Temporal control and compensation for perturbed voicing feedback

Takashi Mitsuya[a)]
School of Communication Sciences and Disorders and National Centre for Audiology, University of Western Ontario, Elborn College, Room 1207, London, Ontario, N6G 1H1, Canada

Ewen N. MacDonald
Department of Electrical Engineering, Technical University of Denmark, Ørsteds Plads, Building 352, Room 116, 2800 Kongens Lyngby, Denmark

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

(Received 3 September 2013; revised 28 March 2014; accepted 3 April 2014)

Previous research employing a real-time auditory perturbation paradigm has shown that talkers monitor their own speech attributes such as fundamental frequency, vowel intensity, vowel formants, and fricative noise as part of speech motor control. In the case of vowel formants or fricative noise, what was manipulated is spectral information about the filter function of the vocal tract. However, segments can be contrasted by parameters other than spectral configuration. It is possible that the feedback system monitors phonation timing in the way it does spectral information. This study examined whether talkers exhibit a compensatory behavior when manipulating information about voicing. When talkers received feedback of the cognate of the intended voicing category (saying "tipper" while hearing "dipper" or vice versa), they changed the voice onset time and in some cases the following vowel. © *2014 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4871359]

PACS number(s): 43.70.Mn, 43.70.Bk [ZZ] Pages: 2986 2994

I. INTRODUCTION

The timing of speech is coordinated over many temporal scales ranging from milliseconds in individual phones and consonant clusters (Kent and Moll, 1975) to seconds in the rhythmical structure for phrases and sentences (Lehiste, 1970). Historically, this precise timing is explained by one of three accounts. (1) Speech timing is attributed to some kind of clock mechanism within the nervous system (e.g., Ivry and Richardson, 2002; Buonomano and Laje, 2010). (2) Speech timing is controlled by detailed motor scripts or intrinsic motor organizations (e.g., MacNeilage, 1970; Tuller and Kelso, 1984; Fowler, 1986). (3) Speech timing of future movements is regulated by feedback from ongoing movements (e.g., Washburn, 1916). In their strictest sense, none of these frameworks has been entirely adequate to explain the intricacies of the temporal structure of speech. However, aspects of all of these ideas appear in recent accounts of the temporal control of behavior (e.g., Buhusi and Meck, 2005).

Following the theoretical criticism of Lashley (1951), sensory feedback has long been dismissed as the source of accurate timing spanning short temporal scales such as occurs in individual consonant coordination (cf. temporal coordination of pitch control with auditory feedback has been studied extensively; e.g., Hain *et al.*, 2001; Liu *et al.*, 2009; Behroozmand *et al.*, 2009; and Cai *et al.*, 2011). However, recent developments in models of feedforward processing in speech movement (e.g., Guenther *et al.*, 1998; Houde and Nagarajan, 2011; Kröger *et al.*, 2009) have led us to revisit the importance of sensory feedback in the control of articulatory coordination for segments. Evidence suggests that talkers monitor their own voice and the perception of self-produced sounds influences many acoustic parameters of speech. For example, it has been shown that talkers exhibit compensatory behavior not only for altered suprasegmental parameters such as loudness (Bauer *et al.*, 2006) and pitch (Burnett *et al.*, 1998; Jones and Munhall, 2000), but also for segmental parameters such as vowel formant frequencies (Houde and Jordan, 1998; Purcell and Munhall, 2006; Villacorta *et al.*, 2007; MacDonald *et al.*, 2010, 2011; Cai *et al.*, 2010; Mitsuya *et al.*, 2011, 2013) and fricative noise (Shiller *et al.*, 2009; Casserly, 2011).

To date, the only segmental parameters that have been examined are acoustic frequency manipulations (formant and spectral noise). However, there are other acoustic parameters that define speech segments such as voicing. All languages have voicing contrasts. For plosives or stop consonants, if vibrations of the vocal folds are present when the consonantal gestures are made and the air is released through the constricted point of the vocal tract, it is perceived to be voiced, whereas if such vibration occurs with a temporal lag, it is perceived to be voiceless. The relative timing between the plosive release and the initiation of voicing is called voice onset time (VOT). Detailed articulatory parameters for this measure are not language universal (see Lisker and Abramson, 1964,

a)Author to whom correspondence should be addressed. Electronic mail: tmitsuya@uwo.ca

 0001-4966/2014/135(5)/2986/9/$30.00

for a review). For example, in Romance languages such as Spanish and French, voicing generally precedes the plosive burst for the voiced consonants, while for voiceless consonants, voicing follows shortly after the plosive burst. In English, voicing follows shortly after the plosive release for the voiced plosives (thus, the VOT value that is typically positive), which is comparable to the voiceless category in Romance languages, whereas for English voiceless consonants, voicing starts much later than the plosive release and it is usually accompanied with aspiration. Even within a language, a voicing category may consist of multiple allophones (e.g., in English, when voiceless plosives occur in a syllable medial position, they are usually unaspirated). In short, although the voicing category is not strictly defined by phonatory timing relative to other gestures in a language universal manner, timing still remains the main acoustic cue to distinguish the categories.

Timing parameters and their control for voicing have been under extensive investigation for a long time. VOT as one of the strongest cues to distinguish the voicing category has been decomposed into its acoustic (e.g., Lisker, 1986) and articulatory constituents and their interactions and coordination (e.g., Löfqvist and Yoshioka, 1984). In terms of articulatory postures, VOT is determined by the interarticulator timing of laryngeal and oral gestures plus the conditions of air pressure across the vocal folds. Löfqvist and Yoshioka (1981, 1984) reported that even though the magnitude of glottal opening during consonant production was variable, the temporal relationship between supralaryngeal and laryngeal gestures was precisely coordinated. This coordination of laryngeal and oral gestures of voicing was demonstrated to be remarkably flexible by Munhall *et al*. (1994). Using a motor perturbation paradigm, Munhall *et al*. (1994) examined the coordination of laryngeal and supralaryngeal gestures. They reported that when the lower lip was mechanically perturbed, interfering with the bilabial closure for /p/ in an /i'pip/ context, not only did the lips show compensatory movements to achieve the closure, but also the onset of glottal abduction was delayed, prolonging the vowel. The following bilabial closure was shorter than the unperturbed one, causing a slightly longer VOT. However, the glottal abduction movement did not change its timing or duration relative to oral movements. Taken together, it is clear that the control of voicing as a phonemic outcome is not solely dictated by voicing as an independently prescribed movement; rather, the interaction of a host of many gestures around the plosive consonant fluidly coordinate their timing to control the category of voicing.

