Please Say What This Word Is – Vowel-extrinsic normalization in the sensorimotor control of speech

Nicolas J. Bourguignon[1,2,4], Shari R. Baum[3,4], Douglas M. Shiller[1,2,4]

[1]School of speech-language pathology and audiology, Université de Montréal, Canada; [2] CHU Sainte-Justine Research Centre, Montreal, Canada; [3]School of Communication Sciences and Disorders, McGill University, Montreal, Canada; [4]Centre for Research on Brain, Language and Music, Montreal, Canada.

Abstract

The extent to which the adaptive nature of speech perception influences the acoustic targets underlying speech production is not well understood. For example, listeners can rapidly accommodate to talker-dependent phonetic properties – a process known as *vowel-extrinsic normalization* – without altering their speech output. Recent evidence, however, shows that reinforcement-based learning in vowel perception alters the processing of speech auditory feedback, impacting sensorimotor control during vowel production. This suggests that more automatic and ubiquitous forms of perceptual plasticity, such as those characterizing perceptual talker normalization, may also impact the sensorimotor control of speech. To test this hypothesis, we set out to examine the possible effects of *vowel-extrinsic normalization* on experimental subjects' interpretation of their own speech outcomes. By combining a well-known manipulation of vowel-extrinsic normalization with speech auditory-motor adaptation, we show that exposure to different vowel spectral properties subsequently alters auditory feedback processing during speech production, thereby influencing speech motor adaptation. These findings extend the scope of perceptual normalization processes to include auditory feedback and support the idea that naturally occurring adaptations found in speech perception impact speech production.

Acknowledgments: This work was supported by the National Institutes of Health (NIDCD-R01DC012502), the Natural Sciences and Engineering Research Council (NSERC-Canada) and les Fonds Québécois de Recherche, Nature et Technologies (FQRNT-Québec). Thanks to

Malcolm Slaney (Machine Hearing Research Group, Google, Mountain View, CA) for providing the voice recordings used as stimulus materials for this research, and to John F. Houde for his thoughtful comments on the manuscript.

1. Introduction

Acoustic speech signals vary greatly between talkers (Peterson & Barney 1952), yet the processes of speech perception are equipped with adaptive mechanisms to reduce the impact of this variability. Evidence for rapid accommodation of the speech perceptual system to talker-dependent phonetic properties, such as foreign accents (Kraljic *et al*., 2008; Samuel & Kraljic, 2009) or differences in age and gender (Peterson & Barney, 1952), attests to the high degree of plasticity of speech perception in the face of individual phonetic idiosyncrasies. In the case of vowels, whose primary categorical determinants are their resonant frequencies (or *formants*: F1, F2, F3, etc., cf. Ladefoged, 2001), this perceptual accommodation is achieved, in part, by *vowel-extrinsic normalization*, whereby the target formants for different speech sound categories are adjusted on the basis of the entire range of formants for a given vocal tract (Johnson, 2008)[1]. An influential demonstration of this phenomenon emerged from Ladefoged & Broadbent's (1957, 1989) vowel identification experiments, in which listeners judged the identity of a word (e.g., "bit", "bet") following brief exposure to carrier-phrases (*Please say what this word is…*) containing different fundamental and formant frequency patterns. Their results revealed a clear influence of the carrier-phrase on subsequent vowel identification, indicating that listeners use spectral properties from the immediate context as a frame of reference for their analysis of subsequent vowels. Such normalization processes have since then been replicated in numerous studies (Ainsworth, 1975; Dechovitz, 1977; Nearey, 1989; Johnson, 1990) and support a perspective on speech perception as a context- and talker-contingent process (Dahan *et al.*, 2008; Goldinger, 1998; Nygaard *et al*., 1994; Sjerps *et al.*, 2013).

Within the study of the role of perceptual information in speech production (Hickok *et al.*, 2011), a question of interest concerns the extent to which context-dependent plasticity in speech perception may also transfer to production. There is a great deal of evidence for context-dependent variability in speech production related to factors intrinsic to the talker, including co-articulation with neighboring phonemes and context-dependent changes in speaking rate, stress

[1] Vowel-*extrinsic* normalization can be contrasted with the idea of vowel-*intrinsic* normalization (e.g., Miller, 1989) in which there is no perceptual adaptation *per se.* Rather, speech perception is presumed to take advantage of relatively stable relationships between acoustic properties (e.g., ratios of formants) that vary less between talkers. Both are supported empirically and considered complementary by many researchers.

and intonation patterns (e.g., Daniloff & Moll, 1968; Byrd, 2000; Miller *et al*., 1984). In contrast, evidence of speech production plasticity related to external factors, whereby auditory exposure to speech stimuli influences speech-motor patterns, is more modest. Exposure to syllables containing initial voiceless stop consonants (e.g., [pi]) has been shown to alter voice onset time in subjects' subsequent production of the same syllable (Cooper & Lauritsen, 1974). Similarly, the production of a phonemic string can be affected by visual exposure to faces articulating slightly different phonemic sequences (Gentilucci & Cattaneo, 2005; Kerzel & Bekkering, 2000). In contrast with these demonstrations, however, a substantial degree of stability in the acoustic targets of speech production remains in the face of external contextual variability, as evidenced by data from second language acquisition. For example, production of a phoneme contrast (e.g., changing voice-onset-time for distinguishing voiced vs. voiceless consonants) shows more limited differences between talkers' native and second language when the second language has been learned later in life (in adulthood), suggesting the reliance on a single articulatory phonetic inventory for both languages (Flege, 1991). Furthermore, changes in speech motor patterns following perceptual adaptation to non-native phonetic contrasts have been reported only after extended periods of laboratory-based training (Bradlow *et al*., 1997, 1999; Wang *et al*., 2003), if ever at all (Kraljic *et al*., 2008; Samuel & Kraljic, 2009).

