was held constant for 20 trials. In the final phase (trial 91–120: Return), the perturbation was removed at trial 91 and normal feedback was provided for the final 30 trials.

B. Results

The average of the last 15 trials of Baseline was calculated for each participant. This baseline average was subtracted from the $F1$ value of each trial to indicate changes in $F1$ production (i.e., normalized $F1$). The group averages of $F1$ change can be seen in Fig. 2 where both 0DL and 100DL

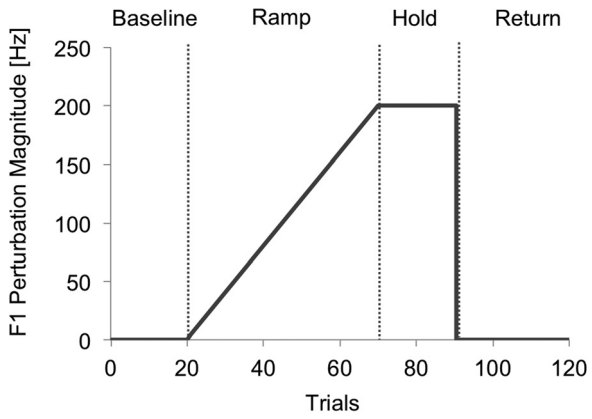

FIG. 1. Feedback perturbation applied to the first formant. The vertical dotted lines denote the boundaries of the four experimental phases: Baseline, Ramp, Hold, and Return (from left to right).

groups changed their production of $F1$ in response to the $F1$ perturbation they were receiving, but with different magnitudes. The last 15 of the normalized $F1$ values in the Hold phase were averaged for each participant then multiplied by –1 to estimate a magnitude of compensation. As can be seen in Fig. 3 and 0DL [$\bar{X} = 57.8$ Hz, standard error (s.e.) = 7.0 Hz] induced a significantly larger formant compensation than 100DL [$\bar{X} = 11.0$ Hz, s.e. = 4.4 Hz; $t(38) = 5.66$, $p < 0.001$]. But importantly, in both conditions, the change was significantly different from zero [0DL: $t(19) = 8.25$, $p < 0.001$; 100DL: $t(19) = 2.51$, $p = 0.02$].

We examined when participants began changing their $F1$ production during the Ramp phase. As in Mitsuya *et al.* (2013, 2015), threshold was defined as the first instance of the first three consecutive formant productions that were lower than 3 s.e. from the baseline average. All speakers in 0DL yielded a threshold point ($\bar{X} = 36.9$th trial, s.e. = 1.9 trial, or 67.6 Hz perturbation) while 17 speakers in 100DL yielded such a point ($\bar{X} = 37.9$th trial, s.e. = 3.5 trial, or 71.6 Hz perturbation). To conduct a paired sample t-test, thresholds for the three speakers who did not yield a threshold point in 100DL were omitted from the 0DL set.

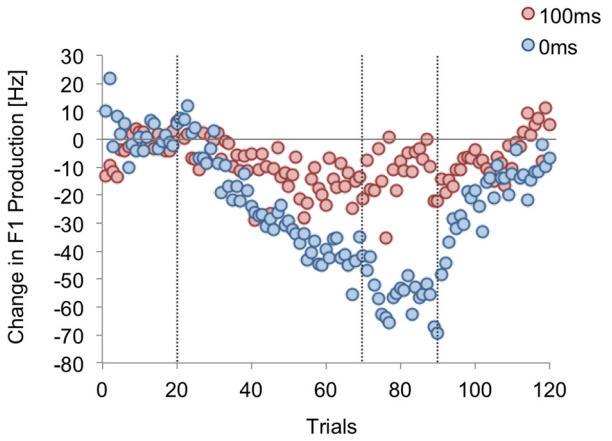

FIG. 2. (Color online) Group average trials of change in first formant production of the vowel /ɛ/ in experiment 1. The blue circles are the formant values in the 0 ms (0DL), whereas the red circles are the 100DL condition. The vertical dotted lines denote the boundaries of the experimental phases.

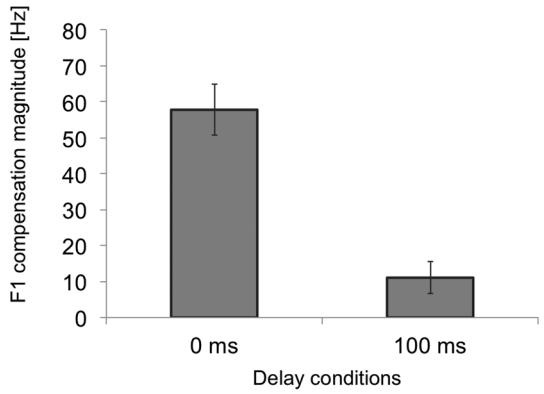

FIG. 3. Average compensation magnitude of the first formant of the vowel /ɛ/ in experiment 1. The 0 ms denotes the 0DL condition, whereas the 100 ms denotes the delayed condition (100DL).

Thresholds were not significantly different between 0DL and 100DL [$t(16) = 0.32$, $p > 0.05$].

