Experimental Brain Research (2019) 237:3155–3163
https://doi.org/10.1007/s00221-019-05657-1

RESEARCH ARTICLE

The influence of bistable auditory feedback on speech motor control

Takashi Mitsuya[1] · K. G. Munhall[2]

Received: 10 September 2018 / Accepted: 18 September 2019 / Published online: 3 October 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

The precision of speech production is strongly influenced by the auditory feedback of our voice. Studies have demonstrated that when speakers receive perturbed auditory feedback, they spontaneously change their articulation to reduce the difference between the intended sound and what was heard. For controlling the accuracy of vowel and consonant production, this corrective behavior reflects the intended sound's category represented in the mind. This in turn suggests that sounds that are not represented as a category would be controlled differently. The current study investigated the effect of linguistic status on controlling the production of a sound. Participants used an apparatus to produce a bistable sound, which could be heard either as a vowel or an acoustic buzz depending on the instructions. During the production of the target sound, their auditory feedback was perturbed and the change in acoustics in response to the perturbation was measured. The results showed that the group producing a linguistic target exhibited an error-reducing behavior similar to what was reported in natural vowel production, whereas the group producing a non-linguistic acoustic target exhibited an error-amplifying behavior, even when the majority in this group had noticed the perturbation and consciously tried to counteract it. Our findings indicate that access to a stable representation differentiates the perceptual organization of auditory feedback, and consequently changes how the production of the sound is controlled.

Keywords Speech production · Speech motor control · Sensorimotor compensation · Auditory motor compensation

Introduction

Across many species, vocal production involves generating a sound and filtering that sound by its passage through the vocal tract (Taylor and Rebby 2010). This source-filter model has been the working model for speech acoustics for decades (Chiba and Kajiyama 1941; Fant 1960), but remarkably, there are still many unknowns. Within this theoretical framework, recent efforts have focused on the control system that can account for consistent articulation and the categorical production of speech sounds (Hickok 2014; Guenther 2016).

For motor control systems, the use of sensory feedback is important to correct ongoing and future actions (e.g., Todorov 2004), and for speech production, the sounds that a speaker produces are thought to be the primary feedback signal. For example, adult hearing loss changes many aspects of vocal control over time (e.g., fundamental frequency, intensity, timing, and spectral precision), indicating the importance of auditory feedback for maintaining articulatory precision (Waldstein 1990; Cowie and Douglas-Cowie 1992; Schenk et al. 2003). Laboratory studies have also demonstrated that normal-hearing speakers change their articulation when real-time perturbations are applied to their vocal pitch, vowel formant frequencies as well as fricative frequency distribution to counteract such perturbations (Burnett et al. 1998; Jones and Munhall 2000; Houde and Jordan 1998; Purcell and Munhall 2006; Shiller et al. 2009; Casserly 2011).

However, unlike the frequent observation of near-complete compensation for visual and force feedback perturbations for limb movement (e.g., Shadmehr and Mussa-Ivaldi 1994), auditory feedback perturbations of speech, particularly perturbations of vowel quality, elicit significantly smaller partial compensations (e.g., MacDonald et al. 2010; Mitsuya et al. 2015). There are a number of

✉ K. G. Munhall
kevin.munhall@queensu.ca

[1] Department of Speech and Hearing Sciences, University of Washington, 1417 NE 42nd Street, Seattle, WA 98105-6246, USA

[2] Department of Psychology, Queen's University, Humphrey Hall, 62 Arch St., Kingston, ON K7L 3N6, Canada

possible explanations for these data including reduced gain for auditory feedback, multiple timescales of adaptation (see Hopp and Fuchs 2004 for a discussion of partial adaptation in saccadic perturbations) and differences in sensory reliance (Purcell and Munhall 2006; Tremblay et al. 2008; Nasir and Ostry 2006; Lametti et al. 2012). Lametti et al. (2012) showed that some subjects, weighted somatosensory feedback more than auditory feedback in speech, while others showed the opposite pattern.

In addition, it has been suggested that the size of compensation for vowel formant perturbations is influenced by the categorical nature of vowels (Mitsuya et al. 2011, 2013, 2015; Reilly and Dougherty 2013; Niziolek and Guenther 2013). For example, Mitsuya and his colleagues (e.g., Mitsuya et al. 2013) have shown that larger compensation behavior is elicited if the perturbed feedback sounds more like a categorically different vowel. This result suggests that compensatory behavior is not independent of the category boundaries of speech. While small, sub-categorical perturbations elicit compensations to maintain the accuracy of speech production (e.g., MacDonald et al. 2010), suggesting that the sensorimotor system in speech is also structured to help maintain the categorical representations.

If production control of a speech sound is influenced by a speaker's language sound categories, then a sound that is not categorized linguistically may be controlled differently. Consequently, responses to auditory feedback perturbations would depend on the linguistic status of the sound being produced. To test this question, a method is required to examine a vocal production that can be produced as a linguistic target and a non-linguistic target. In speech perception, a similar need has been addressed with the use of sine wave speech (Remez et al. 2001). This minimal acoustic signal, where the rich acoustic structure of speech is replaced with a set of time-varying sine waves can be perceived as speech or electronic music depending on instructions (Remez et al. 1981). This perceptual bistability of sine wave speech allows the examination of how comprehension influences the broad neural networks involved in auditory perception (e.g., Khoshkhoo et al. 2018) and as well permits the study of the basic structure of speech representations (e.g., Remez et al. 2001).