In the particular case of vowel production, the apparent stability of acoustic speech targets is further supported by the observation that adult talkers seem to rely on stable and accurate acoustic-phonetic representations in the planning and maintenance of speech movements (Guenther, 2006; Houde & Nagarajan, 2011). The organization of speech production around well-defined acoustic targets has been the focus of numerous behavioral studies involving altered auditory feedback (AAF) during vowel production (e.g., Houde & Jordan 1998; Purcell & Munhall 2006a; Rochet-Capellan & Ostry, 2011; Tourville *et al*., 2008; Villacorta *et al*. 2007). For example, perturbations of vowel resonant properties (e.g., a decrease in F1) of a speaker's auditory feedback yield a compensatory change in speech output (F1 *increase*) when such perturbations are introduced unexpectedly during sustained phonation (e.g., Tourville *et al*., 2008; Purcell & Munhall, 2006b), or when auditory feedback changes are maintained over prolonged periods of word production practice (e.g., Houde & Jordan 1998; Villacorta *et al*., 2007). Neurophysiologically, the comparison between expected and perceived speech auditory feedback is associated with a dampening of auditory cortical responses during active production

relative to listening to speech under normal auditory feedback conditions (Houde *et al*., 2002). Furthermore, this dampening effect is reduced when speech auditory feedback deviates from the expected pattern, as when producing less prototypical productions (Niziolek *et al*., 2013) or when auditory feedback is altered or delayed (Christoffels *et al.*, 2007; Hashimoto & Sakai, 2003; Heinks-Maldonado *et al.*, 2006; Tourville *et al.*, 2008).

More recently, however, evidence has shown that the auditory-perceptual representations of vowels underlying speech motor control may not be as stable as previously thought, but rather can be directly influenced by short-term changes in auditory perceptual processing. Specifically, two recent studies (one involving adults and one with 5-7-year-old children) have used a brief, intensive period of reinforcement-based perceptual training to alter participants' categorization of a vowel contrast (between the vowels [ɛ], as in "head" and [æ] as in "had"; Lametti *et al*. 2014; Shiller & Rochon, 2014). The perceptual training was carried out prior to a test of speech motor adaptation to AAF involving the same vowel contrast. Both studies found that the perceptual modulation immediately transferred to participants' perception of auditory error during the AAF task, altering the amount of speech motor adaptation to a degree that was proportional to the perceptual shift. These findings potentially run counter to current models of speech motor planning and control, as they suggest that the sensorimotor processes guiding speech production are not, in fact, insulated from the plasticity that characterizes the auditory-perceptual representations of phoneme categories. Even more striking is the possibility raised by these studies that more automatic and ubiquitous forms of perceptual plasticity – such as vowel normalization for differences in talker characteristics – may also impact the sensorimotor control of speech production, notably by influencing talkers' perception of their own auditory-perceptual errors. Interestingly, vowel-extrinsic normalization has so far only been examined as a mechanism supporting the decoding of other talkers' speech, and has never been considered outside of the context of exogenous speech perception. The possibility that such talker-contingent normalization processes might alter the processing of self-generated auditory feedback, and hence speech motor control, would thus be notable for two reasons: first, it would extend, for the first time, the scope of perceptual normalization processes to include the domain of self-generated auditory feedback; second, it would provide strong support for the idea that short-term auditory-perceptual plasticity transfers to the sensory processes guiding speech motor control.

The present study directly examined the influence of vowel-extrinsic normalization processes on the sensorimotor control of speech production by combining two distinct paradigms: (1) the normalization of vowel perception to differences in formant properties of extrinsically presented speech (i.e., the approach developed by Ladefoged & Broadbent, 1957, 1989), and (2) adaptation of speech production to AAF. Participants read aloud visually presented words containing the vowel [ɛ] (e.g., "bet") under conditions of normal or altered auditory feedback. The real-time feedback alteration involved a decrease in F1 frequency, resulting in a vowel perceived to be closer to [ɪ] (e.g., "bit"). Immediately prior to the production of each word, participants heard a phrase spoken with one of two different formant patterns, simulating differences in vocal tract properties of different talkers. These stimuli are close replications of carrier-phrases previously shown to induce changes in the perception of vowels along the [ɛ-ɪ] continuum (Ladefoged, 1989, see below). As described schematically in Figure 1A, we predicted that the carrier-phrases would similarly influence participants' perception of their *own* vowel formants during word production, and that the resulting change would impact the degree of motor adaptation to their F1-altered feedback. Our specific prediction was that participants exposed to the carrier-phrase containing relatively *low* formant values would perceive their own altered vowel as comparatively *higher* in F1 (closer to the target vowel [ɛ]), thereby diminishing the perceived impact of the auditory feedback manipulation and reducing the required degree of motor adaptation (bright red upward arrow in Figure 1A). Conversely, participants exposed to the carrier-phrase containing relatively *high* formant frequencies would perceive their own altered vowel as comparatively *lower* in F1 (further from the target vowel [ɛ]), thereby enhancing the perceived impact of the auditory feedback manipulation and increasing the required degree of speech motor adaptation (bright blue upward arrow in Figure 1A). Empirical support for these predictions would indicate that the talker-dependent properties of acoustic speech signals perceived immediately prior to word production serve as a frame of reference for the perception of self-generated speech outcomes, thereby influencing speech motor patterns.