As can be seen in Fig. 2, the conditions appeared to have different rates of compensation. The group average of $F1$ change was fitted with a linear model for each condition. For 0DL, the model yielded an unstandardized slope coefficient of –1.06 ($R^2 = 0.83$). This slope indicates a compensation of –1.06 Hz for every 4 Hz perturbation, or 26.5% of perturbation applied being compensation. On the other hand, for 100DL, the model yielded an unstandardized slope coefficient of –0.38 ($R^2 = 0.43$) which is –0.38 Hz change per every 4 Hz perturbation (or 9.5% of perturbation being compensated).

We also examined whether speakers changed their $F2$ in response to $F1$ perturbation, to rule out the possibility that speakers, especially in 100DL, tried to compensate for the delay in more than one way. Compensation magnitude was calculated in the same way as $F1$ described above. A one-sample t-test was performed to test whether the change in $F2$ production in response to the perturbation was significantly different from zero in both 0DL and 100DL conditions. The results revealed that neither conditions yielded a significant change of $F2$ [0DL: $\bar{X} = 10.13$ Hz, s.e. = 9.85 Hz, $t(19) = 1.03$, $p = 0.32$; 100DL: $\bar{X} = 9.98$ Hz, s.e. = 5.94 Hz, $t(19) = 1.68$, $p = 0.11$].

Because many of our speakers reported that they thought the feedback signal was louder in 100DL than in 0DL, we examined speakers' voice amplitude during the experiments by averaging root-mean-square microphone signal voltage for the vocalic portion of each utterance. As can be seen in Fig. 4(a), our speakers produced a higher voice amplitude in 100DL throughout the experiment. The amplitude measure of each speaker was averaged for the last 15 trials of Baseline, and compared across delay conditions. In 0DL, average voice amplitude was 63.2 dBA SPL (s.e. = 0.9 dBA SPL), whereas in 100DL, the average was 65.8 dBA SPL (s.e. = 1.0 dBA SPL) and a paired sample t-test revealed that the difference was significant [$t(19) = 7.45$, $p < 0.001$].

As can be seen in Fig. 4(a), the change in voice amplitude of the first several trials may suggest that the condition difference in voice amplitude is due to a third possibility.

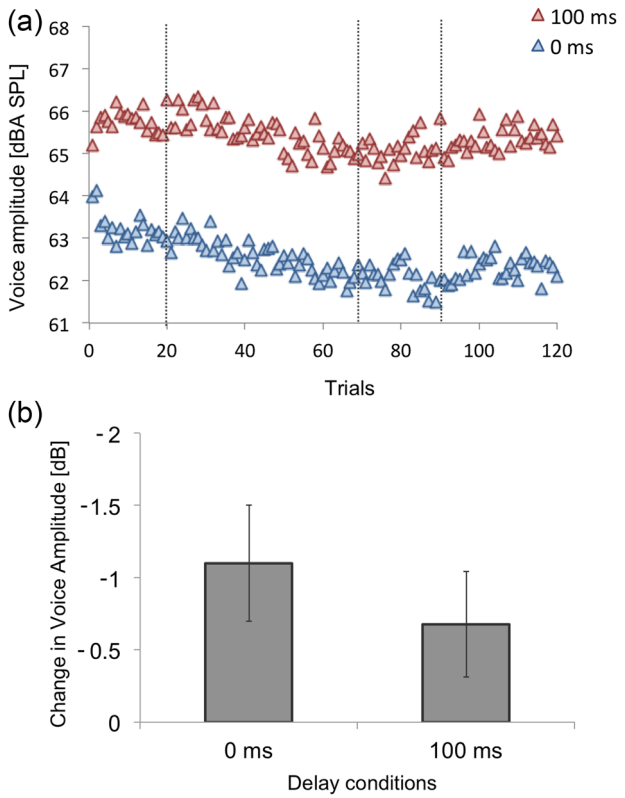

FIG. 4. (Color online) (a) Group average of voice amplitude (dBA SPL) of the vowel /ɛ/ in experiment 1. The blue triangles are the voice amplitude in the 0DL condition, whereas the red triangles are the 100DL condition. The vertical dotted lines denote the boundaries of the experimental phases. (b) Group average of change in voice amplitude from baseline to hold phases. The error bar represents 1 s.e.

Speakers might have decreased their voice amplitude slightly due to their auditory feedback being presented at a relatively high level (0DL: trials 1–2 compared to subsequent trials), and that they increased their voice amplitude due to the delay (100DL: trial 1 compared to later trials).

Interestingly, speakers' voice amplitude appeared to decrease in both conditions as more perturbation was applied [see Fig. 4(a)]. To test this, we normalized their voice amplitude by subtracting the baseline average from each individual voice-amplitude token. The normalized voice amplitude of the last 15 trials of the Hold phase was averaged for each speaker [see Fig. 4(b)] and submitted to a one sample t-test to examine whether each condition's voice amplitude changed significantly from zero. Only 0DL yielded a significant change [0DL: $\bar{X} = -1.10$ dB, s.e. = 0.40 dB, $t(19) = -2.74$, $p = 0.013$; 100DL: $\bar{X} = -0.68$ dB, s.e. = 0.37 dB, $t(19) = 1.85$, $p = 0.080$, see Fig. 4(b)]. However when the two conditions were compared against each other with a paired sample t-test, the amplitude changes were not significantly different [$t(19) = 0.76$, $p = 0.46$]. Decreasing voice amplitude would inevitably decrease the level of auditory feedback, in which case, the robustness of the perturbation we were delivering might have also lessened slightly. To examine this relationship, we calculated correlation between voice amplitude decrease during the Hold phase and magnitude of compensation. Neither condition yielded a significant correlation [0DL: $r(20) = -0.13$, $p = 0.58$; 100DL: $r(20) = 0.22$, $p = 0.35$]. This