Here, we use a novel apparatus with which sounds were produced manually with both speech and nonspeech targets. The apparatus was made with an electrolarynx, a device designed for laryngeal cancer patients who have been treated by laryngectomy, and a pliable tube (Fig. 1). This device has a similar construction to von Kempelen's speaking machine and Wheatstone's "Euphonia" (see Dudley and Tarnoczy 1950). The electrolarynx served as a voice source and the tube as a filter. Thus, the tool functioned as a mechanical version of the source/filter theory of speech sound production (Fant 1960), although our participants were not given such information prior to the experiment. By constricting (i.e., squeezing) the pliable tube by hand while pressing the voicing button of the electrolarynx, this apparatus produced a sound with various resonance characteristics. The sound could be heard as a buzzing sound, or a vowel-like sound depending on instructions. Thus, a unique opportunity was provided to compare acoustically similar sounds with different linguistic statuses and to test how the production of the sound is controlled depending on its linguistic status.

The current study investigated whether the linguistic status of a sound being produced would influence the control of the sound. This was measured by the response to real-time formant perturbation of the auditory feedback participants received while manually producing either a vowel-like sound or a nameless acoustic target. A formant has a variety of meanings in speech research (Titze et al. 2015), but we

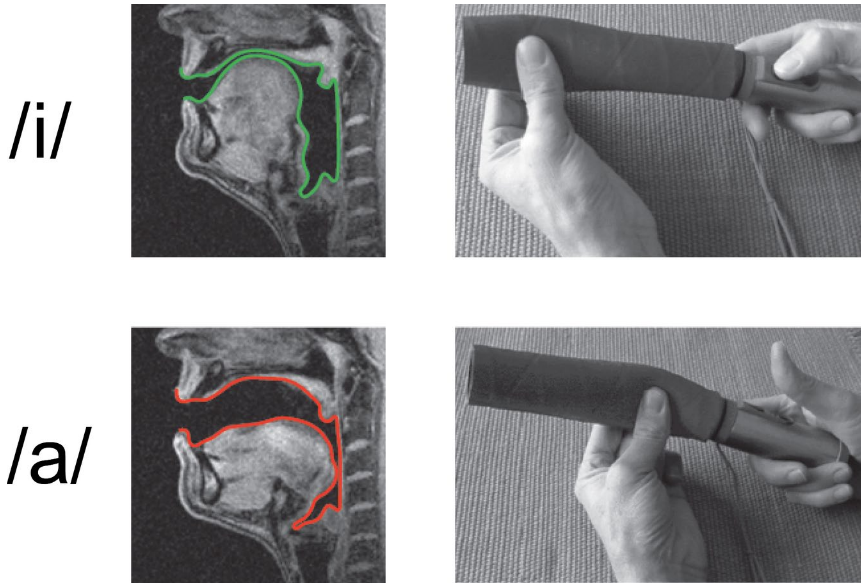

Fig. 1 Comparison of natural vocal tract and the artificial vocal tract during vowel production—midsagittal MRI images of the vocal tract producing a vowel. The oral cavity configuration is traced in colored lines (left). Artificial vocal tract producing a similar sounding vowel (right)

are using it here to refer to the resonant frequencies of an acoustic tube. The column of the air in the vocal tract and in our device resonates at a set of frequencies that are determined by the size and shape of the tube. We estimate these resonances from the emitted acoustics and call the spectral peaks formants too. In speech, vowels are produced by different vocal tract configurations (see Fig. 1) and the resulting acoustics show different patterns of peaks in the spectral envelope. These spectral peaks are labeled from the lowest to the highest in frequency as the first (*F*1), second (*F*2), third (*F*3) formants, etc. The first two formants have been the focus of considerable research concerning the perception of vowels and consonants.

Participants were randomly split into one of the two groups. One group was instructed to produce a vowel-like sound (Linguistic group; LIN hereafter) using the novel apparatus. The other group, on the other hand, was instructed to produce a nameless acoustic target, a sample of which was given by an experimenter (acoustic; ACO, hereafter). While participants were producing the target sound, the resonance characteristics (the first and the second formants; *F*1 and *F*2 hereafter, respectively) of their auditory feedback were perturbed in real-time, and their change in sound production (i.e., change in resonance characteristics) in response to the perturbation was compared across the groups.

Methods

Participants

Thirty male native Canadian English speakers in the community of Queen's University in Ontario, Canada participated in the current study. All participants spoke English as their first and primary language. Fourteen were assigned to be in LIN (mean age 20.1 years) and the remaining 16 were assigned to be in ACO (mean age 23.8 years). None reported having hearing and/or speech impairment. All had normal audiometric hearing thresholds within the range of 500–4000 Hz (≤ 20 dB hearing level). The experimental protocol was approved by the General Research Ethics Board at Queen's University. Informed consent was obtained from all individual participants included in the study.

Materials

The apparatus: artificial vocal tract

An electrolarynx (Servox Inton by Servona) and a pliable and elastic synthetic rubber tube (17.0 cm in length with an inside diameter of 3.18 cm) were used. The electrolarynx was inserted to the rubber tube, which made the length of the artificial "vocal tract" approximately 15.0 cm. This length is similar to the vocal tract length of young adult males (19–25 years; see Fitch and Giedd 1999). The fundamental frequency (*F*0) of the electrolarynx was set approximately at 80 Hz, much lower than the average *F*0 of a male voice (110–120 Hz). This was our effort to prevent the sound from being processed as human-like unless the participants were given instructions that the sound might be heard as a human-like sound.