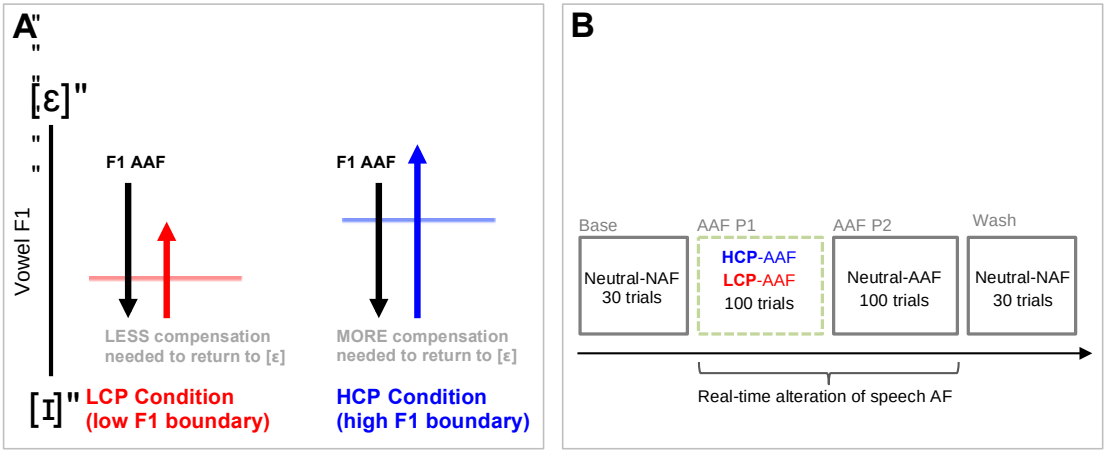

Figure 1 – (A) Illustration of the experimental manipulation and corresponding predicted group differences in the magnitude of adaptation to the altered F1 feedback (F1 AAF) as a function of exposure to the Low carrier-phrase (LCP, red) and High carrier-phrase (HCP, blue). The black downward arrows represent the magnitude and direction of the alteration in participants' F1 auditory feedback. The red and blue upward arrows represent the predicted magnitude and direction of the compensatory response to the feedback alteration in the speech output of LCP and HCP participants respectively: Specifically, for participants exposed to the LCP, the perceptual [ɛ-ɪ] boundary (light red horizontal line) is shifted lower in F1, resulting in subjects perceiving their productions as relatively high in F1 (closer to [ɛ]). With vowel perception shifted closer to the target vowel [ɛ], the required magnitude of the compensatory motor response needed to restore the vowel to [ɛ] under altered feedback conditions (bright red vertical arrow) is reduced. In contrast, for participants exposed to the HCP, the perceptual [ɛ-ɪ] boundary (light blue horizontal line) is shifted higher in F1, resulting in subjects perceiving their productions as relatively low in F1 (closer to [I]). With vowel perception shifted away from the target vowel [ɛ], the required magnitude of the compensatory motor response needed to restore the vowel to [ɛ] under altered feedback conditions (bright blue arrow) is increased. (B) Experimental protocol. The initial *Baseline* and final *Washout* phases correspond to words produced under normal auditory feedback (NAF) after hearing the Neutral carrier-phrase. The AAF-P1 and AAF-P2 phases correspond to words produced under AAF. AAF-P1 (green dashed box) is the phase where groups differ in terms of the carrier-phrase (i.e., HCP vs. LCP). This is therefore the phase during which participants are predicted to differ in their amount of adaptation to AAF. Groups are expected to converge in their magnitude of adaptation during AAF-P2 as the carrier-phrases are set back to *Neutral* in both groups.

2. Materials and Methods

2.1. Participants

Twenty male, native speakers of English (age 18-30) without history of speech, language or hearing disorders took part in the study. Sample sizes were based upon prior studies of sensorimotor adaptation in speech (in our lab and others) demonstrating significant group differences with 10-20 subjects in each condition (Bourguignon *et al*., 2014; Lametti *et al*., 2014; Purcell and Munhall, 2006a; Rochet-Capellan & Ostry, 2011; Shiller & Rochon, 2014; Villacorta *et al*., 2007). All participants passed a pure tone hearing screening (threshold < 30 dB HL at octave frequencies between 250 and 4000 Hz) and provided written informed consent prior to testing. Procedures were approved by the Institutional Review Board, Faculty of Medicine, McGill University.

2.2. Stimuli and group assignment

Participants were randomly assigned to a High carrier-phrase (*High-CP*) and a Low carrier-Phrase (*Low-CP*) group and underwent an identical series of tasks involving the production of monosyllabic words containing the vowel [ɛ] (*bet*, *gem*, *neck*, *mess, peck*, *pen*, *pet*, *tech* and *ten*), first under conditions of normal auditory feedback (NAF, 30 trials), and then during two periods of altered auditory feedback (AAF, 100 trials each). The real-time acoustic manipulation carried out during the AAF condition involved a decrease in F1 frequency, yielding a perception of the vowel [ɛ] as being closer to [ɪ] (see *Speech motor adaptation* below and Figure 1A for detail).

Each word production trial began with the auditory presentation of the carrier-phrase "*Please say what this word is…*". The phrases used in the study feature voice recordings of Peter Ladefoged (made available to us courtesy of Malcolm Slaney, Machine Hearing Research Group, Google, Mountain View, CA). Three different versions of the carrier-phrase were used in the experiment (*Neutral*, *Low* and *High*), characterized by systematic differences in F0 as well as in formant frequencies (see Table 1). The different F0 and formant values were obtained from the same speaker through changes in laryngeal and articulatory configuration (e.g., lip spreading) during speech production, yielding controlled yet naturalistic variations in vowel spectra (see

Ladefoged, 1989 for a similar method). The carrier-phrase labeled "Neutral" was produced without any articulatory modifications.

Table 1 – Average fundamental and formant frequencies characterizing the three carrier-phrases used in the present study. All parameters were estimated in *Praat* (Boersma, 1993). All units are in Hz.

	F0	**F1**	**F2**	**F3**
LOW	98.4	496.8	1706.3	2694.0
HIGH	114.2	566.6	1893.8	2874.1
"NEUTRAL"	103.7	519.3	1743.4	2595.7

2.3. Speech motor adaptation task

Participants in both groups produced a total of 260 target words drawn from the stimulus list of 9 possible [ɛ] words presented in a randomized order. All participants underwent the following sequence of auditory feedback and carrier-phrase conditions (see Figure 1B): (1) an initial set of 30 baseline trials under normal auditory feedback and preceded by the *Neutral* carrier-phrase (NAF-Neutral); (2) a set of 100 trials under conditions of AAF preceded by the *High* or *Low* carrier-phrase, depending on the group (High-AAF or Low-AAF); (3) a set of 100 words under AAF preceded by the Neutral carrier-phrase (Neutral-AAF), and finally (4) a washout phase of 30 trials under NAF with the Neutral carrier-phrase (Neutral-AAF). Speech was recorded in a quiet testing room using a head-mounted microphone (C520, AKG, Germany) and digitized at 16-bit / 44.1 kHz on a PC using custom software written in Matlab (Mathworks, MA). Auditory speech signals were presented to participants using circumaural headphones (880 pro, Beyerdynamic, Germany).