The fact that mechanical perturbations elicit compensatory coordination of multi-articulator movements for voicing suggests that such articulatory maneuvers are influenced by sensory information, thus other form of sensory feedback such as auditory feedback may also be crucial for VOT control. Lane and Perkell (2005) reviewed the studies that examined acoustic and articulatory differences of VOT production due to the lack of auditory feedback among deaf individuals. The VOT difference between voiced and voiceless consonants was reduced (Lane *et al*., 1994; Waldstein, 1990) and often, the voiced/voiceless distinction was wrongly produced (Cowie and Douglas-Cowie, 1983) after individuals lost their hearing. Moreover, in a study with prelingually deafened children, Higgins *et al*. (2001) turned on and off the cochlear implants to examine the effect of auditory feedback (or the lack of) on their speech acoustics. Although the finding was not consistent across their two children examined, at least one child's VOT production for /p/ was significantly reduced the day the implant was turned off. Another study by Jones *et al*. (2003), compared the intraoral pressure as well as its duration during /p/ and /b/ production in the /CVCV/ context (puppy vs baby) among children with normal hearing versus children with a cochlear implant. With children with a cochlear implant, these parameters were measured when the implant was on and off. The changes due to the removal of auditory feedback were not consistent across the examined children however, it seemed to introduce articulatory adjustments, implying that auditory information is used for articulatory distinction between the voicing categories, even though each of the children made use of the information for different articulatory maneuvers. The results of those studies strongly suggest that (1) without being able to hear their own speech, the voicing contrast cannot be monitored precisely and the phonetic details for the contrast are no longer maintained in articulation and (2) auditory information is used to control/adjust voicing contrast.

If the function of the speech production error correction system is to monitor and adjust *any* speech sound that we produce, then the system has to monitor and adjust articulatory/phonetic details of voicing as well. Thus, people should exhibit compensatory behavior when they receive perturbed voicing feedback. In the present study, we examined whether talkers exhibited compensatory production when they received feedback that was not congruent with the voicing production that they intended, using the same general paradigm of the real-time perturbed auditory feedback experiments. Although VOT is a strong acoustic signature for the perception of voicing, as Lisker (1986) discussed, there are a host of many other acoustic parameters associated with voicing. Thus, we decided not only to manipulate VOT; rather, we used large perturbations cross-categorical feedback perturbations in which talkers heard the voicing cognate opposite to the one they were producing. The use of a cross-categorical feedback perturbation allowed us to examine the temporal organization of syllable constituents. Although the paradigm itself is comparable to that of F0 (Burnett *et al*., 1998; Jones and Munhall, 2000) and vowel formant perturbation (Houde and Jordan, 1998; Purcell and Munhall, 2006; Villacorta *et al*., 2007; MacDonald *et al*., 2010, 2011; Mitsuya *et al*., 2011, 2013), real-time modifications of temporal parameters are not easily implemented. That is, it is difficult to lengthen ongoing temporal parameters, and, it is impossible to shorten them. In order to solve this problem, we pre-recorded talkers' utterances, and selected five productions of each token that had VOTs closest to the mean of each voicing cognate to be used as playback tokens during perturbation. This is similar to the manipulation used in the study by Mochida *et al*. (2010). But unlike their study where articulatory measurements were measured, we investigated

how acoustic details were changed due to the introduction of perturbed auditory feedback of voicing category.

II. METHODS

A. Participants

Thirteen female undergraduate students [mean age = 19.4; standard deviation (s.d.) = 1.0] at Queen's University in Canada participated. All were native Canadian English talkers with normal hearing thresholds within the range of 500 4000 Hz (<20 dB hearing level). None reported a history of language or speech impairments.

B. Equipment

Participants were tested in a sound isolated booth (Industrial Acoustic Company). They wore a headset microphone (Shure WH20) and the microphone signal was amplified (Tucker-Davis Technologies MA3 microphone amplifier), low-pass filtered with a cutoff frequency of 4500 Hz (Krohn-Hite 3384 filter), digitized at 10 kHz and the detection of voicing was processed in real-time (National Instruments PXI-8106 embedded controller) such that a playback token was gated when the talker's utterance onset was detected. This resulted in a delay of 3 5 ms, but this magnitude of delay is less than observed in all auditory perturbation studies. While talkers were receiving the playback token, their actual microphone signal was not fed through the headphones. Feedback (regardless of whether it was normal or playback token) was amplified and mixed with noise (Madsen Midimate 622 audiometer), and presented over headphones (Sennheiser HD 265) such that the feedback signal and noise were presented at approximately 80 and 50 dBA sound pressure level, respectively.

C. Target words

Because we used natural productions of syllables, temporal parameters of the playback tokens differed more than simply by onset VOT. If the syllabic duration of a voicing cognate pair is similar, the vowel with a voiceless plosive has a longer vowel duration. Perturbation studies such as Munhall *et al.* (1994) and studies of compensatory shortening (Fowler, 1981; Munhall *et al.*, 1992) have reported that the duration of speech segments is controlled not for each segment but for a proportional relationship of the segments that comprise a larger unit, such as syllable. The organization of constituents within a syllable as a larger unit may also be examined using the real-time auditory feedback paradigm. In order to measure duration of each constituent in a syllable, it was necessary to have a consonantal coda to have a concrete and consistent offset of the voicing. Monosyllabic CVC codas may be deemphasized, in which case the duration of the vowel may also become more variable. For this reason, we selected the trochaic disyllabic words, "tipper" and "dipper," as our test words for the experiment.

In the perturbation phase of the experiment, while talkers were repeating the word "tipper" they started to hear their pre-recorded production of "dipper." In the contrasting condition, while saying "dipper," they received "tipper" as the perturbed feedback. We measured a change in VOT, the following vowel, coda closure and the syllable duration due to the introduction of the playback token perturbations.

D. General procedure

Each participant was tested twice over the course of two weeks. The first session, session 1, was to collect their playback tokens, and session 2 was the experimental session.

1. Session 1

a. Baseline. Each participant was tested individually. They sat in front of a monitor on which a target word ("tipper" or "dipper") was presented for 1 s with an interstimulus interval of approximately 1.5 s. Participants were instructed to produce the prompted word trochaically at a normal pace without gliding their pitch. A total of 100 utterances were produced for each of the words. After completing the 100 tokens of one word, they repeated the same procedure with the other target word. The order of the word presentation was counterbalanced across participants. Participants wore headphones, through which they heard the produced sounds without any perturbation. Within a week's time, participants returned to participate in session 2.

b. Playback tokens. For each utterance, automated estimates of burst and voicing onsets were generated using the algorithm of Das and Hansen (2004). These estimates were reviewed and, if necessary, adjusted by hand. From session 1, each participant's average VOT of each voicing category was measured. Outliers (tokens with VOT values greater or smaller than mean +3.0 s.d. or smaller than −3.0 s.d.) were removed from assessing the mean VOT. Five tokens with VOT values closest to the mean of the voicing category were chosen for each individual as playback tokens for the second session. The amplitude of these selected tokens was normalized, thus there was no difference in peak amplitude across the playback tokens.