Examples of the first three formants (i.e., *F*1, *F*2, and *F*3) that the artificial vocal tract produced are depicted in Fig. 2. The purpose of the figure is solely to show that both constriction depth and constriction location influence the formant patterns produced by the device. Vowel categories correlate with *F*1 and *F*2, which are generally related to the phonetic value of vowel height and frontedness of the tongue constriction (respectively). However, different articulatory maneuvers can produce a similar formant structure (e.g., Perkell et al. 1993, 1997). Similarly, formants can also be controlled in more than one way with the artificial vocal tract used in the current study.

Equipment

Equipment used for signal processing in this experiment was the same as that reported in Purcell and Munhall (2006). Participants sat in front of a computer monitor with headphones (Sennheiser HD 265) in a sound-attenuated booth (Industrial Acoustics Company). A headset microphone (Shure WH20) was placed in front of the participants approximately 10 cm from the open end of the artificial vocal tract. The microphone signal was amplified (Tucker-Davis Technologies MA 3 microphone amplifier), low-pass

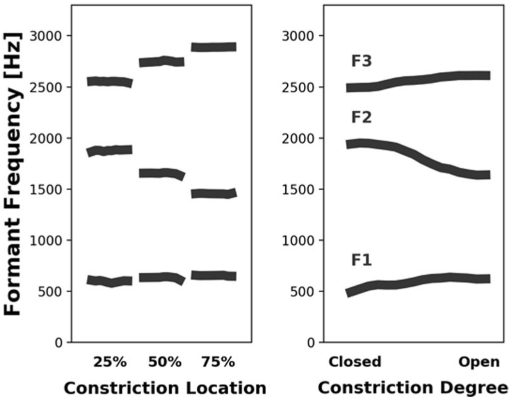

Fig. 2 Profile of the first three formants of the sound (left) produced by the artificial vocal tract with three different constriction locations (approximately 25%, 3.75 cm; 50%, 7.5 cm; 75%, 11.25 cm) along the tube length scaling from 0% (0 cm; sound source) to 100% (15 cm, open end), while the constriction degree was approximately the same and with varying degrees of constriction (right) continuously from open to closed at 50% (7.5 cm) location

filtered with a cutoff frequency of 4.5 kHz (Krohn-Hite 3384 filter), digitized at 10 kHz, and filtered in real time to produce formant shifts (National Instruments PXI-8106 embedded controller). The manipulated signal was then amplified and mixed with speech-shaped noise (Madsen Midimate 622 audiometer). The processed signal was presented over the headset at approximately 80 dBA sound pressure level (SPL) with 50 dBA SPL speech-shaped noise.

Procedure

General procedure

Both ACO and LIN completed a production experiment in which they were instructed to produce a target sound by manually manipulating the artificial vocal tract. LIN was instructed that they would produced an/ɛ/like sound, while ACO produced a nameless sound that had similar acoustic characteristics to the sound LIN produced (example of which was presented by the experimenter prior to the experiment). The LIN also completed a perception task before the production experiment. The order of the perception and production experiments was fixed, because the premise of the perception task was (1) to introduce the sounds produced by the artificial vocal tract as vowel-like sounds, so that the participants in LIN would form a linguistic target for the production experiment and (2) to measure their ability to perceive the sounds as vowel-like sounds.

Perception experiment (LIN only)

A categorization task was performed using three prerecorded sounds produced by the same artificial vocal tract. The three tokens were produced by a single constriction of approximately the same degree at three different locations of the pliable tube of the artificial vocal tract. The formant values of these sounds are summarized in Table 1, and their distribution in the $F1/F2$ acoustic space is shown in Fig. 3. Sound 1 was labeled as the vowel in “had”, sound 2 as the vowel in “head”. While these two sounds had vastly different $F2$ values, their $F1$ values were very similar. Sound 3’s $F1$ was much lower than sounds 1 and 2, while its $F2$ was somewhere in the middle of the other two sounds. This is a slightly centralized variant of/ɪ/found in many dialects of English (see Wells 1982), especially when/ɪ/is produced in continuous or coarticulated speech. We labeled this sound as “the vowel in ‘hid’”, so that the participants could have a label to distinguish this sound from the other sounds.

Table 1 First formant ($F1$) and the second formant ($F2$) in Hertz (Hz) values of the three sounds used in the perception experiment

Sound 1 (had)		Sound 2 (head)		Sound 3 (hid)	
$F1$ (Hz)	$F2$ (Hz)	$F1$ (Hz)	$F2$ (Hz)	$F1$ (Hz)	$F2$ (Hz)
696.90	1525.53	662.13	1641.61	582.21	1569.31

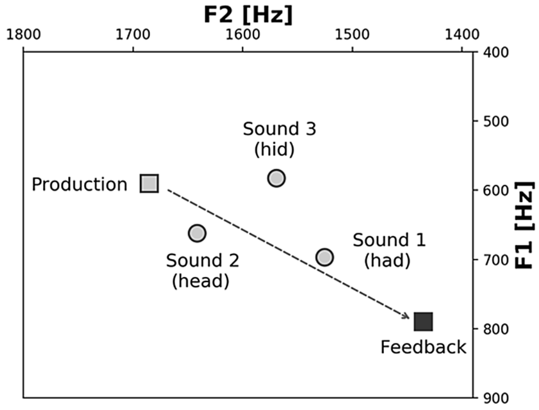

Fig. 3 Distribution of the sounds in the $F1/F2$ acoustic space. The circles indicate the stimuli used in the perception experiment. The light square indicates the average formant values of the sound produced by our participants during the production part of the experiment (without perturbation) using the artificial vocal tract, whereas the filled square indicates the feedback with a perturbed feedback with $F1+200$ Hz and $F2-250$ Hz based on the average production

The amplitude and duration of these tokens were normalized. Each of the three tokens was presented 20 times in a random order, and participants’ task was to identify which of the three vowels that they just had heard by pressing one of the three keys on a computer keyboard.