2.4. Real-time alteration of speech

As in prior studies (e.g., Bourguignon *et al.*, 2014; Mollaei *et al.*, 2013), the auditory feedback manipulation was introduced at maximum level following the baseline period and corresponded to a 30% decrease in F1 (average shift: 180.2 Hz), inducing the perception of a vowel closer to [ɪ]. A detailed description of the system has been published previously (Bourguignon *et al.*, 2014). Briefly, the microphone signal was amplified and split into two channels, one providing an

unprocessed signal and the other altered using a digital signal processor (VoiceOne, TC Helicon) to decrease the frequency of all vowel formants. The vowel formant alteration was restricted to F1 by splitting both signals into non-overlapping low- and high- frequency components (Wavetek 753 low/high pass filter), and then mixing the low-frequency portion of the processed signal with the high-frequency portion of the unprocessed signal. The two signals were filtered using a 1100Hz cutoff, which lies roughly half-way between the first and second formant values for the production of the vowel [ε] in adult males (based upon pilot studies). The total signal processing delay was less than 15 ms.

Participants were asked to keep a constant speaking volume throughout the task. They were aided in this by a digital VU meter presented on the computer screen (showing current and peak acoustic signal level during each trial). Participants were instructed to maintain a target level on the display, which was adjusted at the beginning of the experiment to correspond to a comfortable speaking volume.

2.5. Acoustic analysis

For each word produced in the speech adaptation task, a 30 ms segment centered around the midpoint of the vowel was selected. Mean F1 and F2 frequency for each segment was then estimated by LPC analysis in Matlab. LPC parameters were chosen on a per-participant basis to minimize the occurrence of spurious formant values. F0 was also estimated for each vowel center using an autocorrelation method in *Praat* (Boersma, 1993). Values of F0, F1 and F2 frequency were used to directly compare vowel acoustic properties between conditions (NAF, High/Low - AAF and Neutral-AAF), and between groups (Low carrier-phrase and High carrier-phrase). Acoustic changes in the vowel during the speech adaptation task were computed by means of a "F1 compensation index," reflecting the proportion change in frequency relative to the mean values during the baseline NAF phase (averaged over trials 11-30). Increases in F1 thus correspond to values greater than 1. Differences in speech adaptation between the two groups were evaluated at two key time points: (1) at the end of the first production phase under conditions of altered feedback (High-AAF or Low-AAF; averaged over trials 111-130), and (2) at the end of the second phase under altered feedback (Neutral-AAF; averaged over trials 211-230). The reason for analyzing the final twenty trials in each phase was to avoid trials in which

participants' performance may not have stabilized in the Baseline phase on the one hand, and to measure the maximal effect at the end of the training period for the AAF phases on the other.

2.6. Test of perceptual vowel normalization

The carrier-phrases used in the test of speech motor adaptation were close replications of those used in a prior study of speech perception demonstrating the perceptual normalization of vowels with naturally produced stimuli (Ladefoged, 1989). However, they were not the exact stimuli used in that study. In order to verify the perceptual effect of these sentences in the context of the current experimental setup, a control study was carried out involving 14 native English-speaking listeners (7 male and 7 female, aged 18-30), without history of speech, hearing or language disorders none of whom participated in the test of speech motor adaptation. All listeners passed a pure-tone hearing screening at the time of testing.

Participants listened to a series of target words over headphones (880 pro, Beyerdynamic, Germany), each one preceded by the presentation of the High or Low version of the carrier-phrase (see above for a description of the phrases). Two blocks of 70 trials were carried out, one involving the presentation of the High carrier-phrase, and the other involving the Low carrier-phrase. The block order was counterbalanced among participants. The target words were drawn from a 7-step acoustic continuum between [bɛt] (*bet*) and [bɪt] (*bit*), created by applying step-wise (20 Hz) decreases in F1 frequency to a recording of the word "bet" (produced by the same speaker as the carrier-phrases). The decrease in F1 was carried out using the same experimental procedure as the real-time alteration of formant frequency in the test of speech motor adaptation (described above). Following each presentation of the carrier-phrase and target word, participants had to identify the target vowel as either [ɛ] (*bet*) or [ɪ] (*bit*) by pressing the appropriate key on a keyboard. Key order was counterbalanced between participants. For each of the two carrier-phrase conditions, the 7 target stimuli were presented 10 times each in a fully randomized order (totaling 140 trials).

Each participant's data were analyzed by computing, for each carrier-phrase condition, the proportion of [ɛ] responses for each of the 7 stimuli. The effect of the carrier-phrase was then evaluated at three key points along the continuum: the [ɪ]-endpoint (mean of stimuli 1 and 2), the middle (mean of stimuli 3-5) and the [ɛ]-endpoint (mean of stimuli 6 and 7), using a two-way

repeated-measures ANOVA.

3. Results

3.1. Test of perceptual vowel normalization

From prior studies of perceptual normalization (Ladefoged & Broadbent 1957; Ladefoged, 1989) it was predicted that the identification of the target vowels would be systematically influenced by the preceding carrier-phrase. Specifically, vowels following the phrase containing *high* fundamental and formant frequencies would tend to be identified as containing a relatively low F1, corresponding to the vowel [ɪ]. Conversely, vowels following the *low* carrier-phrase would tend to be perceived as relatively high in F1, corresponding to the vowel [ɛ]. The results of the present perception test match this prediction (Figure 2). The proportion of [ɛ] responses is consistently higher for all vowel stimuli in the context of the Low carrier-phrase (red line), compared to the High carrier-phrase (blue line). A two-way repeated-measures ANOVA showed a reliable main effect of STIMULUS (F[2,26]=32.30, $p < 0.01$) and a reliable main effect of CARRIER-PHRASE (F[1,13]=13.18, $p < 0.01$). While the mean proportion of [ɛ] responses was greater for all stimuli under the Low carrier-phrase condition (i.e., the two identification curves never cross), a significant interaction effect between factors was observed (F[2,26]=13.07, $p < 0.01$), corresponding to a greater effect of carrier-phrase on the classification of the first and second stimuli (i.e., the [i] end of the continuum, cf. Figure 2). This larger effect may be due to the fact that the [i] stimuli were modified versions of the naturally produced [ɛ] endpoint stimulus, rendering them somewhat more ambiguous. However, post-hoc pairwise comparisons carried out between carrier-phrase conditions using the Holm-Bonferroni method at each region of the continuum ([ɪ]-endpoint, middle and [ɛ]-endpoint) revealed that all differences were statistically reliable ($p < 0.05$).