2. Session 2: Experimental condition

When participants came back for the second session, they performed the same task as the first session, producing 100 utterances of each of the words "tipper" and "dipper." The order of the word types was the same from session 1. During the second session, however, the 100 utterances of each word were divided into three experimental phases: the baseline phase (utterances 1 20), the perturbation phase (utterances 21 60), and the return phase (utterances 61 100). During the baseline phase, participants were given normal feedback through the headphones so that they heard their voice as they produced the target word. This matched the conditions for session 1. During the perturbation phase, however, they heard their own pre-recorded playback tokens of the other voicing category as they produced the target word on the monitor. For example, while they were producing the word "tipper," they heard their voice saying "dipper" and vice versa. The order of presentation of the five playback tokens was not randomized. Thus, there were sequential

presentations from token 1 through 5, which repeated during this phase. These playback tokens were played at 80 dB. In the return phase, the participants received normal feedback just as in the baseline phase.

III. RESULTS

A. Session 1

With the data collected from the first session, the utterances 11 100 were used for the analysis to calculate the durational parameters of the first syllable of the two target words. For the target word "dipper," many talkers produced a few instances of negative VOT with which the onset of phonation precedes the plosive burst and continues through the plosive closure and the burst. The mean occurrence of negative VOT production was 3.0 times over a course of 90 utterances, and there were no systematic patterns for these negative VOT productions. These tokens were excluded from the analyses because (1) the values of negative VOT were, in general, quite large, and they would have skewed the mean value of positive VOT for /d/ and (2) the idiosyncratic/unsystematic productions of negative VOT are not representative of talker's standard (positive) VOT for /d/ in the word onset context.

We examined various durational parameters for the test tokens, including VOT of the onset consonant, vowel duration of the vowel in the first syllable, coda closure (offset of the vowel to the onset of /p/ burst of the second syllable) as well as syllable duration of the first syllable (from the onset of initial alveolar plosive to the onset of the plosive of the second syllable; see Fig. 1).

On average, all of the durational parameters were different across the two target words (the order of the presentation was not significant, $p > 0.05$) as summarized in Table I. Although the syllabic duration for "tipper" was significantly longer than that of "dipper," this effect was due to the longer VOT since all other parameters of "tipper" were significantly shorter than those of "dipper." This temporal structure of syllable constituents across voicing cognates has previously been reported by Allen and Miller (1999). Although the increase in VOT from voiced to voiceless consonant is associated with a shortened vowel duration, this reduction is less than the increased amount of VOT resulting in a slightly longer syllable duration for voiceless tokens.

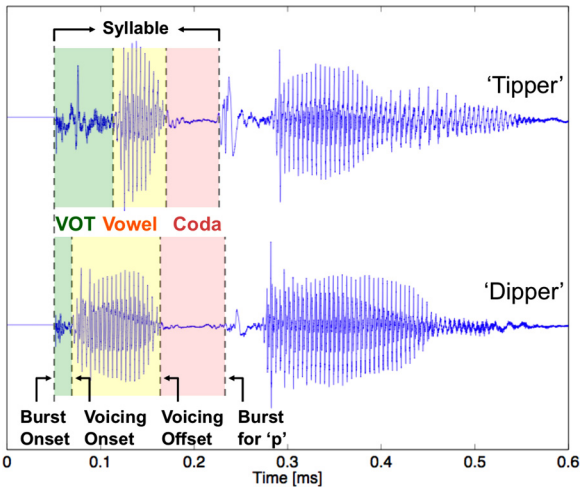

FIG. 1. (Color online) Illustration of VOT, vowel, and coda closure for the first syllable of the utterances of "tipper" (top) and "dipper" (bottom).

The durational parameters of the target words changed slightly over the course of 90 productions of the tokens. A series of regression analyses revealed that some of the parameters changed significantly. As seen in Table I, however, the unstandardized coefficients were all rather small. For the token "dipper," vowel and syllable were lengthened by 4.32 and 6.48 ms, respectively, over the course of 90 trials. For the token "tipper," VOT was shortened by 2.7 ms, while coda closure and syllable duration were lengthened by 6.75 and 3.87 ms, respectively. These changes (i.e., shortened VOT with lengthened coda closure and syllable) are inconsistent with the changes associated with speaking rate. That is, if talkers were producing the segment slower, indicated by the lengthened syllable, other durational parameters should also be lengthened (see Kessinger and Blumstein, 1998, for a review). Although there is no clear explanation to account for these temporal changes, this might have been due to (1) fatigue from repetitive productions and (2) some unknown effect of repetitive production itself. Repetitive listening to a syllable has been reported to lead to intermittent changes in the perception of the constituent phoneme (verbal transform; Warren, 1961; Tuller *et al.*, 1997). Possibly, repetitive production might have shifted perception of the sound the talkers were producing which in turn might have affected the production of the sound.

B. Session 2

1. Comparison across sessions

As in session 1, the first 10 tokens of the 100 trials were removed from the analyses. Most of our talkers produced some instances of negative VOT for /d/ during the experiment. The mean frequency of negative VOT production during this session was 7.5 times over the course of 90 trials and there was no apparent pattern for such productions. However, in this part of the experiment, there was the introduction of the perturbation. Thus, it was possible that the experimental phase (i.e., perturbation playback) could have caused the production of negative VOTs. In order to verify that there was no difference in the number of negative VOT productions across the experimental phases, we calculated the proportion of negative VOT occurrence in each experimental phase for each talker, and then performed a repeated-measures analysis of variance (ANOVA) on the proportion of negative VOT values with the experimental phase as a within-participant factor. The results showed that there was no effect of phase [$F(2, 24) = 1.38$, $p > 0.05$]. Thus, we excluded these negative VOTs from data for the word "dipper" from further analysis.

We also compared each talker's average production of the last 10 utterances (utterance 11 20) of the baseline phase of the second session with those of the first session to examine whether the participants produced target words differently between the two sessions. Among VOT, vowel duration, coda closure duration, and the overall syllable

TABLE I. Average temporal parameters (in ms) of the words "tipper" and "dipper" collected in session 1. *B* represents unstandardized regression coefficient from trial 11 to 100.