Production experiment

For both LIN and ACO group, the task of the production experiment was to produce a consistent sound by squeezing the rubber tube of the artificial vocal tract when a visual prompt (+) was presented on the monitor. Both groups were given a sample sound produced by an experimenter for several times prior to the experiment. The description of the target sound, however, differed across the groups. For LIN, the task was to mimic and produce a vowel-like sound found in the word “head” (hereafter/ɛ/) that the experiment had just produced. Participants in ACO, on the other hand, were instructed to mimic and produce the sample sound produced by the experimenter. Both groups were given approximately 1 min of a familiarization period, during which they learned to use the artificial vocal tract and practiced producing the target sound. After this period, a practice session with 30 trials started. The visual prompt was presented for 2.5 s with an inter-stimulus interval (ISI) of 1.5 s. Participants were specifically instructed to let go of the manual constriction completely during the ISI so that their fingertips and/or hand would not be touching the rubber tube. This was to ensure that consistent productions of the target sound would not be simply due to a static articulation. In the practice session, some participants simply released the tube and others released and moved their hands away. However, the participants' hand movements during the experiment were not monitored. No perturbation was applied to their auditory feedback during the practice session.

After the practice session, a fabric screen was draped in front of the participants. They were instructed to put their arms under the screen, so that they could not see their hands, while they were still able to see the visual prompt on the monitor above the screen. The experimental session consisted of 120 trials with four experimental phases. The first 20 trials (baseline), participants received normal feedback (i.e., no auditory feedback manipulation but with a speech-shaped masking noise). In the second phase (trials 21–70, ramp), incremental changes of 4 Hz for *F*1, and − 5 Hz for *F*2 were applied for each trial over the course of the 50 trials. At the end this phase, the maximum perturbation of 200 Hz for *F*1 and − 250 Hz for *F*2 was applied. With a naturally spoken vowel/ɛ/, this perturbation would make the formant structure more like that of a vowel/æ/. In the third phase (trials 71–90, hold), the maximum perturbation was held constant. Finally, in the last phase (trials 91–120, return), the perturbation was removed all at once at the 91st trial and no perturbation was applied to the auditory feedback for the remaining trials of the experiment; thus, the participants received abrupt change from the maximum perturbation to normal feedback (Fig. 4).

When feedback perturbations are being introduced in real time in an experimental series, the resulting behavior can be due to both real-time corrective compensations and short-term adaptations in the sensorimotor system. Without removal of the feedback or return of sensory feedback to normal, it is difficult to distinguish the contributions of each process. The purpose of this manuscript is not to determine the potential contributions of sensory-based adjustments and learned adjustments to the observed behavior. We are interested in how instructions or experimental set can change the behavior that results from auditory feedback perturbations. As a result, we will use a single term, compensation, to refer to behavior in the ramp and hold phases and acknowledge that the neural processes underlying this behavioral change may also involve short-term learning.

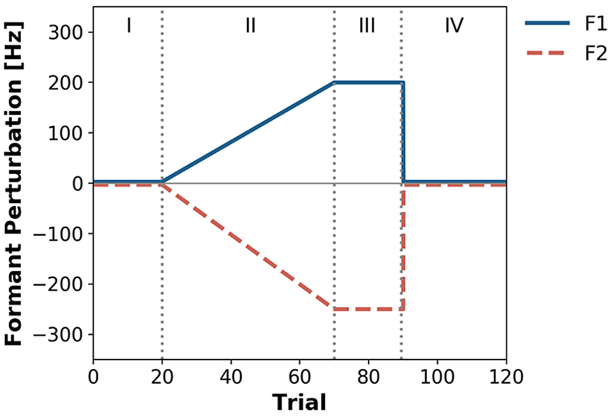

Fig. 4 Feedback perturbation applied to the first (solid blue) second (dotted red) formants. The roman numerals indicate the four experimental phases: (1) baseline; (2) ramp; (3) hold; and (4) return. The vertical dash lines denote the boundaries of the phases

Acoustic processing and analysis

Online formant shifting and detection of voicing

Detection of a sound segment was done using a statistical, amplified-threshold technique, and the real-time formant shifting was done using an infinite impulse response filter. The Burg algorithm (Orfanidis 1988) was used to estimate formants and this was done every 900 µs. Based on these estimates, filter coefficients were computed, such that a pair of spectral zeroes was placed at the location of the existing formant frequency and a pair of spectral poles was placed at the desired frequency of the new formant.

Offline formant analysis

Offline formant analysis was done using the same method reported in Munhall et al. (2009). An automated process estimated sound segment boundaries in each production based on the harmonicity of the power spectrum. These estimates were then manually inspected and corrected if required. An algorithm similar to that used in online shifting was used to estimate the first three formants from the first 25 ms of a sound segment. The estimation of formants was repeated with the window of 1 ms until the end of the sound segment. For each segment, an average value for each formant was calculated from 40 to 80% of the segment duration. These estimates were then inspected, and if any estimates were incorrectly categorized (e.g., *F*1 being mislabeled as *F*2, etc.), they were corrected by hand.