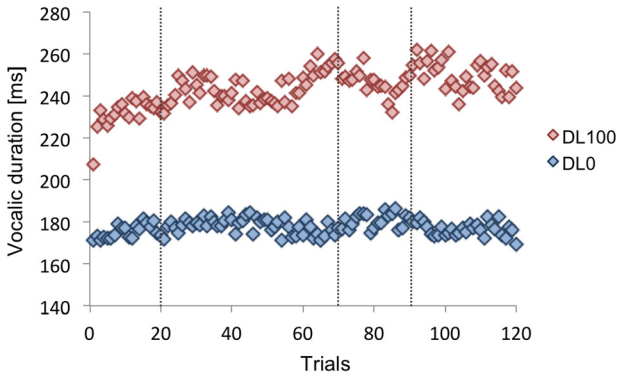

FIG. 5. (Color online) The group average vocalic duration across trials. Red diamonds indicate 100DL condition and blue diamonds indicate the 0DL condition. The vertical dotted lines denote boundaries of the experimental phases.

observation was consistent with a recent report by Mitsuya and Purcell (2016).

Finally, we examined vocalic duration (see Fig. 5). The average of the last 15 trials of the Baseline, Hold, and Return phases were calculated for each speaker and the average vocalic durations were submitted to a 2 (delay condition) × 3 (experimental phase) repeated measures analysis of variance. It was revealed that the effect of delay condition was significant ($F[1,119] = 66.14$, $p < 0.001$), that 100DL produced much longer productions ($\bar{X} = 242.17$ ms, s.e. = 14.48 ms), than 0DL ($\bar{X} = 178.14$ ms, s.e. = 8.48 ms). In addition, the effect of the experimental phase also yielded significance [$F(2,38) = 3.36$, $p = 0.045$]. However, *post hoc* analysis with a Bonferroni correction (alpha set at 0.0067 for three contrasts) did not yield significance. Possibly this procedure was too conservative to detect a significant effect; however, the p-values of the three contrasts indicated that if there was a significant difference, it would have been the Baseline and the Hold phases ($p = 0.025$), while the other two contrasts, p-values were larger than 0.05. An interaction between the main effects was not significant ($F[2,38] = 1.55$, $p = 0.23$).

If speakers produced longer productions, then they would have received equally longer perturbed portions of the feedback, compared to those who produced shorter productions. This implies that longer production would have given the speech control system more or stronger evidence of incongruency between planned versus heard production within each delay condition. Consequently, vocalic duration might have been correlated with $F1$ compensation magnitude. To examine this relationship, we calculated correlation of each speaker's average vocalic duration in the Hold phase and her compensation magnitude. Neither condition yielded a significant result (0DL: $r[20] = 0.37$, $p = 0.11$; 100DL: $r[20] = 0.26$, $p = 0.27$).

C. Discussion

Max and Maffett (2015) reported that speakers' compensation behavior in response to formant perturbation was eliminated with 100 ms auditory feedback delay. However, the current study found a small yet significant compensation in response to $F1$ perturbation with the same amount of delay. In addition, the delay conditions did not differ in compensation threshold, that is, the speakers in both conditions started changing $F1$ production approximately at the same trial. This result indicates that error detection was unaffected by auditory delay, and the reduction of compensation magnitude was mainly due to a smaller adaptation factor.

We suspect that the difference in the compensation results between the current study and that of Max and Maffett (2015) is largely due to the nature of the perturbation applied to the auditory feedback. However, there are a few other considerations. First, the current study tested 20 speakers whereas Max and Maffett (2015) examined eight. Vowel production is inherently variable (see Peterson and Barney, 1952, for a review), and there is quite a variable response to formant perturbation using our technique from almost complete compensation to a response in the same frequency direction as the perturbation (following, instead of compensating; see Fig. 3 in MacDonald *et al.*, 2011, for a distribution of compensation behavior). If only a small group average compensation magnitude was present with a delay, testing a larger sample would make detection of the effect more feasible. Second, the signal presentation level was slightly different from Max and Maffett's (2015) studies. They presented their auditory feedback at 75 dB SPL with 68 dB SPL pink noise, while we presented our processed signal at 80 dBA SPL with 50 dBA SPL speech shaped noise. The difference in the signal-to-noise ratio along with the signal level itself might have masked the unprocessed bone-conducted signals differently.

The change in voice amplitude during the current experiment needs further examination. First, there was a significant difference in voice amplitude between the delay conditions. This observation is consistent with what has been reported in the delayed auditory feedback literature (see Siegel *et al.*, 1980; Howell *et al.*, 1983; Howell and Archer, 1984; Howell and Powell, 1987). While the airborne feedback was delayed in the 100DL, the speakers' cochleae still received simultaneous bone-conducted sound of the utterances. The speech planning system's expectation of voice amplitude would be higher than the sensed amplitude with limited air-conducted feedback. Perhaps the rapid 100DL increase of voice amplitude seen in Fig. 4(a) was adaptive learning of voice amplitude in response to the absence of amplified airborne feedback to offset the difference between predicted versus actual sensory feedback. However, the current design does not allow us to conclude whether it was the case that (1) the 0DL condition produced lower voice amplitude due to the robust signal presentation at approximately 80 dBA SPL, (2) the 100DL condition produced higher voice amplitude due to the delay, or (3) a combination of both.