	VOT		Vowel		Coda closure		Syllable	
	Mean	*B*	Mean	*B*	Mean	*B*	Mean	*B*
tipper (s.d.)	51.8 (2.4)	0.03[a]	57.6 (2.6)	0.003	99.5 (3.1)	0.075[a]	208.8 (3.4)	0.043[a]
dipper (s.d.)	16.9 (1.0)	0.008	74.3 (2.6)	0.048[a]	109.4 (2.7)	0.015	200.6 (4.0)	0.072[a]

[a]$p < 0.05$.

duration, only the coda closure was significantly longer in the session 2 with both "tipper," [$t(12) = -2.23$, $p < 0.05$; mean difference = 7.24 ms, s.d. = 11.69 ms] and "dipper" [$t(12) = -2.49$, $p < 0.05$; mean difference = 6.85 ms, s.d. = 9.91 ms] being longer in session 2. Overall, participants were producing the test tokens similarly across the two sessions.

2. Compensatory production

Figure 2(a) shows the group average VOT over the course of 100 utterance. As can be seen, the talkers changed their VOT production immediately after the introduction of a cross-categorical feedback. The VOT for /t/ became longer while that for /d/ became shorter. The compensatory production reached a plateau within 20 utterances after the introduction of the perturbation. Moreover, once the perturbation was removed, talkers de-adapted the VOT production and the VOT slowly returned to the baseline. This adaptation pattern is consistent with the pattern reported with other acoustic parameters, such as formant compensatory production (Houde and Jordan, 1998; Purcell and Munhall, 2006; Villacorta *et al.*, 2007; Munhall *et al.*, 2009; MacDonald *et al.*, 2010, 2011; Mitsuya *et al.*, 2011, 2013).

In order to quantify this pattern, VOT was averaged across participants and trials for the three experimental phases: baseline (utterances 11 20), perturbation (41 60), and return (81 100). The last 20 utterances from the perturbation and return phases were analyzed because talker's production was stabilized during this part of the phase. A repeated measures ANOVA was conducted on VOT, vowel, coda closure, and syllabic duration with the experimental phase as a within-participant factor.

In the condition where talkers were producing "tipper" while receiving "dipper" during the perturbation phase (tipper condition, hereafter), the phase effect was significant on VOT [$F(2, 24) = 20.65$, $p < 0.01$] and syllabic duration [$F(2, 24) = 6.54$, $p < 0.01$], whereas there was no phase effect on vowel and coda closure (both $p > 0.05$). *Post hoc* analyses with Bonferroni correction (α set at 0.016 for three comparisons) were performed in order to compare the phases for the word "tipper." The VOT of the perturbed phase was 60.53 ms (s.d. = 15.84) and this was significantly longer than that of the baseline phase ($X = 50.27$ ms, s.d. = 13.75; $t[12] = -4.72$, $p < 0.016$), and the return phase [$X = 50.56$ ms, s.d. = 15.91;

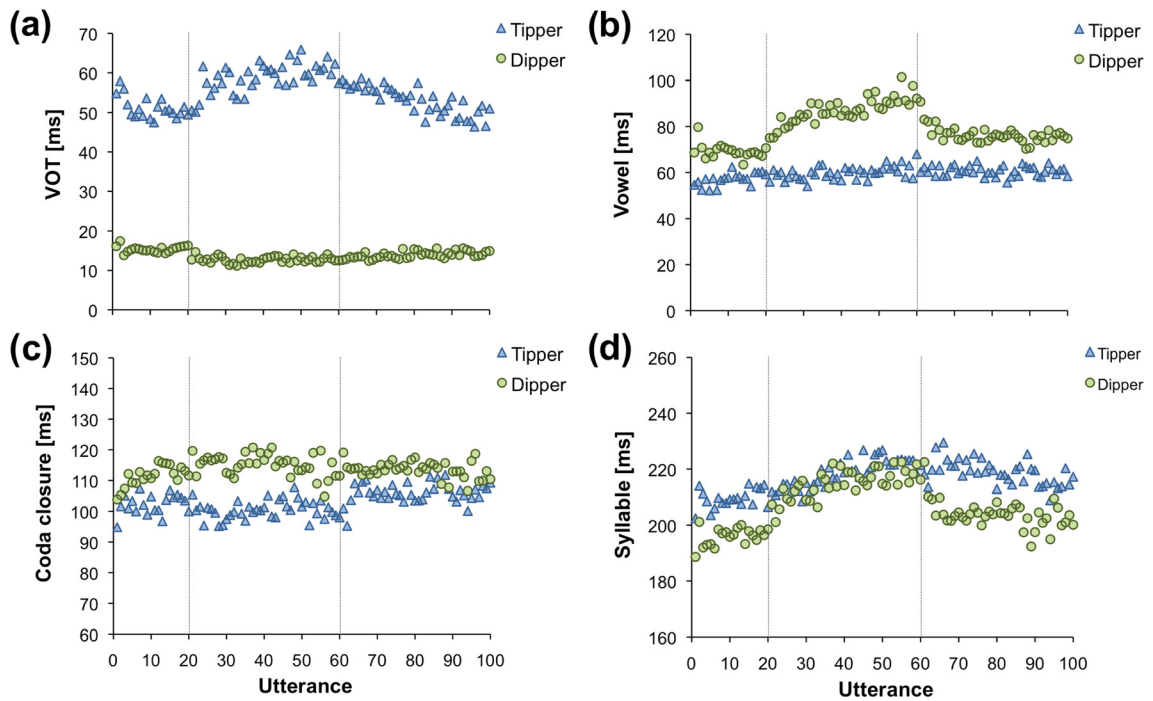

FIG. 2. (Color online) Group average of temporal parameters (in ms) over the course of the experiment of "tipper" (triangles) and "dipper" (circles); (a) VOT, (b) vowel, (c) coda closure, and (d) syllable. The vertical lines indicate the boundaries of the three experimental phases.

$t(12) = 5.15$, $p < 0.01$]. However, the difference between the baseline and return phases was not significant ($p > 0.016$). For syllable duration, the perturbation phase [$X = 222.16$ ms, s.d. = 29.18; $t(12) = -4.15$, $p < 0.016$] was significantly longer than the baseline phase ($X = 211.55$ ms, s.d. = 29.27), however, the return phase ($X = 217.55$ ms, s.d. = 30.02) was not significantly different from either phases (both $p > 0.016$).