Results

Perception experiment (only LIN)

All three sounds were labeled correctly as the intended English vowel category with "had,/æ/" having the highest accuracy (92.9%) followed by "head/ɛ/" (80.7%) and "hid/ɪ/" (76.1%). For each of the three sounds, the percentage of correctly identified sounds was tested against the percentage of confusion responses given for the other two vowels. When sound 1 (had,/æ/) was presented, our participants were reliably able to identify the sound as/æ/[$\overline{X}$ = 92.9%, standard error (SE hereafter) = 2.3%] and this proportion was significantly different from other response types [sound 2 (head/ɛ/): $\overline{X}$ = 4.6%, SE = 1.7%, $t(13) = 23.20$, $p < 0.0001$; sound 3 (hid/ɪ/): $\overline{X}$ = 2.5%, SE = 1.5%, $t(13) = 25.85$, $p < 0.0001$]. With sound 2 (head/ɛ/), participants identified the sound mostly as/ɛ/($\overline{X}$ = 80.7%, SE = 5.7%) and this proportion was significantly different from the/æ/responses [$\overline{X}$ = 5.00%, SE = 1.7%, $t(13) = 10.89$, $p < 0.0001$]. Although there were some "head" responses with sound 3 (hid/ɪ/), ($\overline{X}$ = 22.1%, SE = 6.39), significantly more responses were "hid" [$\overline{X}$ = 76.1%, SE = 6.6%; $t(13) = 4.16$, $p = 0.0011$]. This indicated that our participants could identify and distinguish all sounds reliably. Most importantly, sound 1 (/ɛ/-like) and sound 2 (/æ/-like) were categorized with little confusion.

Production experiment

The baseline average of formant values was calculated for $F1$ and $F2$, for each of the participants based on the last 15 trials of the baseline phase (i.e., trials 6–20). To minimize the influence of the subjects adjusting to the task and hearing feedback through headphones, the first five trials were not included in the Baseline calculations. The raw formant values in Hz were normalized by subtracting the speaker's baseline average from the value of each trial. The group averages of the normalized $F1$ and $F2$ data are plotted in Fig. 5. The magnitude of change for each participant was calculated by averaging the normalized formant values of the last 15 trials of the hold phase. To measure the compensation at the largest perturbation, the steady-state estimate of the hold phase did not include the first five trials which could be influenced by the transition from ramp to hold.

As can be seen in Fig. 5, the changes in formant production between the two groups were vastly different. While both groups changed their behavior with the different feedback conditions, the direction and pattern of change were not the same for the two instruction conditions. Two-way analyses of variance (time: baseline, hold × instruction: LIN, ACO) were carried on each of the formants separately. Both formants showed significant time × instruction interactions [$F1$: $F(1,28) = 10.64$, $p = 0.003$; $F2$: $F(1,28) = 9.67$, $p = 0.004$]. Simple main effects showed that these interactions were caused by the significant differences in the hold phase between the two subject groups, see Fig. 5.

The above analyses clearly indicate that the two groups behaved differently in their $F1$ and $F2$ production in response to real-time formant perturbation. However, $F1$ and $F2$ were analyzed independently, making it unclear how each participant responded in both $F1$ and $F2$ simultaneously. To capture each participant's change in both formants simultaneously, the distribution of compensation data (the average of the last 15 trials of the hold phase) was plotted in the $F1/F2$ acoustic space (Fig. 6). It appeared that participants in LIN tended to show behavior that was considered compensation, whereas participants in ACO tended to cluster more in the "following" area. A Kuiper's test, which examines a directional distribution of data across groups, was indeed significant ($V = 0.74$, $p = 0.002$), demonstrating that the two groups changed formant values in different directions when an identical auditory perturbation was delivered.

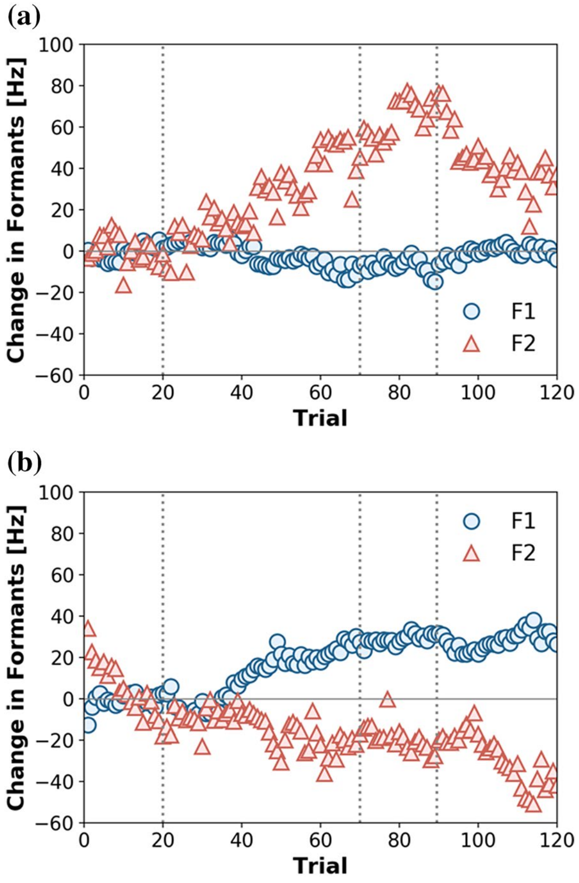

Fig. 5 Group average of change in the first and second formant productions. The blue circles are the first formant values, whereas the red triangles are the second formant values of **a** linguistic condition (LIN) and **b** acoustic condition (ACO). The vertical dotted lines denote the boundaries of the experimental phases