Second, the results of the current data, as well as Mitsuya and Purcell's (2016) data suggest that decreased voice amplitude during the perturbation phase was not likely due to change in articulatory posture because of no significant correlation between compensation magnitude and the change in voice amplitude for either condition (0DL: $r[20] = 0.07$, $p = 0.78$; 100DL: $r[20] = 0.12$, $p = 0.62$), but was likely due to a change in feedback. One of the ways in which feedback might have influenced voice amplitude was

the feedback amplitude. However, Mitsuya and Purcell (2016) determined that the formant perturbation processes did not influence the feedback amplitude in systematic or substantial ways. Another possibility is that differences between the intended versus the heard (perturbed) signals caused the speech motor control to reduce the input level of auditory feedback. As the discrepancy became larger, the reduction was increased. Because somatosensory feedback is congruent with the expected feedback, it is possible that the system may shift the relative contributions of auditory versus somatosensory feedback by reducing the auditory input level.

One might argue that with delays of 100 ms, the speakers have intentionally avoided changing their speech production when they noticed their auditory feedback was delayed; however, this speculation is unlikely for two reasons. First, awareness of delayed feedback has been reported to be variable (Natke and Kalveram, 2001). Moreover, spontaneity judgment of vocal production and auditory feedback was reported to be at chance with a delay of 100 ms (Yamamoto and Kawabata, 2011). Thus, some of our speakers were likely not aware of a delay of 100 ms, and hence conscious control should not have occurred. Second and more importantly, formant compensation has been reported to be highly automatic (Munhall *et al.*, 2009); thus, even if speakers were aware of the auditory delay, reduced compensation was not likely due to volitional control.

Overall, we have evidence that with our $F1$ perturbation technique, speakers still exhibited some compensatory formant production when auditory feedback was delayed by 100 ms. However, the significant reduction in compensation magnitude in the delay condition still raises questions about the modulation of formant compensation as a function of auditory delay. Experiment 2 was carried out to examine this by incrementally introducing smaller auditory delays.

III. EXPERIMENT 2

In this part of the study we examined how compensation behavior is modulated with different amounts of feedback delay, by incrementally decreasing delay from 100 ms while simultaneously applying a large formant perturbation. Specifically, we examined the modulation of compensation magnitude as a function of auditory delay. The reason that we decreased the delay, instead of increasing it, was due to the slow time course of de-adaptation. When a large perturbation is introduced, speakers generally start adapting their formants within several trials from the onset of the perturbation (MacDonald *et al.*, 2010), whereas when a large perturbation is removed after speakers have adapted to the perturbation, it often takes many more trials for them to de-adapt (Purcell and Munhall, 2006). The current design would allow us to capture a sudden change in compensation magnitude more easily, particularly if the delay were to affect the behavior in a non-linear fashion.

A. Methods

The same participants from experiment 1 took part in experiment 2. The time between experiments was approximately 1 week for most of the participants. One did not return to the study; her data were excluded when experiments 1 and 2 were compared. The equipment and signal processing in experiment 2 were identical to those of experiment 1.

Speakers produced the word "head" 130 times, and the experiment started with a 100 ms delay with no formant perturbation for the first 20 trials. At trial 21, participants' $F1$ were increased by 200 Hz all at once. This perturbation was held constant throughout the experiment. Every 10 trials beginning with trial 31, the auditory delay was reduced by 10 ms (i.e., trials 21–30 had a 100 ms delay, trials 31–40 had a 90 ms delay and so on). At trial 121, the delay was eliminated while the 200 Hz $F1$ perturbation was still being applied (see Fig. 6).

B. Results

Participants' formant values were normalized by subtracting the mean of the last 15 baseline trials as described in experiment 1. The group average of normalized $F1$ production can be seen in Fig. 7. Participants, on average, increased their compensation magnitude gradually as auditory delay decreased incrementally. From trial 21 through 130, the group average compensation magnitude yielded a linear model with an unstandardized slope coefficient of -0.47 ($R^2 = 0.85$), verifying an increase in compensation magnitude (more negative Hz values) as the experiment progressed. The average compensation magnitude with 100 ms delay (trials 26–30: the first 5 perturbed trials were excluded because speakers might still have been adjusting their articulation in response to a sudden introduction of a delay, hence, they might not have fully compensated during this period, as reported in MacDonald *et al.*, 2010) was 16.85 Hz (s.e. = 6.4 Hz), which was significantly different from zero [$t(18) = 2.63$, $p = 0.01$], replicating the effect that we found in experiment 1. Moreover, the compensation magnitude with 100 ms delay was not significantly different across the two experiments ($t[18] = 0.83$, $p = 0.42$). Average compensation magnitude with no delay trials (trials 121–130) was also estimated using the last 5 trials of the phase (i.e., trials 126–130) because it is possible that speakers might have been changing their articulation in response to no delay from 10 ms delay during this period. The average compensation magnitude with no delay was 55.64 Hz (s.e. = 7.3 Hz) and it was significantly different from zero

