Focal Manipulations of Formant Trajectories Reveal a Role of Auditory Feedback in the Online Control of Both Within-Syllable and Between-Syllable Speech Timing

Behavioral/Systems/Cognitive

Focal Manipulations of Formant Trajectories Reveal a Role of Auditory Feedback in the Online Control of Both Within-Syllable and Between-Syllable Speech Timing

Shanqing Cai,[1,2] Satrajit S. Ghosh,[1,2] Frank H. Guenther,[2,3,4] and Joseph S. Perkell[1,2]

[1]Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology (MIT), and [2]Speech and Hearing Bioscience and Technology Program, Harvard–MIT Division of Health Sciences and Technology, Cambridge, Massachusetts 02139, and [3]Department of Speech, Language, and Hearing Sciences and [4]Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215

Within the human motor repertoire, speech production has a uniquely high level of spatiotemporal complexity. The production of running speech comprises the traversing of spatial positions with precisely coordinated articulator movements to produce 10–15 sounds/s. How does the brain use auditory feedback, namely the self-perception of produced speech sounds, in the online control of spatial and temporal parameters of multisyllabic articulation? This question has important bearings on the organizational principles of sequential actions, yet its answer remains controversial due to the long latency of the auditory feedback pathway and technical challenges involved in manipulating auditory feedback in precisely controlled ways during running speech. In this study, we developed a novel technique for introducing time-varying, focal perturbations in the auditory feedback during multisyllabic, connected speech. Manipulations of spatial and temporal parameters of the formant trajectory were tested separately on two groups of subjects as they uttered "I owe you a yo-yo." Under these perturbations, significant and specific changes were observed in both the spatial and temporal parameters of the produced formant trajectories. Compensations to spatial perturbations were bidirectional and opposed the perturbations. Furthermore, under perturbations that manipulated the timing of auditory feedback trajectory (slow-down or speed-up), significant adjustments in syllable timing were observed in the subjects' productions. These results highlight the systematic roles of auditory feedback in the online control of a highly over-learned action as connected speech articulation and provide a first look at the properties of this type of sensorimotor interaction in sequential movements.

Introduction

During the production of running speech, the brain is faced with the complex task of rapidly and accurately sequencing movements of multiple articulators (e.g., tongue and lips). The complexity of speech articulation can be described in two domains: (1) the spatial domain, which concerns the millimeter precision of articulatory positions (or the corresponding acoustic correlates such as resonance peaks in the speech sound spectrum), and (2) the temporal domain, including the precise timing between successive articulatory movements. Articulation occurs at a dauntingly high rate, with the durations of single phonemes (vowels or consonants) in the vicinity of 70–110 ms in running speech (Crystal and House, 1988). Rich afferent information through sensory channels is available to the brain during speech production. In particular, auditory feedback—the speaker's auditory perception of his or her own speech—has been shown to play important roles in speech motor control (Burnett et al., 1998; Houde and Jordan, 1998; Purcell and Munhall, 2006). However, given the high rate at which speech movements are sequenced and the relatively long latency of the auditory feedback control loop [120 ms or greater during quasistatic speech sounds (Burnett et al., 1998; Purcell and Munhall, 2006; Tourville et al., 2008)], it remains unclear whether ongoing speech movements, especially their timing, can be affected by auditory feedback (cf., Lane and Tranel, 1971; Borden, 1979; Howell and Sackin, 2002).

It is technically challenging to manipulate auditory feedback during running speech in precisely controlled ways. The effects on ongoing speech by less technically challenging alterations of auditory feedback, such as masking noise and delayed auditory feedback, have been studied carefully (Fairbanks, 1955; Van Summers et al., 1988), but these gross, nonspecific alterations are of limited value in understanding speech under ordinary circumstances. Limb reaching, an intensively studied form of biological movement, uses visual feedback to guide spatial trajectories online (Desmurget and Grafton, 2000; Saunders and Knill, 2003). However, this feedback-based regulation may not extrapolate to speech because of the unique spatiotemporal complexity and constraints in speech production. An approach to dealing with such issues is to use more focal

Received July 17, 2011; revised Sept. 8, 2011; accepted Sept. 29, 2011.

Author contributions: S.C., S.S.G., F.H.G., and J.S.P. designed research; S.C. performed research; S.C. analyzed data; S.C., S.S.G., F.H.G., and J.S.P. wrote the paper.