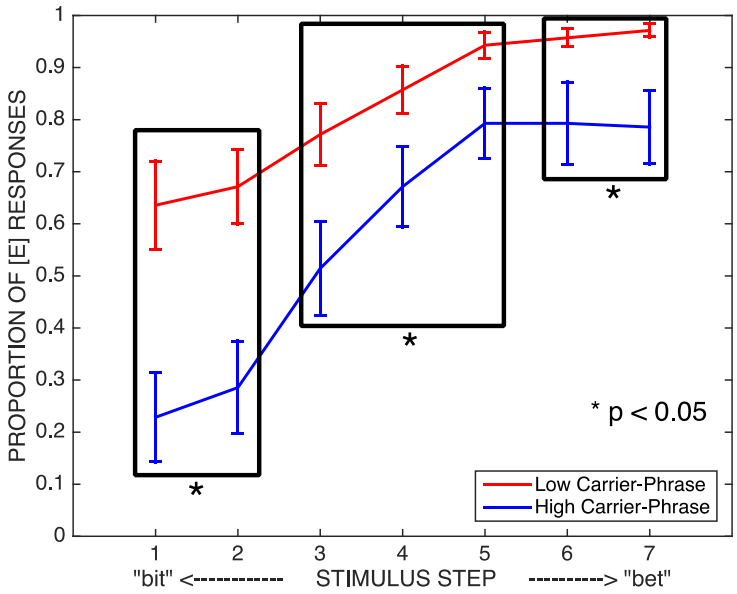

Figure 2 – Mean response pattern for the test of perceptual vowel normalization. The perception of different vowels along an [ɛ-ɪ] continuum is influenced by the spectral properties of a preceding carrier-phrase. The three comparisons of interest ([ɪ]-endpoint, middle, and [ɛ]-endpoint) are indicated by black boxes. As can be seen, the proportion of [ɛ]-responses provided after exposure to the High carrier-phrase (blue line) is significantly lower than the proportion of [ɛ]-responses provided after exposure to the Low carrier-phrase (red line) for all tokens along the [ɛ-ɪ] continuum. This result replicates Ladefoged & Broadbent's original vowel-extrinsic normalization effect, whereby vowels processed after exposure to carrier-phrases with relatively high spectral properties will be identified as comparably lower in their formant values (in this instance, less [ɛ]-like) and vice versa.

3.2. Production Baseline

In order to ensure that the two groups were comparable in their production of [ɛ] during the Neutral-NAF baseline phase, mean F0, F1 and F2 values were compared between groups using independent-samples *t*-tests. No reliable baseline differences were observed between the groups for F0 ($t[18] = 1.63$, $p = 0.12$), F1 ($t[18] = 0.50$, $p = 0.62$), or F2 ($t[18] = 0.80$, $p = 0.43$). The mean values of F0, F1 and F2 were, respectively, 116, 611 and 1681 Hz for the Low carrier-phrase group and 104, 595 and 1649 Hz for the High carrier-phrase group.

3.3. Speech Adaptation

The results of the speech adaptation task for the two groups (*High* and *Low carrier-phrase*) are shown in Figure 3 with mean changes in formant values relative to baseline at the end of the two AAF phases shown in Figure 4. Overall, a compensatory increase in F1 frequency can be observed in response to the F1 auditory feedback manipulation for both groups. By the end of the first AAF phase, during which the two carrier-phrases differed for the two groups, the magnitude of the F1 compensation can be seen to diverge between the two phrase conditions, with the *High carrier-phrase* group showing a relatively large F1 increase, and the *Low carrier-phrase* showing a smaller increase. By the end of the second AAF phase, during which both groups were presented with the *Neutral* carrier-phrase, the magnitude of the compensatory change can be seen to converge once again.

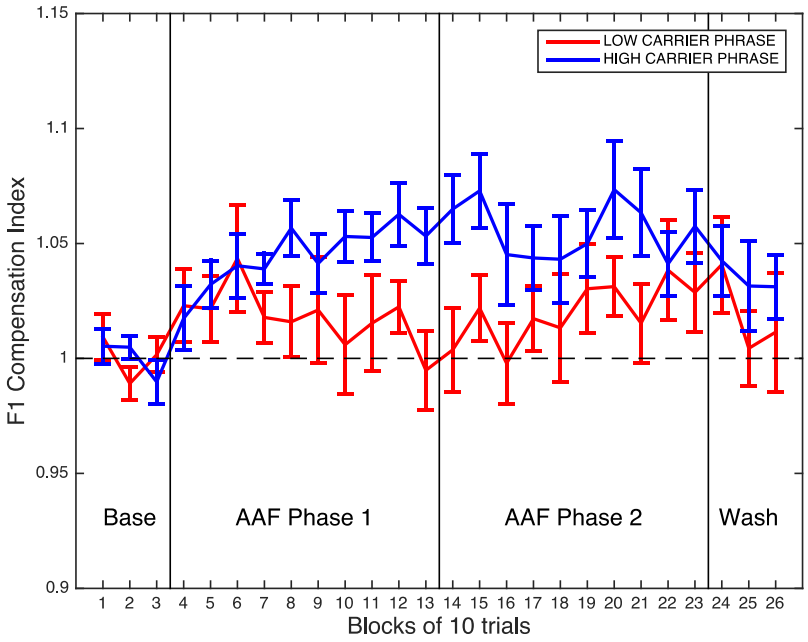

Figure 3 – Compensatory changes in speech motor output in response to the real-time manipulation of F1 auditory feedback. The mean F1 compensation index (reflecting the proportion change relative to baseline) is seen to diverge during the first altered auditory feedback phase (AAF Phase 1) between the group exposed to the *High* carrier-phrase (blue line) and the group exposed to the *Low* carrier-phrase (red line). The degree of compensation converges by the end of the second AAF phase, during which both groups were exposed to the same (*Neutral*) carrier-phrase. The difference between groups in the magnitude of the motor-compensatory response reflects the differential effect of the carrier phrases on subjects' perception of their own auditory feedback, either enhancing or diminishing the perceived auditory error during vowel production (see text for details).