In the condition where talkers were producing "dipper" while they receive "tipper" feedback (dipper condition, hereafter), an experimental phase effect was observed for VOT [$F(2,24) = 8.90$, $p < 0.01$], vowel duration [$F(2,24) = 32.24$, $p < 0.01$], and syllable duration [$F(2,24) = 8.93$, $p < 0.01$]. *Post hoc* analyses revealed that VOT of the perturbation phase [$X = 12.90$ ms, s.d. = 3.76] was significantly shorter than that of baseline [$X = 15.38$ ms, s.d. = 4.59; $t(12) = -4.72$, $p < 0.016$] and return phase [$X = 14.34$ ms, s.d. = 3.94; $t(12) = 5.15$, $p < 0.016$]. On the other hand, the vowel duration in the perturbation phase ($X = 90.24$ ms, s.d. = 22.36) was significantly longer than the baseline [$X = 68.08$ ms, s.d. = 15.22; $t(12) = -7.08$, $p < 0.016$] and the return phase [$X = 75.05$ ms, s.d. = 19.66; $t(12) = -4.72$, $p < 0.016$]. However, the baseline and return phases were not significantly different ($p > 0.016$). Similarly, the duration of syllable in the perturbation phase ($X = 217.72$ ms, s.d. = 25.09) was significantly longer than the baseline [$X = 197.12$ ms, s.d. = 21.88; $t(12) = -4.08$, $p < 0.016$], but was not different from the return phase ($X = 202.76$ ms, s.d. = 26.68, $p < 0.016$). The difference in syllable duration of the baseline and the return phases was also not significant ($p > 0.016$).

The magnitude of compensation was calculated by the difference between the average VOT during the last 20 tokens from the perturbation phase and the average VOT of the last 10 tokens of the baseline phase. In the tipper condition, the difference was multiplied by -1 since the compensatory VOT was longer than the baseline VOT. A paired sample *t*-test was conducted to examine the magnitude of compensation across the two conditions, and revealed that, the compensatory production with the word "tipper" ($X = 10.23$ ms, s.d. = 7.81) was significantly larger than that for the word "dipper" [$X = 2.47$ ms, s.d. = 2.54; $t(12) = -3.437$, $p < 0.05$]. We speculated that the smaller change in VOT with /d/ was due to the fact that the VOT for /d/ is already short, and that there is not much temporal allowance for shortening the already short temporal parameter. Thus, we examined the proportional change of the VOT production instead of the raw temporal scale. The VOT results were normalized by subtracting the talker's baseline average, which was calculated from the last 10 utterances of the baseline phase (i.e., utterances 11 20), from VOT value of each utterance and then dividing by the baseline average. The proportional results for each utterance, averaged across talkers, can be seen in Fig. 3. Interestingly, the magnitude of compensation for the perturbed VOT feedback was found to be similar across the two conditions, and it was roughly 15 20 % (dipper: $X = 15.68\%$, s.d. = 12.63; tipper: $X = 21.82\%$, s.d. = 19.19) and the difference across the voicing cognates was not significant [$t(12) = -1.09$, $p > 0.05$]. Both the time course pattern as well as the symmetry across the shift conditions observed here is similar to those of compensatory production for formant perturbation (Purcell and Munhall, 2006; Villacorta *et al.*, 2007; Munhall *et al.*, 2009; Mitsuya *et al.*, 2011, 2013).

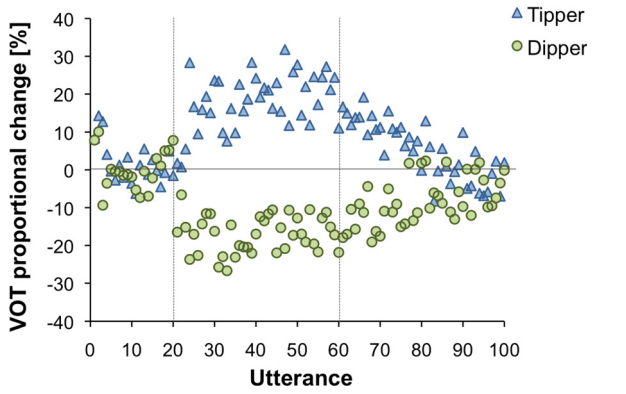

FIG. 3. (Color online) Average normalized VOT of utterances of "tipper" (triangles) and "dipper" (circles) over the course of the experiment. For each individual, the VOT measurements were normalized by diving by the average VOT of the last ten utterances of the baseline phase. The vertical lines indicate the boundaries of the three experimental phases.

In the current experiment, the magnitude of the *perturbation* applied varied across talkers since the perturbed playback tokens had a different VOT, depending on a talker's mean VOT production from the first session. Thus, the VOT change in absolute time does not necessarily capture how much talkers compensated relative to the magnitude of perturbation applied. In order to normalize the magnitude of compensation, we divided the magnitude of compensation by the difference between the mean VOT for /t/ and /d/ collected from the first session for each of the talkers, then averaged across them. In the tipper condition, the magnitude of change was on average 30.1% (s.d. = 27.4), however, in the dipper condition, the VOT change was much smaller ($X = 6.7\%$, s.d. = 5.0).

The compensatory production de-adapted in the return phase once the perturbations were removed. With the proportional change in VOT, we ran a curve fitting analysis on the trial 61 through 80. For both dipper and tipper conditions, the best fitting model was the linear model [model for dipper: $F(1, 18) = 4.57$, $p < 0.05$, B (unstandardized slope) = 0.458; tipper: $F(1,18) = 21.42$, $p < 0.001$, $B = -0.583$]. Because the slopes were rather mild, these results indicate that de-adaptation was gradual, with no rapid or exponential reduction of adaptation. The slow de-adaptation implies that the changes in VOT we observed in the perturbation phase were indeed the result of an adaptive learning (i.e., updated articulatory posture) based on the feedback they had been receiving. Interestingly, a slow and or otherwise a gradual de-adaptation was not observed with vowel and coda closure.

It is also important to note that during the post-experiment interview, all talkers reported that they had noticed that they heard their own voice producing the opposite voicing cognate.

IV. DISCUSSION

The purpose of the current investigation was to examine how people compensate for an opposite voicing cognate

feedback using the real-time auditory feedback paradigm. The paradigm of the current study employed a pseudo-real time manipulation in which the participants heard their own pre-recorded voice producing a cognate of the voicing category during the perturbation phase. The results showed that talkers changed the temporal structure of the intended syllable when they heard themselves say the word onset plosive of the opposite voicing cognate. While they were producing "tipper" and heard themselves produce "dipper," they lengthened the VOT as if they were trying to compensate for the short VOT feedback they received. Similarly, when producing "dipper" with "tipper" feedback, the talkers shortened their VOT as well as lengthened the vowel, making the VOT proportionally shorter. Moreover, the data showed that the proportional change of VOT was found to be similar in both conditions of perturbation. The importance of auditory feedback had been suggested for voicing control, however, it was unknown whether or not talkers would adapt to perturbed voicing category feedback in a single session and how sensitive the adaptation system would be to such a perturbation. Our results clearly indicate that the speech motor control system utilizes auditory feedback for voicing, and the pattern of compensatory production was similar to results reported in studies of spectral perturbation with only a partial compensation being observed (e.g., Houde and Jordan, 1998; Purcell and Munhall, 2006; Villacorta *et al.*, 2007; Munhall *et al.*, 2009; MacDonald *et al.*, 2010, 2011; Cai *et al.*, 2010; Mitsuya *et al.*, 2011, 2013).