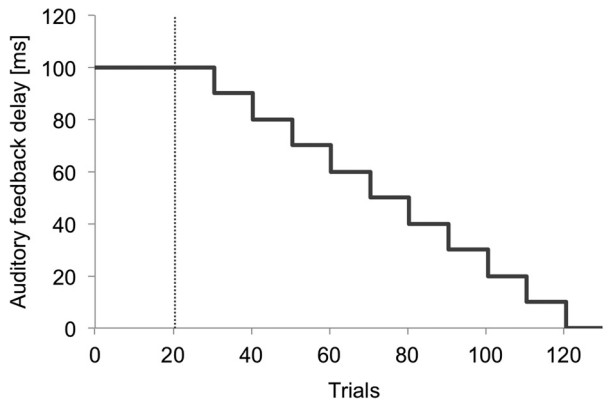

FIG. 6. Feedback delay applied to the auditory feedback (in ms). The vertical dotted line indicates the onset of +200 Hz $F1$ perturbation.

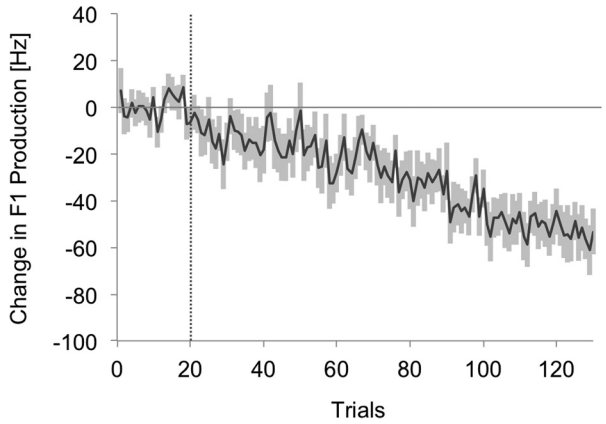

FIG. 7. Group average of change in first formant production of the vowel /ɛ/ in experiment 2. The line indicates the group average, whereas the shading indicates 1 s.e. The vertical dotted line denotes the onset of 200 Hz $F1$ perturbation.

[$t(18) = 7.59$, $p < 0.001$]. This magnitude was not different from the observed magnitude in experiment 1 during the hold phase of the 0DL condition [$t(18) = 0.61$, $p = 0.55$].

As in experiment 1, we also analyzed speakers' voice amplitude. The group average voice amplitude can be seen in Fig. 8. The baseline voice amplitude average of experiment 2 (trials 6–20) was 66.6 dBA SPL (s.e. = 1.1 dBA SPL) and was significantly higher than that of 0DL in experiment 1 [trials 6–20: $t(18) = -6.263$, $p < 0.001$], but it was not different from that of 100DL in experiment 1 [$t(18) = 1.68$, $p = 0.11$].

With the same definition used to estimate the threshold for formant compensation in experiment 1, we estimated the trial at which each speaker started decreasing voice amplitude (voice amplitude threshold). On average, our speakers started decreasing voice amplitude at trial 62.7 (s.e. = 9.2 trial), which corresponds to a 60 ms delay. Therefore, we estimated that voice amplitude threshold to be 60 ms.

C. Discussion

The compensation magnitude of $F1$ production in response to an increase $F1$ by 200 Hz was found to be linearly modulated by the amount of auditory delay such that compensation behavior increased as the delay amount decreased. One possible explanation for the results is that auditory delay

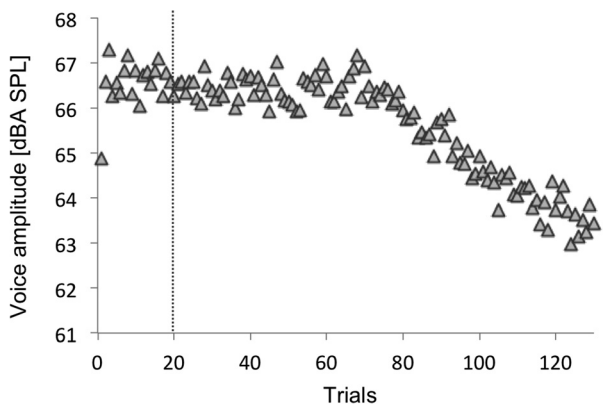

FIG. 8. Group average of voice amplitude (dBA SPL) of the vowel /ɛ/ in experiment 2. The vertical dotted line denotes the onset of 200 Hz $F1$ perturbation.

changes the contribution of somatosensory versus auditory feedback used to monitor the accuracy of speech production. With an auditory delay, an error assessment is initiated with somatosensory feedback before auditory feedback becomes available. Simultaneous sensory feedback may be preferred or weighted more than sensory feedback that is not simultaneous when multiple sensory feedback modalities are available for assessment of the accuracy of motoric behavior. Thus, as the temporal disparity between somatosensory feedback and auditory feedback becomes larger due to an auditory delay, the reliance on somatosensory feedback might become larger because it is deemed the more reliable sensory feedback due to temporal congruency.

Reduced compensation magnitude may also be modulated by how much unperturbed body/bone-conducted signal reached speakers' cochleae. The longer the delay is, the longer the duration of unperturbed speech signal reaching the cochleae before the onset of perturbed air-conducted feedback. How bone-conducted feedback influences formant production and compensation during the delay is not well understood. However, it is possible to examine it by presenting a higher level of noise while a delay is applied in order to mask bone-conducted sound reaching the cochleae, similar to the signal-to-noise ratio used in Max and Maffet's (2015) study. Setting the noise level so as not to interfere with air-conducted feedback requires future investigation.