The reliability of the carrier-phrase effects was evaluated using a 2-way mixed-factorial ANOVA, focusing on the magnitude of the compensatory response at the end of AAF-Phase 1 (different carrier-phrases between groups) and the end of AAF-Phase 2 (same carrier-phrase between groups). In the analysis, GROUP (*High* vs. *Low carrier-phrase*) is a between-participants factor and PHASE (AAF-Phase 1 and AAF-Phase 2) is a within-participants factor. Neither the main effect of GROUP nor PHASE was significant (GROUP: $F(1,18)= 2.46$, $p = 0.13$; PHASE: $F(1,18) = 0.89$, $p = 0.36$), however the 2-way interaction effect was reliable ($F(1,18) = 4.79$, $p < 0.05$). Post-hoc pairwise comparisons were carried out using the Holm-Bonferroni procedure. A reliable difference between groups was found at the end of AAF-Phase 1 ($p < 0.05$), but not at the end of AAF-Phase 2 ($p = 0.68$).

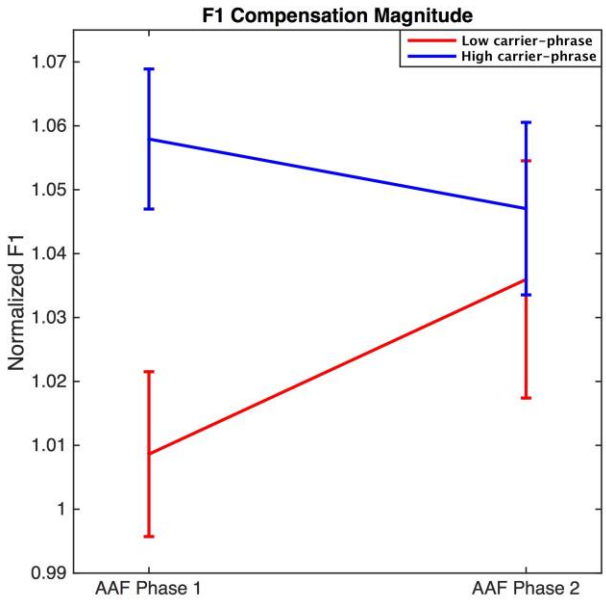

Figure 4 – The mean compensation effect at the end of AAF-Phase 1 and AAF-Phase 2. Consistent with predictions, the *High* carrier-phrase group exhibited greater compensation than the *Low* carrier-phrase group during AAF-Phase 1. The effect is seen to diminish by the end of AAF-Phase 2 (during which both groups were exposed to the *Neutral* carrier-phrase).

While a between-group effect was noted for the compensatory change in F1, no systematic difference was observed between the carrier-phrase conditions for fundamental frequency (F0) or

F2 (see Figure 5). A 2-way mixed-factorial ANOVA examining the change in these acoustic parameters at the end of AAF-Phase 1 and AAF-Phase 2 showed no significant main or interaction effects for either F0 (GROUP: $F(1,18)= 0.11$, $p = 0.75$; PHASE: $F(1,18) = 0.323$, $p = 0.23$; Interaction: $F(1,18) = 0.18$, $p = 0.68$) or F2 (GROUP: $F(1,18) = 0.28$, $p = 0.60$; PHASE: $F(1,18) = 0.90$, $p = 0.35$; Interaction: $F(1,18) = 0.04$, $p = 0.84$).

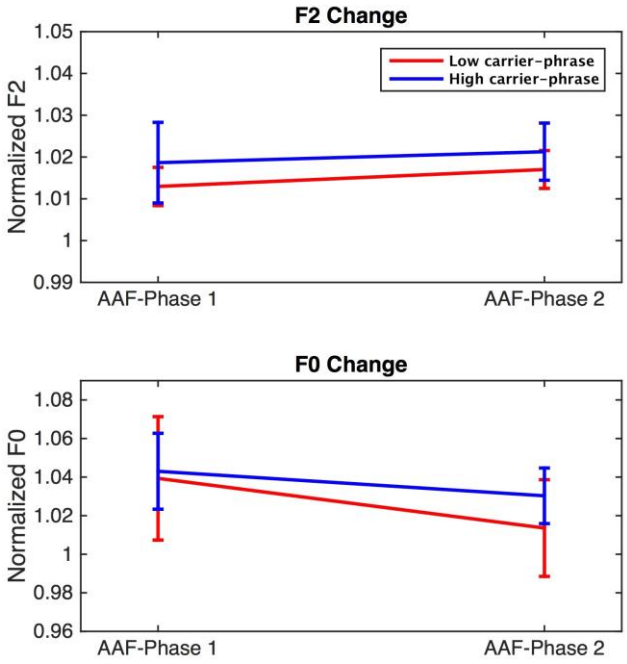

Figure 5 – Mean change in F2 and F0 (proportion relative to baseline) is shown for the High carrier-phrase (blue) and Low carrier-phrase (red) groups at the end of AAF-Phase 1 and AAF-Phase 2. In contrast with the observed changes in F1, no difference between groups was observed for either of these acoustic measures during the two phases.

Finally, in order to evaluate the reliability of the F1 compensation effects in each group associated with the feedback alteration, a separate analysis was carried out in which the F1 compensation index (proportion change relative to baseline) at the end of each AAF phase was directly compared with the baseline value of 1. At the end of AAF Phase 1 (trial block 13, see Figure 3), a reliable adaptation effect was observed for the High-sentence group ($t[9]=4.09$, $p < 0.01$) but not for the Low-sentence group ($t[9]=0.032$, $p = 0.97$). In contrast, at the end of the AAF Phase 2 (trial block 23), both groups exhibited compensation effects that differed reliably

from baseline (High-sentence group: $t[9]=3.04$, $p < 0.05$; Low-sentence group: $t[9]=2.27$, $p < 0.05$).