This work was supported by NIH Grants R01-DC0001925, R01-DC007683, and R56-DC010849; and NSF Doctoral Dissertation Research Improvement Grant 1051566. We thank David Ostry and Jason Tourville for helpful comments on this manuscript.

Correspondence should be addressed to Shanqing Cai, Massachusetts Institute of Technology, MIT Building 36 Room 585, 50 Vassar Street, Cambridge, MA 02139. E-mail: cais@mit.edu.

DOI:10.1523/JNEUROSCI.3653-11.2011

 0270-6474/11/3116483-08$15.00/0

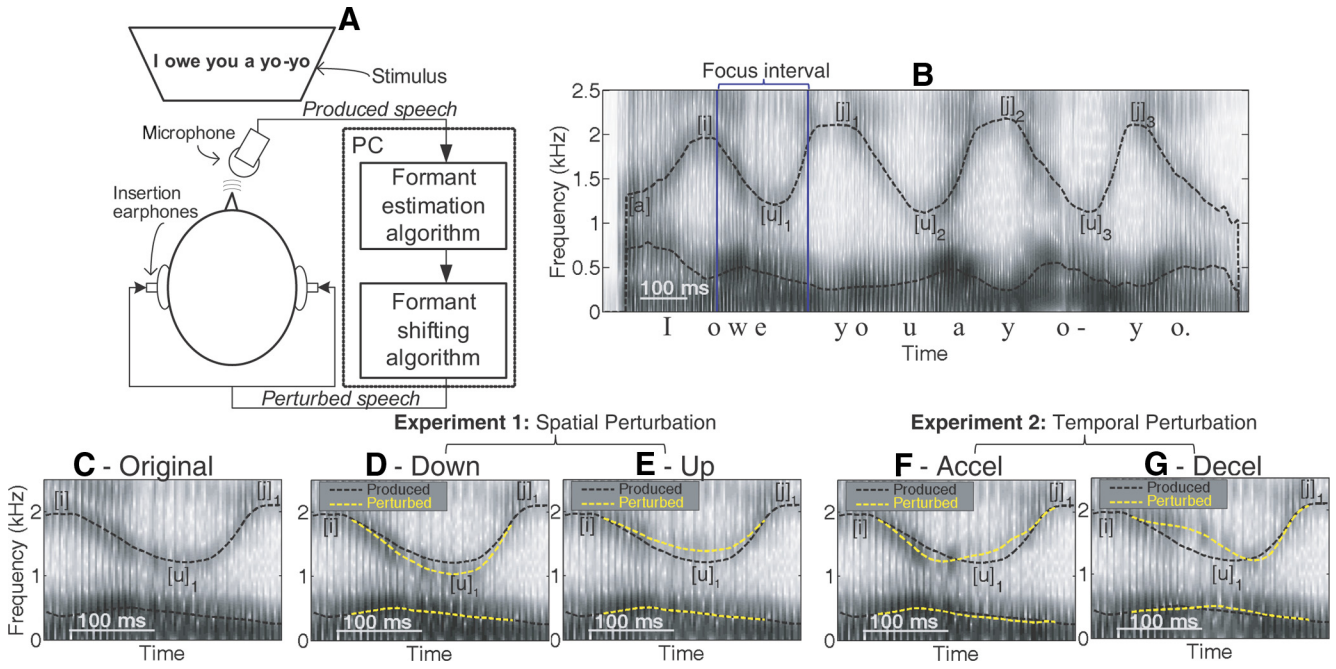

Figure 1. Examples of the spatial perturbations to the auditory feedback of the F2 trajectory. ***A***, A schematic diagram showing the setup for auditory feedback perturbation. ***B***, An example spectrogram of the stimulus utterance "I owe you a yo-yo" with the F1 and F2 trajectories (dashed black lines) overlaid and the F2 extrema labeled. The blue vertical lines highlight the focus interval, that is, the interval during which the auditory perturbation occurs. ***C***, A zoomed-in view of the focus interval of the example in ***B***. ***D***, ***E***, Resultant spectrograms of Down (***D***) and Up (***E***) perturbation on the same sound as shown in ***C***. The yellow dashed lines show the shifted F2. For comparison, the black dashed lines in ***D*** and ***E*** are identical to those in ***C***. ***F***, ***G***, Examples of the temporal perturbations Accel and Decel, in the same format as ***D*** and ***E***.

and subtle perturbations of auditory feedback and observe their effects on multisyllabic articulation.