Producing a sound using an apparatus like the one used here is mechanically and cognitively similar to playing a musical instrument; thus, the length of musical training (or the lack thereof) that our participants had had might have influenced (1) the variability of the sounds that they produced and (2) their response to the perturbation. The groups did not differ in the self-reported length of music training they had received [$t(28) = 0.59$, $p = 0.56$; LIN: 4.82 years, standard deviation (SD hereafter) = 4.9 years; ACO: 5.68 years, SD = 3.0 years]. In addition, we tested for correlations between the amount of musical training and production

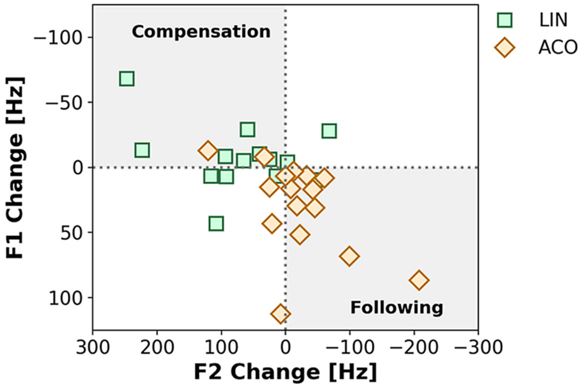

Fig. 6 Distribution of compensation data in the *F*1/*F*2 acoustic space, where the *x*-axis is reversed *F*2, and the *y*-axis is reversed *F*1. The green squares are the data points of the linguistic condition (LIN), whereas the orange diamonds represent the acoustic condition (ACO). The shaded area in the left upper quadrant indicates responses to the perturbation that are considered "compensations", while the shaded area in the lower right quadrant is where the responses are considered "following"

variability as indexed by the Baseline standard deviation. No significant correlation was found both in LIN [$F1$: $r(14) = -0.05$, $p = 0.88$; $F2$: $r(14) = 0.01$, $p = 0.98$] nor in ACO groups [$F1$: $r(16) = -0.23$, $p = 0.40$; $F2$: $r(16) = 0.14$, $p = 0.62$]. Similarly, compensation magnitude yielded no significant correlations with their musical training/experience in ACO [$F1$: $r(16) = 0.01$, $p = 0.97$; $F2$: $r(16) = 0.30$, $p = 0.26$], and in LIN [$F1$: $r(14) = -0.26$, $p = 0.37$; $F2$: $r(14) = 0.27$, $p = 0.36$].

Finally, during debriefing only four participants in LIN reported a slight change(s) in auditory feedback during the experiment, and none of them correctly identified the nature of the perturbation (i.e., change in "vowel" quality). In contrast, 14 out of the 16 participants in ACO reported that they thought they had been making different sounds during the middle part of the experiment. Out of these 14 participants, 12 of them reported that they had consciously tried to maintain consistent productions by moving the constriction location and/or its degree. None in this group reported that the sound they were producing sounded like a vowel when they were explicitly asked.

Discussion

The purpose of speech production is to deliver an intended message to the listener. To keep the message intact, the speech motor control system must be sensitive to variation around the phonemic categories of the language (Mitsuya et al. 2011, 2013; Niziolek and Guenther 2013; Reilly and Dougherty 2013). The current investigation examined this directly using a novel apparatus to manually produce a sound, which could be heard as a vowel-like sound or an acoustic buzz sound depending on the particular instructions. Production control was measured by participants' response to the real-time formant perturbation of the auditory feedback they received. The results unequivocally showed that the different sets of instructions given to our participants—whether they would be producing a vowel-like sound or mimicking an acoustic exemplar, made the response to the same feedback manipulation quite different.

The response pattern exhibited by LIN is relatively easy to interpret. The participants in this group tended to change the formant values to offset the perturbation (i.e., compensate), similar to the compensatory response observed with a naturally produced vowel (Purcell and Munhall 2006; Mitsuya et al. 2015). Even though they had little experience with the apparatus used in the experiment, their response indicates that they were able to learn the relationship between different manual articulations; however, subtle they might have been, and the resulting sounds remarkably quickly. Except for the initial 1-min exploration period, our participants were not able to freely manipulate the apparatus to sample the full spectrum of the sounds that it could produce. Thus, it is evident that the articulation–acoustic mapping was learned from the production variability during the 30 practice trials and the 20 baseline trials of the experiment. With the new mapping, the long-term knowledge of sound categories drove the compensatory behavior.

Another similarity between the LIN group' behavior and what has been reported in natural vowel formant production is participants' unawareness of the perturbation, and whether awareness was a prerequisite for the compensatory behavior. Purcell and Munhall (2006) reported that only 8% of their participants correctly identified a change in vowel quality when the formant of their auditory feedback was perturbed, and their awareness was not related to compensation behavior. Similarly, in Mitsuya et al.'s (2013) *F*2 perturbation study, only 2 of 35 participants reported the exact nature of the perturbation, and they did not compensate more or less compared to the rest of the participants. These findings strongly suggest that articulatory adjustments to maintain linguistic targets based on auditory feedback do not require active cognitive or conscious effort. In fact, it has been shown that even when subjects are made aware of the perturbation and are instructed to consciously ignore the change in auditory feedback, compensations are still observed (Munhall et al. 2009).