Because we did not test delays longer than 100 ms, it is unclear whether a significant compensatory formant production would still be observed beyond this delay amount (as mentioned in Sec. II C). Given that the vocalic duration of the tested word was generally less than a few hundred milliseconds, and that at least 165 ms is needed for on-going feedback-based corrective responses (e.g., approximately 165 ms in their shift up condition by Tourville *et al.*, 2008), it is uncertain that delays much longer than 100 ms would still elicit compensation production via feedforward updating. Further experiments are needed to directly examine this.

Speakers' voice amplitude showed a slightly different pattern than that of formant compensation magnitude. Voice amplitude stayed constant until the delay was reduced to approximately 60 ms. A delay shorter than 60 ms showed a clear linear reduction in voice amplitude (in dB) such that as the delay became shorter, speakers' voice amplitude became lower. Without an obvious change in $F1$ compensation behavior around the delay of 60 ms, one might speculate that the amount of auditory delay affected the control of formant and voice amplitude differently. Although the observable behavior might be somewhat different between these two parameters in the current experiment, the underlying neural circuitry and mechanism of formant compensation and processing delayed auditory feedback have been reported to be similar [Formant perturbation, e.g., Niziolek and Guenther (2013), Zheng *et al.* (2013); Delayed auditory feedback, e.g., Hashimoto and Sakai (2003), Takaso *et al.* (2010)]. There does not appear to be a large difference in how quickly behavioral responses are observed for formant perturbations (e.g., approximately 165 ms by Tourville *et al.*, 2008) and intensity perturbations (e.g., approximately 130 ms by Bauer *et al.*, 2006). The shared neural circuitry along with similar latencies

does not necessarily mean that the two parameters are functionally coupled; further examinations are needed to help understand the interplay of formant and voice amplitude control.

IV. GENERAL DISCUSSION

Temporal congruency of sensory feedback and the action that generates it is one of the important cues for the control system to process feedback as generated by the self (Weiskrantz *et al.*, 1971). In the present paper, we examined how the speech motor control system processes auditory delay using magnitude of compensatory production for perturbed *F*1 feedback as a behavioral index. In experiment 1, a small yet significant compensatory formant production was observed with a 100DL, contrary to the previous study by Max and Maffett (2015). This difference is likely due to the use of different formant perturbation techniques between the studies.

When considering error monitoring and feedback processing, we have to keep in mind that language production involves information organized across many different temporal spans, and that timing of response and error tolerance may differ for different temporal intervals. Even within one level of language production such as speech motor control, feedback processing might vary with the type of parameter that is modified. The timing of auditory feedback processing for *F*0 or speech amplitude signals that are part of a prosodic contour may be longer than changes affecting a single phonemic status. In the former case, information must be integrated over time to determine if the prosodic pattern is incorrect. Even at a single time scale, compensatory behavior may depend on the type of manipulation(s) of the feedback or the types of errors. Changing vocal tract length (changing all formants such as in Max and Maffett's study) might have been more difficult for the control system to resolve articulatorily, compared to a single articulatory correlate of vowel openness (i.e., *F*1 in the present study). As a result, compensation for perturbations of single formants may be given preference. Additionally, the task participants are performing may change the priority or need for compensation. The *F*1 perturbations in the current studies change the lexical target while the shift in all formants produces a qualitative change in the utterance's sound with a subtle change to the phonemic or lexical target. The importance of each kind of perturbation could presumably be modified by task demands (such as vowel, voicing category and lexical contrast), which have been demonstrated by Mitsuya and his colleagues (2011, 2013, 2014).

The graded compensation magnitude as a function of delay reported in experiment 2 suggests that multimodal feedback information might be combined in a weighted fashion. For any single utterance, the speech motor system has access to kinesthetic information as well as bone-conducted and airborne sound. In normal conditions, feedback from these information sources is combined. With delay, the initial part of the vowel will yield no perturbed auditory feedback, but some veridical feedback through bone conduction and headphone "leakage." The duration of this decreased perturbed feedback will vary as a function of delay. Perhaps incongruency of sensory feedback may influence the assessment of self-generation of the motor behavior.

In the context of self-generation assessment of motor behavior, the sensitivity to temporal congruency of articulation and auditory feedback has been demonstrated neurally in a pitch-shifted auditory feedback study by Behroozmand *et al.* (2011). While their speakers were producing a sustained vowel, the pitch of their voice was perturbed along with various delay amounts. They found that the amplitude of the *P*2 component of auditory evoked potentials was significantly larger during production compared to passively listening to the same stream of sounds that they had previously produced, but only when auditory feedback was delayed (Behroozmand *et al.*, 2011). In addition, when the feedback was congruent with articulation (i.e., no auditory delay), the component was considerably reduced (i.e., suppression), regardless of whether the pitch was shifted or not. Because the suppression effect was not observed in their pitch-shifted no delay condition, contrary to the disappearance of a *N*1 suppression effect with pitch-shifted feedback regardless of feedback delay, Behroozmand *et al.* (2011) speculate that modulation of *P*2 is a motor-related effect, whereby temporal congruence of auditory feedback and the actual articulation (regardless of the auditory/phonetic errors) is detected, and may be critical for the assessment of "self-generation" of motor behavior.