In the current study, we imposed phoneme-specific perturbations of the trajectory of the second formant frequency (F2) in subjects' auditory feedback when they produced a multisyllabic utterance. F2 is one of the most perceptually important resonance peaks in vowel-like speech sounds. It is determined primarily by, and thus reflects, the positions of the tongue and the lips during articulation. The perturbation technique we used was novel in two main respects: it focused on time-varying transitions in multisyllabic speech and it manipulated the timing of events (sped up or slowed down) in auditory feedback. We demonstrate small but significant spatial compensation (Experiment 1) and temporal adjustments (Experiment 2) in parameters of the subjects' articulations, which highlight the use of auditory feedback by the brain in fine-tuning the spatiotemporal parameters of connected speech articulation on a moment-by-moment basis.

Materials and Methods

Subjects. A total of 41 volunteers, naive to the purpose of this study, participated in the two experiments of this study. These subjects were adult native speakers of American English with no self-reported history of speech, language, or neurological disorders. Pure-tone audiometry confirmed that all participants had auditory thresholds within the normal range at 0.5, 1, 2, and 4 kHz in both ears. Thirty (26 male, 4 female; age range: 19.2–42.6 years, median age: 22.8 years) participated in Experiment 1, which involved perturbations of spatial parameters of auditory feedback (see Auditory feedback setup and speech task, below). Twenty-two subjects (20 male, 2 female; age range: 19.2–47.1 years, median age: 23.3 years) participated in Experiment 2, which involved perturbations of the temporal parameters of auditory feedback. Eleven of the 41 subjects participated in both Experiments 1 and 2. The recruitment of more male than female subjects in this study was mainly due to the need to have relatively clean and reliable formant tracking to ensure the quality of the feedback perturbation. The only study related to gender difference in the feedback control of speech production that we are aware of is Chen et al. (2011), which showed slightly greater magnitude (~15%) and longer latency (~13%) of response to perturbation of the auditory feedback of vocal pitch in males than in females. However, to our knowledge, there exists no evidence for qualitative differences in auditory feedback-based articulatory control. Therefore, although caution should be taken when generalizing the quantitative details in the findings of this study to the general population, the qualitative conclusions of this study should be broadly relevant.

The Institutional Review Board of MIT approved the experimental protocols.

Auditory feedback setup and speech task. A schematic diagram of the auditory perturbation setup in both Experiments 1 and 2 is shown in Figure 1*A*. E-A-R insertion earphones (Aearo Technologies) provided the subjects with unperturbed and experimentally perturbed auditory feedback. The level of the auditory feedback provided through the insertion earphones was amplified (+14 dB relative to the level at a microphone fixed 10 cm from the subject's mouth), which, together with the exclusion effect of the ear tips, served to mask the unprocessed auditory feedback. In Experiment 1, spatial perturbations (Down and Up, see Auditory feedback setup and speech task, below) were used; in Experiment 2, temporal perturbations [Accelerating (Accel) and Decelerating (Decel); see Perturbations to the auditory feedback of formant trajectory, below) were used.

In the main part of the experiment, the subject read aloud the multisyllabic utterance "I owe you a yo-yo" 160 times. This stimulus utterance was chosen due to its composition entirely of vowels and semivowels, which led to relatively sustained voicing. This facilitated the real-time algorithmic extraction of formant frequency values and the introduction of perturbations to the spatiotemporal parameters of the formant trajectories. In addition, the F2 trajectory of this utterance contains a number of well defined local minima and maxima (Fig. 1*B*). These local extrema are used as landmarks in defining the spatial and temporal measures of the underlying articulation. For simplicity of notation, we use the symbols provided in Table 1 to represent the extrema in the F2 trajectory during this utterance.

Table 1. Ad hoc phonetic symbols used in the current paper to denote the F2 extrema in the utterance "I owe you a yo-yo" and the baseline values of the F2 at these extrema and their timing with respect to the first maximum ([i])

		Latency re. [i] under noPert (ms; Mean ± 1 SD)		F2 under noPert (Hz;, mean ± 1SD)	
Symbol	Landmark	Experiment 1	Experiment 2	Experiment 1	Experiment 2
[i]	F2 maximum at the end of "I"	0	0	1933 ± 163	1938 ± 123
$[u]_1$	F2 minimum at the end of "owe"	162.8 ± 26.0	160.7 ± 28.2	1144 ± 137	1149 ± 112
$[j]_1$	F2 maximum at the onset of "you"	294.7 ± 31.6	288.7 ± 29.0	2155 ± 199	2179 ± 154
$[u]_2$	F2 minimum at the end of "you"	465.5 ± 39.0	441.9 ± 29.2	1072 ± 127	1069 ± 113
$[j]_2$	F2 maximum at the onset of the 1st "yo"	645.8 ± 56.6	620.7 ± 34.0	2184 ± 231	2211 ± 157
$[u]_3$	F2 minimum at the end of the 1st "yo"	835.4 ± 64.2	794.7 ± 39.3	1184 ± 124	1224 ± 116
$[j]_3$	F2 minimum at the onset of the 2nd "yo"	941.3 ± 66.7	896.4 ± 37.2	2119 ± 201	2124 ± 125