The behavior exhibited by the participants in ACO is rather difficult to explain. First, and most obviously, they exhibited following behavior—they tended to change their acoustics in the direction of the perturbation. In feedback-perturbation experiments, a following response is not the most common observation, but it is not an unheard of response. As MacDonald et al. (2010) summarized for formant perturbations, the size of compensatory responses is, in general, unimodally distributed from the negative (following) to the positive (compensation) with most of the data on the positive compensatory side. Considering that the direction of the change by ACO was statistically different from that of LIN, ACO's behavior is not part of the typical compensation behavior.

Second, and more subtly, ACO's behavior upon removal of the perturbation was different from a typical de-adaptation pattern shown by subjects in a normal speech perturbation study. De-adaptation usually shows a reversal of compensatory behavior, which is generally thought to indicate that the newly learned motor-sensory mapping is decaying or being undone when normal feedback is restored. Our participants in ACO appeared to respond to the offset of perturbation but then shortly after that they resumed diverging from the baseline value again. We do not have a clear explanation for this behavior; however, based on (1) a consistent following response to perturbation and (2) no de-adaptation, we can conclude that the production control of a non-linguistic sound is characteristically different from that of a linguistic target. It should be noted that it is possible that participants with extensive musical training may not act the same as the ACO group tested here, although musical experience in our study did not predict our results.

An unresolved issue with our methods is that LIN group received more sound samples (60 total, 20 of which were the target vowel sound) as part of the perception task prior to the production experiment. This exposure alone might have contributed to the difference in behavior that we observed across the groups, and some may consider this a limitation. However, two things need to be considered. First, the perception task was performed before the production experiment, and the participants in LIN did not know the nature of the subsequent production; thus, participants had no prior knowledge that they would be producing one of the three stimulus sounds that they were categorizing. Second, what the current study examined was the effect of linguistic status of the intended sound production. The extra exposure of the sounds as labeled speech items among LIN ensured that the sounds that they were given were associated as part of the rich representation of the already existing vowel categories. Presumably, ACO could have received the same number of sound samples randomly presented as passive listening (without making any judgment). However, we did not know if they would start perceiving the buzzing sound as vowel-like prior to the production task, which would have compromised the study design. To avoid this, we opted not to present those sounds to ACO. However, in future studies, a labelling task for the ACO group with neutral names (e.g., sound 1, sound 2…) could be tested to eliminate any concerns about more exposure being a cause of the sound mapping.

Task dependence of vocal tract movements is not a new notion (see Bunton 2008 for a review). For example, opening of the jaw appears similar across speech articulation and mastication; however, the movement and its function depend on the task at hand. Jaw opening for speech articulation needs to be intricately coordinated with the movements of respiration, larynx, velum, tongue, and lips. On the other hand, a different coordination of movements is required for mastication (e.g., Gentil and Gay 1986; Ostry and Munhall 1994). In the current study, the function of the gripping movements was assumed to be the same between the groups at least during the baseline phase. However, the actual movements of the grip were not measured, and we have no way of knowing exactly how the manual articulation was coordinated.

Our failure to monitor aspects of the manual behavior is, in retrospect, a lost opportunity. One of the challenges of speech research is the difficulty in measuring the complex kinematics and dynamics of the orofacial system. The paradigm tested here presents a much simpler acoustic tube and the articulations carried out by the hand are visible. Given the similarity of the compensatory behavior when auditory feedback is perturbed, the kinematic and force characteristics of vowel production could in the future be studied in this simple speech analogue system.

When we produce speech, mental representations of language sounds are transformed into physical entities—the movements of the vocal tract and the resultant acoustics. One of the unique aspects of spoken language as a movement system is that the movements are coordinated to produce categories of sounds. Those categories are represented in the mind in complex ways—in multiple dimensions (e.g., a host of acoustic features) and in multiple modalities (e.g., auditory, visual, somatosensory, etc.). This makes speech production the product of a detailed system of such representations. The current study demonstrated that the access to such representations (or the lack thereof) influences how the production of a sound is controlled. Our data show that people who produced a linguistic target responded to perturbations in a similar manner to responses reported in natural speech production. It is likely that the perturbed auditory feedback was organized around strong representations (e.g., phonemically), while the ACO group did not have such consistent internal representations for their targets. This weaker representation presumably made the ACO task more difficult as well different in organization.

Taken together, the current findings provide (1) insights into the importance of understanding the nature of the representations that drive speech production and how they relate to the targets of production and (2) an innovative and unique method using a manual apparatus to examine the nature of speech production representation and its control parameters.