We have years of experience hearing our voice when we speak (both through air- and bone-conducted sound). Recognition of voice as self has been reported to be very accurate (e.g., Maurer and Landis, 1990; Kaplan *et al.*, 2008), which indicates that we are sensitive to the acoustic characteristics of our voice. Many dynamic acoustic cues can be used for voice identity but some cues such as ones that reflect the speaker specific morphology of the vocal tract cannot be controllable. In this sense, those cues may function similarly to uncontrollable auditory delay. Hence, one can argue that self-recognition of voice may be important for assessing whether the feedback is a genuine consequence of self-generated articulation. When we hear someone else's voice as perturbed feedback, speakers may not compensate because the voice they hear is not clearly theirs, and the control system might "reject" it. However, for assessing whether the auditory feedback is a genuine sensory result of actual articulation, temporal congruency might be more important than phonetic details of auditory feedback. This is consistent with evidence that, at least at a conscious level, the ownership of voice is malleable. Zheng *et al.* (2011) showed that when their participants were given auditory feedback of a prerecorded voice of a stranger producing the same word, many of them reported that it was a modified version of their own voice. Zheng *et al.* (2011) also reported that speakers' *F*0 changed due to the introduction of the other voice as feedback. But the nature and mechanism of the change were not fully identified because the speakers changed their *F*0 toward the *F*0 of the feedback voice, as if they were trying to match their *F*0 to the feedback voice's *F*0 (following), instead of moving away from it (compensation). Their study did not manipulate other acoustic parameters; thus it is still

unknown whether speakers would compensate if other perturbation(s) such as vowel formants were perturbed along with voice identity. However, it may be the case that speakers might accept strangers' voices as their own and change their own production to compensate if other acoustic parameters (e.g., *F*1) of the others' voices are perturbed concurrently. If that is the case, phonetic details of self may not be so important but rather matching temporal congruency of the abstract representation of the articulatory trajectories might be what the control system uses for assessment of self-generation. Further examinations are needed to better understand the nature of processing of self-assessment.

ACKNOWLEDGMENTS

We would like to thank Mark Tiede at Haskins Laboratories for use of his offline *F*0 and formant estimation tools, and Emma Bridgwater for editing the manuscript. This research was supported by a Canada Foundation for Innovation the Natural Sciences and Engineering Research Council of Canada (NSERC), and was partially supported by NIH/NIDCD Grant No. R01 DC014510.

Bauer, J. J., Mittal, J., Larson, C. R., and Hain, T. C. (**2006**). "Vocal responses to unanticipated perturbations in voice loudness feedback," J. Acoust. Soc. Am. **119**, 2363–2371.

Behroozmand, R., Liu, H., and Larson, C. R. (**2011**). "Time-dependent neural processing of auditory feedback during voice pitch error detection," J. Cognit. Neurosci. **23**, 1205–1217.

Burnett, T. A., Freedland, M. B., Larson, C. R., and Hain, T. C. (**1998**). "Voice *F*0 responses to manipulations in pitch feedback," J. Acoust. Soc. Am. **103**, 3153–3161.

Casserly, E. D. (**2011**). "Speaker compensation for local perturbation of fricative acoustic feedback," J. Acoust. Soc. Am. **129**, 2181–2190.

Christiansen, M. H., and Chater, N. (**2015**). "The language faculty that wasn't: A usage-based account of natural language recursion," Front. Psychol. **6**, 1182.

Dell, G. S., and Chang, F. (**2014**). "The P-chain: Relating sentence production and its disorders to comprehension and acquisition," Philos. Trans. R. Soc. B **369**, 20120394.

Fahey, R. P., and Diehl, R. L. (**1996**). "The missing fundamental in vowel height perception," Percept. Psychophys. **58**, 725–733.

Fant, G. (**1966**). "A note on vocal tract size factors and non-uniform F-pattern scalings," Speech Trans. Lab. Quart. Prog. Stat. Rep. **4**, 22–30.

Garrod, S., and Pickering, M. J. (**2004**). "Why is conversation so easy?," Trends. Cogn. Sci. **8**, 8–11.

Hashimoto, Y., and Sakai, K. L. (**2003**). "Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: An fMRI study," Hum. Brain. Map. **20**, 22–28.

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

Howell, P., and Archer, A. (**1984**). "Susceptibility to the effects of delayed auditory feedback," Percept. Psychophys. **36**, 296–302.

Howell, P., and Powell, D. J. (**1987**). "Delayed auditory feedback with delayed sounds varying in duration," Percept. Psychophys. **42**, 166–172.

Howell, P., Powell, D. J., and Khan, I. (**1983**). "Amplitude contour of the delayed signal and interference in delayed auditory feedback task," J. Exp. Psychol. Human. **9**, 772–784.

Jones, J. A., and Munhall, K. G. (**2000**). "Perceptual calibration of *F*0 production: Evidence from feedback perturbation," J. Acoust. Soc. Am. **108**, 1246–1251.

Kalinowski, J., Armson, J., Stuart, A., and Gracco, V. L. (**1993**). "Effects of alterations in auditory feedback and speech rate on stuttering frequency," Lang. Speech. **36**, 1–16.