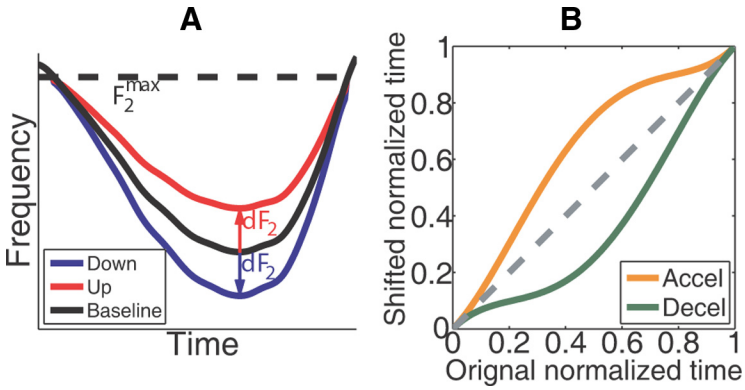

Figure 2. Schematic illustrations of the spatial (Down and Up) and temporal (Accel and Decel) perturbations. ***A***, The mapping between the original and perturbed values of F2 in the spatial (Down and Up) perturbations. ***B***, The time-warping functions used in the temporal (Accel and Decel) perturbations.

Before the main data-gathering phase, the subjects were trained to produce the utterance within a medium range of vocal intensity (74–84 dB SPL, A-weighted) and a medium range of speaking rate (1.2–1.4 s sentence duration). In the main part of the experiment, visual feedback regarding the success or failure of achieving these ranges was provided on the computer monitor after each utterance to ensure relatively consistent vocal intensity and speaking rate throughout the experiment and across subjects. On average, the subjects were able to speak within these ranges of intensity and speaking rate in 90.9% of the trials in Experiment 1 and in 94.8% of the trials in Experiment 2. No trials were excluded from subsequent analysis solely on the basis of intensity and/or speaking rate errors. Trials with audible speech errors and/or disfluencies (0.76% of all trials in Experiment 1 and 0.48% in Experiment 2) were discarded. Trials were arranged into blocks in both Experiment 1 and 2. A block consisted of eight trials, of which six contained no perturbation to auditory feedback (noPert) and the other two contained two opposite types of perturbations. Experiment 1 used spatial perturbations, namely the Down and Up perturbations. Experiment 2 used temporal perturbations, i.e., the Accel and Decel perturbations (see Perturbations to the auditory feedback of formant trajectory, below). The order of the trials in each block was pseudorandomized with the constraint that no two adjacent trials both contain perturbations. To reduce the monotony of the task, a filler sentence [drawn from the Harvard IEEE sentence corpus (Rothauser et al., 1969)], different from the main stimulus utterance, was read aloud by the subject between successive blocks.

Perturbations to the auditory feedback of formant trajectory. The methods for real-time formant tracking and shifting were adapted from a Matlab Mex-based digital signal processing software package that has been described previously (Cai et al., 2010). The latency of the artificial auditory feedback loop was an imperceptible 11 ms. In the current study, the time-varying perturbation to auditory feedback was focused on the section of the stimulus utterance during the second syllable, "owe" ([ou]), and the transition from the end of this syllable to the beginning of the next one, "you" ([ju]), which we refer to as the focus interval (Fig. 1*B*, bracket). The second formant (F2) trajectory of the stimulus utterance contains a number of well defined local maxima and minima, which formed the basis for the online tracking of sentence progress and the online detection of the focus interval.