Acknowledgements This research was supported by a Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC), and partially supported by The National Institute on Deafness and Other Communication Disorders (NIDCD) Grant R01 DC014510. The authors wish to thank Dr. Mark Tiede for the MRI images used in Fig. 1.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

Bunton K (2008) Speech versus nonspeech: different tasks, different neural organization. Semin Speech Lang 29(4):267–275. https://doi.org/10.1055/s-0028-1103390

Burnett TA, Freedland MB, Larson CR, Hain TC (1998) Voice *f*0 responses to manipulations in pitch feedback. J Acoust Soc Am 103:3153–3161

Casserly ED (2011) Speaker compensation for local perturbation of fricative acoustic feed- back. J Acoust Soc Am 129:2181–2190

Chiba T, Kajiyama M (1941) The vowel: its nature and structure. Kaiseikan, Tokyo

Cowie RJ, Douglas-Cowie E (1992) Postlingually acquired deafness: speech deterioration and the wider consequences. Mouton De Gruyter, New York

Dudley H, Tarnoczy TH (1950) The speaking machine of Wolfgang von Kempelen. J Acoust Soc Am 22:151–166

Fant G (1960) Acoustic theory of speech production. Mouton, The Hague

Fitch WT, Giedd J (1999) Morphology and development of the human vocal tract: a study using magnetic resonance imaging. J Acoust Soc Am 106:1511–1522

Gentil M, Gay T (1986) Neuromuscular specialization of the mandibular motor system: speech versus non-speech movements. Speech Commun 5:69–82

Guenther FH (2016) Neural control of speech. MIT Press, Cambridge

Hickok G (2014) The architecture of speech production and the role of the phoneme in speech processing. Lang Cogn Neurosci 29:2–20

Hopp JJ, Fuchs AF (2004) The characteristics and neuronal substrate of saccadic eye movement plasticity. Prog Neurobiol 72:27–53

Houde JF, Jordan MI (1998) Sensorimotor adaptation in speech production. Science 279:1213–1216

Jones JA, Munhall KG (2000) Perceptual calibration of *F*0 production: evidence from feedback perturbation. J Acoust Soc Am 108:1246–1251

Khoshkhoo S, Leonard MK, Mesgarani N, Chang EF (2018) Neural correlates of sine-wave speech intelligibility in human frontal and temporal cortex. Brain Lang 187:83–91

Lametti DR, Nasir SM, Ostry DJ (2012) Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. J Neurosci 32:9351–9358

MacDonald EN, Goldberg R, Munhall KG (2010) Compensation in response to real-time formant perturbations of different magnitude. J Acoust Soc Am 127:1059–1068

Mitsuya T, MacDonald EN, Purcell DW, Munhall KG (2011) A cross-language study of compensation in response to real-time formant perturbation. J Acoust Soc Am 130:2978–2986

Mitsuya T, Samson F, Ménard L, Munhall KG (2013) Language dependent vowel representation in speech production. J Acoust Soc Am 133:2993–3003

Mitsuya T, MacDonald EN, Munhall KG, Purcell DW (2015) Formant compensation for auditory feedback with English vowels. J Acoust Soc Am 138:413–424

Munhall KG, MacDonald EN, Byrne SK, 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

Nasir SM, Ostry DJ (2006) Somatosensory precision in speech production. Curr Biol 16:1918–1923

Niziolek CA, Guenther FH (2013) Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. J Neurosci 33:12090–12098

Orfanidis SJ (1988) Optimum signal processing: an introduction. McGraw-Hill, New York

Ostry DJ, Munhall KG (1994) Control of jaw orientation and position in mastication and speech. J Neurophysiol 71:1528–1545

Perkell JS, Matthies ML, Svirsky MA, Jordan MI (1993) Trading relations between tongue-body raising and lip rounding in production of the vowel/u: a pilot "motor equivalence" study. J Acoust Soc Am 93:2948–2961

Perkell J, Matthies M, Lane H, Guenther F, Wilhelms-Tricarico R, Wozniak J, Guiod P (1997) Speech motor control: acoustic goals, saturation effects, auditory feed- back and internal models. Speech Commun 22:227–250

Purcell DW, Munhall KG (2006) Adaptive control of vowel formant frequency: evidence from real-time formant manipulation. J Acoust Soc Am 120:966–977

Reilly KJ, Dougherty KE (2013) The role of vowel perceptual cues in compensatory responses to perturbations of speech auditory feedback. J Acoust Soc Am 134:1314–1323

Remez RE, Rubin PE, Pisoni DB, Carrell TD (1981) Speech perception without traditional speech cues. Science 212:947–949

Remez RE, Pardo JS, Piorkowski RL, Rubin PE (2001) On the bistability of sine wave analogues of speech. Psychol Sci 12:24–29

Schenk BS, Baumgartner WD, Hamzavi JS (2003) Effects of the loss of auditory feedback on segmental parameters of vowels of postlingually deafened speakers. Auris Nasau Larynx 30:333–339

Shadmehr R, Mussa-Ivaldi FA (1994) Adaptive representation of dynamics during learning of a motor task. J Neurosci 14:3208–3224

Shiller DM, Sato M, Gracco VL, Baum SR (2009) Perceptual recalibration of speech sounds following speech motor learning. J Acoust Soc Am 125:1103–1113

Taylor AM, Rebby D (2010) The contribution of source-filter theory to mammal vocal communication research. J Zool 280:221–236

Titze IR, Baken R, Bozeman K, Granqvist S, Henrich N, Herbst C, Howard D, Hunter E, Kaelin D, Kent R, Kreiman J, Kob M, Lofqvist A, McCoy S, Miller D, Noe H, Scherer R, Smith J, Story BH, Svec J, Ternstrom S, Wolfe J (2015) Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. J Acoust Soc Am 137:3005–3007

Todorov E (2004) Optimality principles in sensorimotor control. Nat Neurosci 7:907–915

Tremblay S, Houle G, Ostry DJ (2008) Specificity of speech motor learning. J Neurosci 28:2426–2434

Waldstein RS (1990) Effects of postlingual deadness on speech production: implications for the role of auditory feedback. J Acoust Soc Am 88:2099–2114

Wells JC (1982) Accents of English, vol 1. Cambridge University Press, Cambridge

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.