4. Discussion

The present study expands on recent demonstrations that short-term changes in phoneme perception can impact the sensorimotor control of speech production (Lametti *et al*., 2014; Shiller & Rochon, 2014). In contrast with these studies, which relied on reinforcement-based training to introduce perceptual changes, the present study examined a form of rapid, automatic perceptual plasticity that accompanies everyday interactions with other talkers: vowel-extrinsic normalization. Specifically, we examined whether the vowel formants characterizing an introductory carrier-phrase would serve as a frame of reference for talkers' perception of their own speech, thereby influencing their degree of motor adaptation to an alteration of auditory feedback during the production of [ɛ]-words. We predicted that a carrier-phrase with higher formants would yield a perception of the self-produced vowel as comparatively low in F1, thereby increasing the perceived auditory error and enhancing the degree of speech motor compensation. In contrast, a carrier-phrase with lower formants was predicted to yield perception of the self-produced vowel as comparatively high in F1, thereby decreasing the perceived error and diminishing the motor compensatory response. Consistent with these predictions, results showed a difference in the degree of F1 compensation between groups when participants were exposed to different carrier-phrases under AAF, and a subsequent convergence in compensation magnitude when both groups were exposed to the same (*Neutral*) carrier-phrase under AAF. This indicates not only that the auditory processing guiding speech production is highly flexible and adaptive under a range of perceptual conditions, but that mechanisms of auditory plasticity previously observed only in the context of extrinsic speech perception (namely, vowel-extrinsic normalization) may also play a role in the sensorimotor control of speech production.

On the surface, there are similarities between the sensorimotor interactions observed in the present study and those associated with *phonetic convergence* – the tendency for speech phonetic properties to converge between conversational partners, including vowel spectral properties, duration, and amplitude (Pardo 2006, 2013; Sato *et al*., 2013). Both phonetic convergence and the present results are characterized by an effect of extrinsic speech signals on

speech motor control. However, phonetic convergence differs from the present findings in a number of important ways. In particular, convergence effects involving vowel formants have been inconsistent among studies, with some showing convergence (e.g., Babel, 2012; Sato *et al*., 2013) and others showing divergence or mixed changes in vowel formant patterns across talkers under varying conditions (Pardo, 2013). Furthermore, while several studies have demonstrated convergence effects under conditions of non-interactive word presentation (e.g., Goldinger, 1998; Sato *et al*., 2013; Shockley *et al*., 2004), psycho-social variables such as attractiveness (Babel, 2012), social closeness (Pardo *et al*., 2012), conversational role and gender (Pardo, 2006) have been shown to modulate the degree of convergence across a range of acoustic parameters. In the present study, participants did not engage in a conversational interaction, but rather were passively exposed to a brief sample of another talker's speech prior to speaking (and notably, that exposure did not include the target word). The resulting effect on speech production was found to be restricted to vowel F1 frequency: the acoustic property that was manipulated systematically and hence the primary driver of the compensatory speech motor response. Participants did not exhibit differential changes in other vowel acoustic properties, such as F0 or F2, which also differed systematically between the High and Low carrier-phrases. The restriction of the effects in the current study to changes in F1 indicates that, rather than simply converging towards the carrier-phrases, participants interpreted their own vowel acoustic error (confined to F1) within a frame of reference provided by the carrier-phrase – a change that parallels the process of vowel-extrinsic normalization.

Although the restriction of the effect in the present study to changes in F1 rules out an explanation in terms of phonetic convergence toward the carrier phrase, our results illustrate a mechanism whereby extrinsic speech properties alter the perception of self-generated speech under somewhat more natural conditions than previously demonstrated (e.g., using intensive, reinforcement-based auditory-perceptual training, cf. Lametti *et al*., 2014; Shiller & Rochon, 2014). It is therefore conceivable that with longer, more varied exposure to extrinsic speech stimuli in naturalistic, social-conversational contexts, adaptive perceptual processes could drive long-term changes in speech output through a combination of vowel-extrinsic normalization and error-correction mechanisms in the speech motor system. Interestingly, such an explanation for speech accommodation (or convergence) has, to our knowledge, not previously been discussed. While some researchers have proposed a purely sensorimotor mechanism to explain phonetic

convergence, it has been framed as a shift in phonetic *targets* toward the model provided by an interlocutor (e.g., Sato *et al*., 2013). The mechanism suggested by the present result, in contrast, is purely perceptual in nature. Specifically, a talker hears his/her own formants as relatively high or low depending on the speech of the interlocutor (through the process of vowel-extrinsic normalization), thereby altering the perceived discrepancy between the perceived feedback and the intended target. Error-minimizing mechanisms would then "drive" speech properties toward those of the interlocutor in response to this perceived error in order to reduce the discrepancy.

Our results may have implications for existing models of speech production, in particular those that highlight a role for auditory-sensory feedback in the fine-tuning of predictive, feed-forward control processes (e.g., the Directions-into-Velocities of Articulators, or DIVA, model, Tourville & Guenther, 2011; or the State Feedback Control, or SFC, model; Houde & Nagarajan, 2011; Houde & Chang, 2015; see also Tian & Poeppel 2010 and Hickok 2012). Within the framework of such models, the context-dependent changes in speech motor adaptation observed in the current study could emerge either early on via contextual influences on the representation of auditory input, or at the earlier-specified prediction of this auditory input, reflecting the system's expectation of the auditory-sensory outcome, given the current speech goal. Since these different levels of representation are believed to occupy different regions of auditory and motor cortex (Guenther *et al.*, 2006; Tourville & Guenther, 2011; Houde & Nagarajan, 2011; Houde & Chang, 2015), these hypotheses yield distinct predictions about the locus of neural activity correlating with the behavioral effect reported here.

In the present study, both groups exhibited sensorimotor adaptation to alterations of auditory feedback, consistent with the idea that the acoustic correlates of phoneme categories operate as the primary targets of speech production. However, the results also indicate that the sensory processes guiding speech production can be rapidly biased or altered. This study joins a small but growing body of work indicating that neural processes once viewed as purely perceptual or cognitive in nature transfer to the sensorimotor processes guiding speech production. For example, recent research has shown that contrasts in the degree of compensation to feedback perturbations similar to those featured here also arise when combined with manipulations affecting *top-down* mechanisms related to the lexical status of the words being produced (Bourguignon *et al.*, 2014). These and the present findings invite the development of

interactive models accounting for the influence of perceptual plasticity (and the factors contributing to it) on the neural control of speech production.

5. References

Ainsworth, W. A. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. Tatham (Eds.). *Auditory analysis and perception of speech*. London, Academic Press, 103-113.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, 40, 177-189.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, 17(1193), 97-110.