Two categories of perturbations, namely spatial and temporal perturbations of auditory feedback, were used in Experiments 1 and 2, respectively. Experiment 1 involved the spatial perturbations Down and Up, which altered the magnitude of F2 at $[u]_1$, viz., the F2 minimum during the syllable "owe," without changing the timing of this minimum. The Down perturbation decreased the magnitude of F2 at $[u]_1$, leading to an exaggeration of the downward sweep of F2 during the syllable "owe" (Fig. 1*D*). The Up perturbation had the opposite effect (Fig. 1*E*). The perturbations were implemented so as to preserve the continuity and smoothness of the formant tracks and ensure naturalness of the altered auditory feedback. In terms of the correspondence to articulatory movements, the Down perturbation created a percept in which the extents of the backward movement of the tongue and/or rounding of the lips during the diphthong [ou] were exaggerated. Conversely, the Up perturbation diminished the auditory percept of the extent of these movements.

The Down and Up perturbations were implemented as a mapping from the original F2 value to the perturbed one during the focus interval. Specifically, the mapping was: $F_2'(t) = F_2(t) - dF_2(t) = F_2(t) - k \times (F_2^{\max} - F_2(t))$, if $F_2(t) < F_2^{\max}$, in which F_2 is the original F2, F_2' is the perturbed F2 in the auditory feedback, k is the coefficient of perturbation (set to 0.25 for both Down and Up perturbations for all subjects in Experiment 1), and $F_2^{\max}$ is the subject-specific perturbation limit, extracted automatically from the practice trials before the main part of the experiment (Fig. 2*A*).

The Accel and Decel perturbations, used in Experiment 2, differed from the spatial perturbations in that they altered the perceived timing of the F2 minimum at $[u]_1$ while approximately preserving the magnitude. An Accel perturbation in the feedback signal led to an advancing of the F2 minimum at $[u]_1$ in time by an average of 45.4 ms (Fig. 1*F*), whereas a Decel perturbation led to a delaying of the F2 minimum at $[u]_1$ in time by an average of 24.6 ms (Fig. 1*G*). Thus, the Accel perturbation led to a perception that the minimum at $[u]_1$ would occur earlier than expected and the Decel perturbation led to a perception that it would occur later than expected. The technical details on these online time-varying spatial and temporal perturbations are described below.

The Accel and Decel perturbations were achieved through time-warping in the focus interval, governed by the following equation,

$$F_2'\left(\frac{t - t_0}{T_{\text{est}}}\right) = F_2\left(W\left(\frac{t - t_0}{T_{\text{est}}}\right)\right), \quad \text{when } t < t_0 + \bar{D},$$

wherein t_0 is earliest time at which $F_2(t) < F_2^{\max}$ is satisfied (i.e., onset of the perturbation), T_{est} is the duration of the focus interval estimated and updated online based on the preceding trials, $W(\cdot)$ is a fourth-order polynomial time-warping function (Fig. 2*B*), and $\bar{D}$ is the subject-specific average duration of the focus interval computed from previous trials, which was updated adaptively during the course of the experiment.