One interesting observation about the current results is that in the condition in which talkers produced "dipper" while they received "tipper" feedback, they not only shortened the VOT but they also lengthened vowel duration when the perturbation was applied. We speculate that the lengthened vowel duration was a secondary way to compensate. Perception of voicing is known to be influenced by the proportion of VOT and vowel in a syllabic structure (Port and Rotunno, 1979; Port and Dalby, 1982). The English VOT for the voiced plosives is typically positive in the onset location of a stressed syllable. Because negative VOTs (prevoicing) are not systematically produced in the onset location, the controllable range of VOT is very limited. Thus, shortening an already short VOT might have been supplemented by lengthening the following vowel so that the proportion of the VOT within the syllable would, in turn, become much shorter. By doing so, the perception of voicing might have been more finely controlled.

The time course of compensatory production as well as the proportional magnitude of compensation reported here are very similar to what has been reported in the studies that employed vowel formant perturbations. Thus, one may speculate that the mechanism underlying the utilization of error feedback is functionally similar for both spectral and temporal information, or possibly that the two aspects of speech share some of the processing components even though the acoustic characteristics are vastly different between spectral and temporal properties.

However, the auditory perturbation applied in this study was pseudo-real time; talkers received pre-recorded sound as auditory feedback synchronous to their actual production. Manipulation had been already done before the experiment, thus, talkers' online articulatory/acoustic compensation for the perturbation was not reflected in the feedback. This raises a few important questions. First, why do speakers control something that they cannot control? In natural speech production control, without any experimental manipulations, auditory feedback *is always* the genuine resultant of articulation. Thus, with extensive previous experience of the use of the feedback to guide articulation, it is more parsimonious for the talker and adaptive to have a system that assumes that the feedback is a result of its own behavior (even with pre-recorded tokens, as long as articulation temporally match with the acoustic they hear). Even temporal contingency is not a strict constraint as delayed auditory feedback can influence articulation with long delays (> 200 ms). It seems that the auditory system acts as if auditory feedback is related to ongoing speech production. Even contingent sine waves that are shifted in frequency produce vocal pitch adjustments (Sivasankar *et al.*, 2005).

The importance of contingency of articulation and auditory feedback has been demonstrated by a study by Zheng *et al.* (2011). They reported that when their speakers were given pre-recorded voice tokens produced by a stranger, their F0 production changed in a systematic way, even though their change in F0 was not reflected by the feedback they were receiving. Moreover, their participants indicated that what they were hearing was their voice despite differences from the speakers' own voice in various acoustic parameters including temporal structure of the tested token. From the current study, and the findings of Zheng *et al.* (2011) together, it seems that "ownership" or the pairing between articulation and auditory feedback is flexibly defined.

In a similar vein, it is arguable that the awareness of the tokens being pre-recorded had any effect on speakers' production. We did not systematically ask our speakers if they "recognized the playback tokens as pre-recorded." However, we interpret the finding of the current study that speakers did change their production to indicate that their auditory feedback was processed as self-produced sounds, regardless of the awareness of the nature of the tokens.

Second, which is related to the first question, how can speakers continue to try to compensate for the perturbation when the compensation is not successful? We need to recognize that in auditory feedback experiments, compensation is not always complete. Therefore, it is more than acoustic faithfulness that the system is trying to achieve. Recent studies indicate that phonemic processes also influence speakers' compensatory behavior (Mitsuya *et al.*, 2011; Mitsuya *et al.*, 2013; Reilly and Dougherty, 2013; Niziolek and Guenther, 2013), as well as somatosensory feedback (e.g., Lametti *et al.*, 2012), suggesting the speech motor target is inherently multivariate/multimodal. As suggested by Feng *et al.* (2011), perhaps information in the auditory domain may be weighed heavily for motor control but it is not the sole contributor.

We believe that it is not only the acoustic disparity (error) the system tries to *reduce* but also it is the phonological/phonemic distinctiveness that the system tries to *maintain*. Because the system has learned the phonological contrastiveness of voicing and phonetic details of such contrastiveness (i.e., a host of many articulatory consequences associated with voicing), commands can be updated without

necessarily checking to see if compensation is successful. It is an open question and one not addressed here whether the feedback system is evaluating the trajectory of compensation or just dealing locally with error correction. If the former is true, then at some point the sense of ownership and contingency of the feedback would break down.

We are not trying to imply that compensation does not rely on auditory feedback. Rather, we suggest that reduction of acoustic error in the feedback domain is not the only goal and/or the source of feedback. Phonological/phonemic representation and articulatory plans of phonemic contrastiveness can contribute to compensatory behavior independently from auditory feedback.

Because the moment-to-moment compensation was not reflected in the auditory feedback, the acoustic changes we observed here are thus due to changes/updates in articulatory plans (feedforward) rather than small adjustment of on-going articulation (error reduction based on feedback). Such articulatory/acoustic changes due to updated speech motor plans are also reflected on the aftereffect we observed. Interestingly, the aftereffect was strongest with the VOT measures, which is indicated by a slow de-adaptation pattern. Although changes were observed with many of these parameters with our manipulations, the different degrees of aftereffect suggests that with the cross-categorical perturbation that we employed here, the speech motor system was more sensitive to the articulatory/phonetic features that differentiated the voicing contrasts of /t/ and /d/, than other temporal differences of the two syllables.

The current data also provide insight into the nature of the target of speech production with the implication that adaptation might be operated at more linguistic level than at the pure acoustic level. Language specific compensatory production has been reported in the vocalization (e.g., Liu *et al.*, 2010; Chen *et al.*, 2010). One example of linguistic influence on compensatory speech production in vowel formant paradigm is that Mitsuya *et al.* (2013) had French talkers and English talkers produce a phonologically identical vowel /ɛ/ while its F2 was lowered. This made the feedback sound slightly rounded. Because French has front rounded vowels, and there is a vowel that sounds similar to the perturbed vowel, French talkers compensated more than English talkers. But what was more interesting is that French talkers also changed their F3 production in response to F2 perturbation. F2 and F3 are known to covary for roundedness. Thus, it was interpreted that these two acoustic parameters were coupled functionally because the language specified the function. This implies that the production target is also multi-dimensionally specified (Mitsuya *et al.*, 2013).