Bourguignon N.J., Baum S.R., & Shiller D.M. (2014). Lexical-perceptual integration influences sensorimotor adaptation in speech. *Frontiers in Human Neuroscience*, 8, Article 208, DOI:10.3389/fnhum.2014.00208.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, 61(5), 977-985.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, 101(4), 2299-2310.

Byrd, D. (2000). Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica*, 57, 3-16

Christoffels, I.K., Formisano, E., Shiller, N.O. (2007). Neural correlates of verbal feeedback processing: An fMRI study employing overt speech. *Human Brain Mapping*, 28(9), 868-879.

Cooper, W.E., & Lauritsen, M.R. (1974). Feature processing in the perception and production of speech. *Nature*, 252, 121-123.

Dahan, D., Drucker S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition*, 108, 710-718.

Daniloff, R., & Moll K. (1968). Coarticulation of lip rounding. *Journal of Speech and Hearing Research*, 11,707–721.

Dechovitz, D. (1977). Information Conveyed by Vowels: A Confirmation. *Proceedings of the 93rd meeting of the Acoustical Society of America*. Pennsylvania State University, PA, 213-219.

Flege, J. E. (1991). Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language. *Journal of the Acoustical Society of America*, 89(1), 395-411.

Gentilucci, M., & Cattaneo, L. (2005). Automatic audiovisual integration in speech perception. *Experimental Brain Research*, 167, 66-75.

Goldinger, S. D. (1998). Echoes of Echoes? An Episodic Theory of Lexical Access. *Psychological Review*, 105(2), 251-279.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, 39, 350-365.

Hashimoto, Y., & Sakai, K.L. (2003). Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: an fMRI study. *Human Brain Mapping*, 20, 22-28.

Harrington, J., Palethorpe, S., & Watson, C.I. (2000). Does the Queen speak the Queen's English? *Nature*, 408, 927.

Heinks-Maldonado, T.H., Nagarajan, S.S., & Houde, J.F. (2006). Magnetoencphalographic evidence for a precise forward model in speech production. *NeuroReport*, 17(13), 1375-1379.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, 13, 135-145.

Hickok, G., Houde, J.F., & Feng, R. (2011). Sensorimotor Integration in Speech processing: Computational Basis and Neural Organization. *Neuron*, 69, 407-422.

Houde, J. & Jordan, M. I. (1998). Sensorimotor Adaptation in Speech Production. *Science*, 279, 1213-1216.

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience*, 5, Article 82, DOI: 10.3389/fnhum.2011.00082.

Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the Auditory Cortex during Speech: An MEG Study. *Journal of Cognitive Neuroscience*, 14(8), 1125-1138.

Houde, J.F. & Chang, E.F. (2015). The cortical computations underlying feedback control in vocal production. *Current Opinion in Neurobiology*, 33, 174-181.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America*, 88(2), 642-654.

Johnson, K. (2008). Speaker normalization in speech perception. In. D. B. Pisoni & R. Remez (Eds.). *The Handbook of Speech Perception*. Blackwell Publishing Ltd. Oxford, UK, 363-389.

Kerzel, D., & Bekkering, H. (2000). Motor Activation From Visible Speech: Evidence From Stimulus Response Compatibility. *Journal of Experimental Psychology: Human Perception and Performance*, 26(2), 634-647.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, 107, 54-81.

Ladefoged, P. (1989). A note on "Information conveyed by vowels". *Journal of the Acoustical Society of America*, 85(5), 2223-2224.

Ladefoged, P., (2001). *Vowels and consonants: An introduction to the sounds of languages.* Oxford, Blackwells.

Ladefoged, P., & Broadbent, D. (1957). Information Conveyed by Vowels. *The Journal of the Acoustical Society of America*, 29(1), 98-104.

Lametti, D. R., Krol, S. A., Shiller, D. M., & Ostry, D. J. (2014). Brief Periods of Auditory Perceptual Training Can Determine the Sensory Targets of Speech Motor Learning. *Psychological Science*, 25(7), 1325-1336.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical Society of America*, 85(5), 2114–2134.

Miller, J., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, 41, 215-225.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, 85(5), 2088-2113.

Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2013). What does motor efference copy represent? Evidence from speech production. *The Journal of Neuroscience*, 33(41), 16110-16116.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5(1), 42-46.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, 119(4), 2382-2393.

Pardo, J.S. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology*, 4, Article 559, DOI: 10.3389/fpsyg.2013.00559.

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, 40, 190–197.

Peterson, G. E., & Barney, H. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, 24(2), 175-184.

Purcell, D. W., & Munhall, K. G. (2006a). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *Journal of the Acoustical Society of America*, 120(2), 966-977.

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

Rochet-Capellan, A., & Ostry, D.J. (2011). Simultaneous Acquisition of Multiple Auditory-Motor Transformations in Speech. *The Journal of Neuroscience*, 31(7), 2657-2662.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception & Psychophysics*, 71, 1207-1218.

Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J.-L., & Nguyen, N. (2013). Converging toward a common speech code: imitative and perceptuo-motor recalibration processes in speech production. *Frontiers in Psychology*, 4, Article 422, DOI: 10.3389/fpsyg.2013.00422.

Shiller, D. M., & Rochon, M.-L. (2014). Auditory-Perceptual Learning Improves Speech Motor Adaptation in Children. *Journal of Experimental Psychology*, 40(4), 1308-1315.

Shockley, K., Sabadini, L., & Fowler, C.A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, 66(3), 422-429.

Sjerps, M. J., McQueen, J. M., & Mitterer, H. (2013). Evidence for precategorical extrinsic vowel normalization. *Attention, Perception and psychophysics*, 75, 576-587.

Tian, X. & Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology*, 1, Art 106. DOI: 10.3389/fpsyg.2010.00166.

Tourville, J. A., Reilly, K. J., & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *NeuroImage*, 39(3), 1429-1443.

Tourville, J.A., & Guenther, F.H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Langauge and Cognitive Processes*, 26(7), 952-981.

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *Journal of the Acoustical Society of America*, 122(4), 2306-2319.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Journal of the Acoustical Society of America*, 113(2), 1033-1043.