Kaplan, J. T., Aziz-Zadeh, L., Uddin, L. Q., and Iacoboni, M. (**2008**). "The self across the senses: An fMRI study of self-face and self-voice recognition," Soc. Cogn. Affec. Neur. **3**, 218–223.

Larson, C. R., Sun, J., and Hain, T. C. (**2007**). "Effects of simultaneous perturbations of voice pitch and loudness feedback on voice F_0 and amplitude control," J. Acoust. Soc. Am. **121**, 2862–2872.

Lee, B. S. (**1950**). "Effects of delayed speech feedback," J. Acoust. Soc. Am. **22**(6), 824–826.

Levelt, W. J. (**1983**). "Monitoring and self-repair in speech," Cognition **14**, 41–104.

MacDonald, E. N., Goldberg, R., and Munhall, K. G. (**2010**). "Compensation in response to real-time formant perturbations of different magnitude," J. Acoust. Soc. Am. **127**, 1059–1068.

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**, 955–965.

Maurer, D., and Landis, T. (**1990**). "Role of bone conduction in the self-perception of speech," Folia. Phoniatr. Logo. **42**, 226–229.

Max, L., and Maffett, D. G. (**2015**). "Feedback delays eliminate auditory-motor learning in speech production," Neurosci. Lett. **591**, 25–29.

Meyer, A. S., Huettig, F., and Levelt, W. J. (**2016**). "Same, different, or closely related: What is the relationship between language production and comprehension," J. Mem. Lang. **89**, 1–7.

Mitsuya, T., MacDonald, E. N., and Munhall, K. G. (**2014**). "Temporal control and compensation for perturbed voicing feedback," J. Acoust. Soc. Am. **135**, 2986–2994.

Mitsuya, T., MacDonald, E. N., Munhall, K. G., and Purcell, D. W. (**2015**). "Formant compensation for auditory feedback with English vowels," J. Acoust. Soc. Am. **138**, 413–424.

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.

Mitsuya, T., and Purcell, D. W. (**2016**). "Occlusion effect on compensatory formant production and voice amplitude in response to real-time perturbation," J. Acoust. Soc. Am. **140**, 4017–4026.

Mitsuya, T., Samson, F., Ménard, L., and Munhall, K. G. (**2013**). "Language dependent vowel representation in speech production," J. Acoust. Soc. Am. **133**, 2993–3003.

Munhall, K. G., MacDonald, E. N., Byrne, S. K., and Johnsrude, I. (**2009**). "Speakers alter vowel production in response to real-time formant perturbation even when instructed to resist compensation," J. Acoust. Soc. Am. **125**, 384–390.

Natke, U., and Kalveram, K. T. (**2001**). "Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables," J. Speech, Lang. Hear. Res. **44**, 577–584.

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

Orfandidis, S. J. (**1988**). *Optimum Signal Processing, An Introduction* (MacMillan, New York).

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Pickering, M. J., and Garrod, S. (**2013**). "An integrated theory of language production and comprehension," Behav. Brain Sci. **36**, 329–347.

Purcell, D. W., and Munhall, K. G. (**2006**). "Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation," J. Acoust. Soc. Am. **120**, 966–977.

Sapir, S., McClean, M. D., and Larson, C. R. (**1983**). "Human laryngeal responses to auditory stimulation," J. Acoust. Soc. Am. **73**, 315–321.

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

Siegel, G. M., Fehst, C. A., Garber, S. R., and Pick, H. L. (**1980**). "Delayed auditory feedback with children," J. Speech, Lang. Hear. Res. **23**, 802–813.

Takaso, H., Eisner, F., Wise, R. J., and Scott, S. K. (**2010**). "The effect of delayed auditory feedback on activity in the temporal lobe while speaking: A positron emission tomography study," J. Speech, Lang. Hear. Res. **53**, 226–236.

Tourville, J. A., Reilly, K. J., and Guenther, F. H. (**2008**). "Neural mechanisms underlying auditory feedback control of speech," Neuroimage **39**(3), 1429–1443.

Traunmüller, H. (**1981**). "Perceptual dimension of openness in vowels," J. Acoust. Soc. Am. **69**, 1465–1475.

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.

Weiskrantz, L., Elliot, J., and Darlington, C. (**1971**). "Preliminary observations of tickling oneself," Nature **230**, 598–599.

Wernicke, C. (**1874/1969**). "The symptom complex of aphasia: a psychological study on an anatomical basis," in *Boston Studies in the Philosophy of Science*, edited by R. S. Cohen and M. W. Wartofsky (D. Reidel, Dordrecht), pp. 34–97.

Yamamoto, K., and Kawabata, H. (**2011**). "Temporal recalibration in vocalization induced by adaptation of delayed auditory feedback," PloS One **6**, e29414.

Zheng, Z. Z., MacDonald, E. N., Munhall, K. G., and Johnsrude, I. S. (**2011**). "Perceiving a stranger's voice as being one's own: A 'rubber voice' illusion?," PloS One **6**, e18655.

Zheng, Z. Z., Vicente-Grabovetsky, A., MacDonald, E. N., Munhall, K. G., Cusack, R., and Johnsrude, I. S. (**2013**). "Multivoxel patterns reveal functionally differentiated networks underlying auditory feedback processing of speech," J. Neurosci. **33**, 4339–4348.