Similar to the current study, Mochida *et al.* (2010) used pre-recorded playback sounds to manipulate feedback timing and feedback syllable. Their talkers kept producing a syllable at a constant interval with auditory feedback. Subjects expected to hear the feedback of the intended syllable, yet they sometimes received the pre-recorded sound of their own voice producing a different syllable. This feedback manipulation was done with various feedback delays to examine how labial movements of an on-going articulation compensated. They found that when the intended syllable was played at −50 ms relative to the actual onset of production, significant changes in labial movements were observed, however, not at −100 ms or prior. Moreover, no changes in the movements were observed with unintended tokens played back independent of the timing. Based on the results, they concluded that congruency in syllabic identity between acoustic prediction and the actual feedback is an important factor for the talkers to exhibit compensatory articulation. These results seem to contradict the current results where our talkers compensated with a large difference between intended versus actual feedback (cross-categorical feedback). The differences in results might have been due to the differences in (1) measurements used (acoustic versus articulatory) and (2) the degree of dissimilarity between the intention and feedback. However, it is important to note that in both studies the employment of pre-recorded feedback has been shown to be effective in eliciting compensatory production. This pseudo-real time auditory perturbation paradigm opens doors to many other investigations of articulatory coordination and the use of auditory feedback in speech production.

Many neural-based speech production models, such as DIVA (Guenther *et al.*, 1998), state feedback control (Houde and Nagarajan, 2011), and the neurocomputational model (Kröger *et al.*, 2009) describe speech production targets as phonemic (or larger units). Despite the differences in the theoretical details in their models, all have a process of error detection for controlling ongoing and future production of speech. The current study as well as Mitsuya *et al.* (2013) showed that both error estimation and error reduction can be achieved with more than one parameter with multi-dimensionally represented phonemes. Models of speech production should take such representations into account in order to model accurate speech production control. Future investigations measuring the movements of the articulatory apparatus are needed in order to fully understand how acoustic configurations for voicing are monitored by the nervous system and how such perceptual processes are transformed into fine adjustment of multi-articulator gestures.

ACKNOWLEDGMENTS

This research was supported by the National Institute of Deafness and Communicative Disorder Grant No. DC-08092, and the National Sciences and Engineering Research Council of Canada.

Allen, J. S., and Miller, J. L. (**1999**). "Effects of syllable initial voicing and speaking rate on the temporal characteristics of monosyllabic words," J. Acoust. Soc. Am. **106**, 2031 2039.

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

Behroozmand, R., Karvelis, L., Liu, H., and Larson, C. R. (**2009**). "Vocalization induced enhancement of the auditory cortex responsiveness during voice F0 feedback perturbation," Clin. Neurophysiol. **120**, 1303 1312

Buhusi, C. V., and Meck, W. H. (**2005**). "What makes us tick? Functional and neural mechanisms of interval timing," Nat. Rev. Neurosci. **6**, 755 765.

Buonomano, D. V., and Laje, R. (**2010**). "Population clocks: Motor timing with neural dynamics," Trends Cognit. Sci. **14**, 520 527.

Burnett, T. A., Freedland, M. B., Larson, C. R., and Hain, T. C. (**1998**). "Voice f0 responses to manipulations in pitch feedback," J. Acoust. Soc. Am. **103**, 3153 3161.

Cai, S., Ghosh, S. S., Guenther, G. H., and Perkell, J. S. (**2010**). "Adaptive auditory feedback control of the production of formant trajectories in the mandarin triphthong /iau/ and its pattern of generalization," J. Acoust. Soc. Am. **128**, 2033 2048.
Cai, S., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (**2011**). "Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within syllable and between syllable speech timing," J. Neurosci. **31**, 16483 16490.
Casserly, E. D. (**2011**). "Speaker compensation for local perturbation of fricative acoustic feedback," J. Acoust. Soc. Am. **129**, 2181 2190.
Chen, Z., Liu, P., Wang, E. Q., Larson, C. R., and Liu, H. (**2010**). "Online monitoring of auditory feedback is sensitive to language experience," J. Acoust. Soc. Am. **127**, 2022 2022.
Cowie, R. J., and Douglas Cowie, E. (**1983**). "Speech production in profound postlingual deafness," in *Hearing Science and Hearing Disorders*, edited by M. E. Lutman and M. P. Haggard (Academic Press, New York), pp. 183 230.
Das, S., and Hansen, J. H. (**2004**). "Detection of voice onset time (VOT) for unvoiced stops (/p/,/t/,/k/) using the Teager energy operator (TEO) for auto matic detection of accented English," in *Proceedings of the 6th Nordic Signal Processing Symposium (NORSIG 2004)*, Espoo, Finland, pp. 344 347.
Feng, Y., Gracco, V. L., and Max, L. (**2011**). "Integration of auditory and somatosensory error signals in the neural control of speech movements," J. Neurophysiol. **106**, 667.
Fowler, C. A. (**1981**). "A relationship between coarticulation and compensa tory shortening," Phonetica **38**, 35 50.
Fowler, C. A. (**1986**). "An event approach to the study of speech perception from a direct realist perspective," J. Phonetics **14**, 3 28.
Guenther, F. H., Hamoson, M., and Johnson, D. (**1998**). "A theoretical investigation of reference frames for the planning of speech movements," Psychol. Rev. **105**, 611 633.
Hain, T. C., Burnett, T. A., Larson, C. R., and Kiran, S. (**2001**). "Effects of delayed auditory feedback (DAF) on the pitch shift reflex," J. Acoust. Soc. Am. **109**, 2146 2152.
Higgins, M. B., McCleary, E. A., and Schulte, L. (**2001**). "Articulatory changes with short term deactivation of the cochlear implants of two pre lingually deafened children," Ear Hear. **22**, 29 41.
Houde, J. F., and Jordan, M. I. (**1998**). "Sensorimotor adaptation in speech production," Science **279**, 1213 1216.
Houde, J. F., and Nagarajan, S. S. (**2011**). "Speech production as state feed back control," Front. Hum. Neurosci. **5**, 82.
Ivry, R. B., and Richardson, T. C. (**2002**). "Temporal control and coordina tion: the multiple timer model," Brain Cognit. **48**, 117 132.
Jones, D. L., Gao, S., and Svirsky, M. A. (**2003**). "The effect of short term auditory deprivation on the control of intraoral pressure in pediatric coch lear implant users," J. Speech Lang. Hear. Res. **46**, 658 669.
Jones, J. A., and Munhall, K. G. (**2000**). "Perceptual calibration of F0 pro duction: Evidence from feedback perturbation," J. Acoust. Soc. Am. **108**, 1246 1251.
Kent, R., and Moll, K. (**1975**). "Articulatory timing in selected consonant sequences," Brain Lang. **2**, 304 323.
Kessinger, R. H., and Blumstein, S. E. (**1998**). "Effects of speaking rate on voice onset time and vowel production: Some implications for perception studies," J. Phonetics **26**, 117 128.
Kroger, B. J., Kannampuzha, J., and Neuschaefer Rube, C. (**2009**). "Towards a neurocomputational model of speech production and perception," Speech Commun. **51**, 793 809.
Lametti, D. R., Nasir, S. M., and Ostry, D. J. (**2012**). "Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback," J. Neurosci. **32**, 9351 9358.
Lane, H., and Perkell, J. S. (**2005**). "Control of voice onset time in the absence of hearing: A review," J. Speech Lang. Hear. Res. **48**, 1334 1343.
Lane, H., Wozniak, J., and Perkell, J. (**1994**). "Changes in voice onset time in speakers with cochlear implants," J. Acoust. Soc. Am. **96**, 56 64.
Lashley, K. (**1951**). "The problem of serial order in behavior," in *Cerebral Mechanisms in Behavior*, edited by L. Jeffress (Wiley, New York), pp. 112 136.
Lehiste, I. (**1970**). *Suprasegmentals* (Cambridge University Press, Cambridge, UK), pp. 1 194.
Lisker, L. (**1986**). "'Voicing' in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees," Lang. Speech **29**, 3 11.
Lisker, L., and Abramson, A. S. (**1964**). "A cross language study of voicing in initial stops: Acoustical measurements," Word **20**, 384 422.
Liu, H., Auger, J., and Larson, C. R. (**2009**). "Voice fundamental frequency modulates vocal response to pitch perturbations during English speech," J. Acoust. Soc. Am. **127**, EL1 EL5.
Liu, H., Wang, E. Q., Chen, Z., Liu, P., Larson, C. R., and Huang, D. (**2010**). "Effect of tonal native language on voice fundamental frequency responses to pitch feedback perturbations during sustained vocalizations," J. Acoust. Soc. Am. **128**, 3739 3746.
Lofqvist, A., and Yoshioka, H. (**1981**). "Interarticulator programming in obstruent production," Phonetica **38**, 21 34.
Lofqvist, A., and Yoshioka, H. (**1984**). "Intrasegmental timing: Laryngeal oral coordination in voiceless consonant production," Speech Commun. **3**, 279 289.
MacDonald, E. N., Goldberg, R., and Munhall, K. G. (**2010**). "Compensation in response to real time formant perturbations of different magnitude," J. Acoust. Soc. Am. **127**, 1059 1068.
MacDonald, E. N., Purcell, D. W., and Munhall, K. G. (**2011**). "Probing the independence of formant control using altered auditory feedback," J. Acoust. Soc. Am. **129**, 955 966.
MacNeilage, P. F. (**1970**). "Motor control of serial ordering of speech," Psychol. Rev. **77**, 182 196.
Mitsuya, T., MacDonald, E. N., Purcell, D. W., and Munhall, K. G. (**2011**). "A cross language study of compensation in response to real time formant perturbation," J. Acoust. Soc. Am. **130**, 2978 2986.
Mitsuya, T., Samson, F., Ménard, L., and Munhall, K. G. (**2013**). "Language dependent vowel representation in speech production," J. Acoust. Soc. Am. **133**, 2993 3003.
Mochida, T., Gomi, H., and Kashino, M. (**2010**). "Rapid change in articula tory lip movement induced by preceding auditory feedback during produc tion of bilabial plosives," PloS One **5**, e13866.
Munhall, K. G., Fowler, C. A., Hawkins, S., and Saltzman, E. (**1992**). "Compensatory shortening in monosyllables of spoken English," J. Phonet. **20**, 225 239.
Munhall, K. G., Lofqvist, A., and Kelso, J. A. S. (**1994**). "Lip larynx coordi nation in speech: Effects of mechanical perturbations to the lower lip," J. Acoust. Soc. Am. **95**, 3605 3616.
Munhall, K. G., MacDonald, E. N., Byrne, S. K., and Johnsrude, I. (**2009**). "Speakers alter vowel production in response to real time formant pertur bation even when instructed to resist compensation," J. Acoust. Soc. Am. **125**, 384 390.
Niziolek, C. A., and Guenther, F. H. (**2013**). "Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations," J. Neurosci. **33**, 12090 12098.
Port, R. F., and Dalby, J. (**1982**). "Consonant/vowel ratio as a cue for voic ing in English," Percept. Psychophys. **32**, 141 152.
Port, R. F., and Rotunno, R. (**1979**). "Relation between voice onset time and vowel duration," J. Acoust. Soc. Am. **66**, 654 662.
Purcell, D. W., and Munhall, K. G. (**2006**). "Adaptive control of vowel formant frequency: Evidence from real time formant manipulation," J. Acoust. Soc. Am. **120**, 966 977.
Reilly, K. J., and Dougherty, K. E. (**2013**). "The role of vowel perceptual cues in compensatory responses to perturbations of speech auditory feedback," J. Acoust. Soc. Am. **134**, 1314 1323.
Shiller, D. M., Sato, M., Gracco, V. L., and Baum, S. R. (**2009**). "Perceptual recalibration of speech sounds following speech motor learning," J. Acoust. Soc. Am. **125**, 1103 1113.
Sivasankar, M., Bauer, J. J., Babu, T., and Larson, C. R. (**2005**). "Voice responses to changes in pitch of voice or tone auditory feedback," J. Acoust. Soc. Am. **117**, 850 857.
Tuller, B., Ding, M., and Kelso, J. S. (**1997**). "Fractal timing of verbal trans forms," Perception **26**, 913 928.
Tuller, B., and Kelso, J. S. (**1984**). "The timing of articulatory gestures: Evidence for relational invariants," J. Acoust. Soc. Am. **76**, 1030 1036.
Villacorta, V. M., Perkell, J. S., and Guenther, F. H. (**2007**). "Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception," J. Acoust. Soc. Am. **122**, 2306 2319.
Waldstein, R. S. (**1990**). "Effects of postlingual deadness on speech produc tion: Implications for the role of auditory feedback," J. Acoust. Soc. Am. **88**, 2099 2114.
Warren, R. M. (**1961**). "Illusory changes of distinct speech upon repeti tion The verbal transformation effect," Br. J. Psychol. **52**, 249 258.
Washburn, M. F. (**1916**). *Movement and Mental Imagery: Outlines of a Motor Theory of the Complexer Mental Processes* (Houghton Mifflin, Boston), pp. 1 252.
Zheng, Z. Z., MacDonald, E. N., Munhall, K. G., and Johnsrude, I. S. (**2011**). "Perceiving a Stranger's Voice as Being One's Own: A 'Rubber Voice'Illusion?" PloS one **6**, e18655.