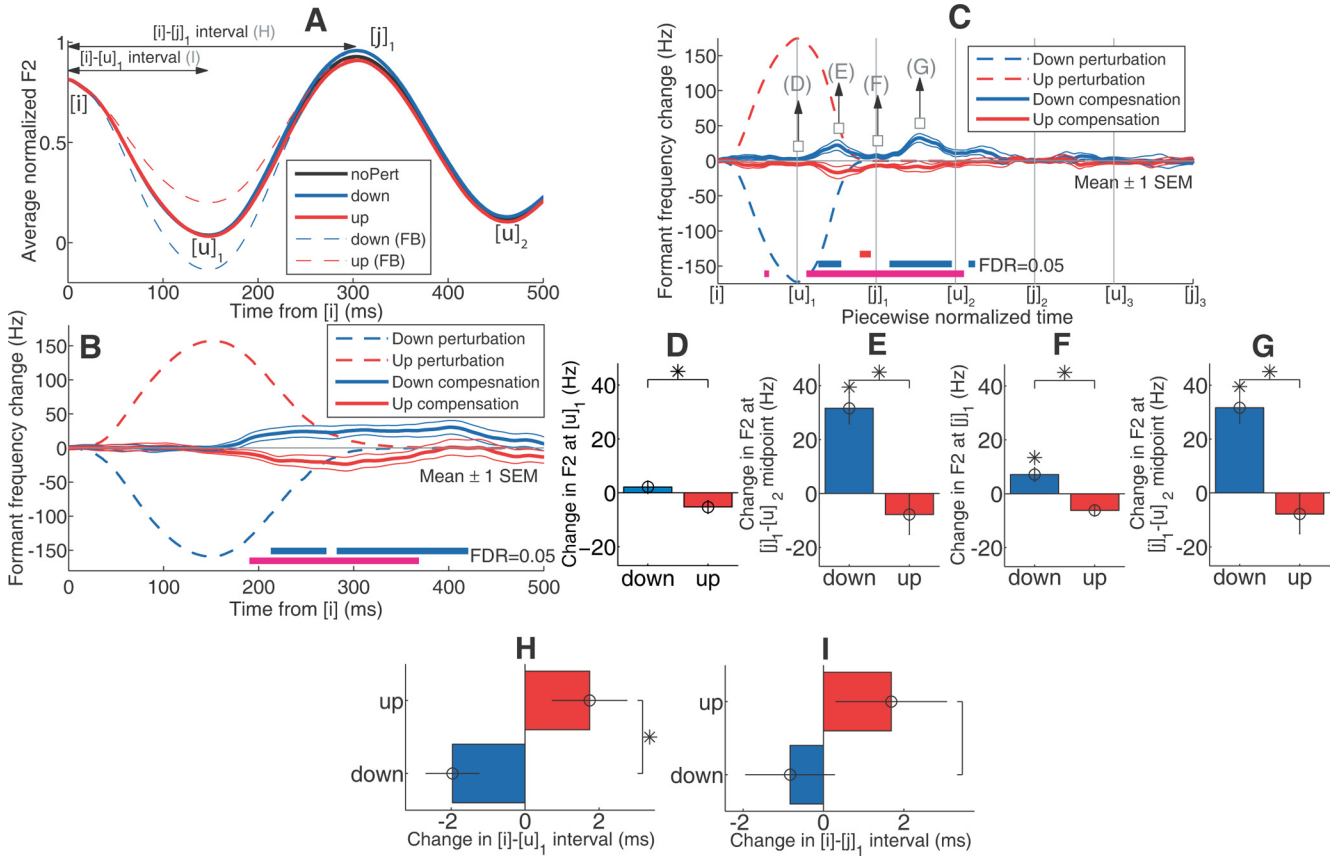

Figure 3. Articulatory compensations under the spatial (Down and Up) perturbations. ***A***, Grand average F2 trajectories aligned at the F2 maximum in [i] of "I". The time axis shows unnormalized (real) time and includes only an early part of the utterance, from [i] to $[u]_2$. Error bands are omitted for clarity of visualization. F2 was amplitude normalized before averaging across subjects. ***B***, Average F2 trajectory changes from the noPert baseline under the Down and Up perturbations. The thin lines show mean ± 1 SEM. The time axis is the same as that of ***A***. The magenta bars indicate the intervals of significant difference between the responses to Down and Up perturbation under paired *t* tests with a statistical threshold of FDR = 0.05. The blue bar shows the comparison between the Down and noPert conditions (FDR = 0.05). ***C***, Average F2 magnitude changes shown on the segment-normalized time axis (for details on the time normalization procedure, see Results). The meaning of the magenta and blue bars are the same as in ***B***; the red bar indicates the comparison between the Up and noPert conditions (FDR = 0.05). The arrows signify the correspondence with the data shown in ***D–G***, which show spatial changes under the perturbations. ***D***, Change in the value of F2 at the minimum in $[u]_1$. ***E***, Change in value of F2 at the temporal midpoint between the F2 minimum in $[u]_1$ and the F2 maximum in $[j]_1$. ***F***, Change at the F2 maximum in $[j]_1$. ***G***, Change at the midpoint between the F2 maximum in $[j]_1$ and the F2 minimum in $[u]_2$ in "you". Error bars indicate ±1 SEM. ***H***, ***I***, Timing changes under the perturbations. ***H***, Change in the [i]–$[u]_1$ time interval. ***I***, Change in the [i]–$[j]_1$ interval. Asterisks, Significant difference at $p < 0.05$ (*post hoc* Tukey's HSD following RM-ANOVA). FB, Feedback.

Perturbations to the trajectory of the first formant (F1) were done in a similar manner. The time-warping function $W(\cdot)$ took different forms for the Decel and Accel perturbations. The green curve in Figure 2*B* shows the time-delaying warping used in the Decel perturbation; the orange curve in the same panel shows the time-advancing function used for the Accel perturbation. The time-warping in the Accel perturbation was noncausal and hence required predictions of future F1 and F2 values. This prediction was achieved by using the average F1 and F2 trajectories from the focus intervals in the previous trials. Due to the naturally occurring trial-to-trial variation in the magnitude of the F2 minimum, a certain amount of mismatch in the value of the F2 minimum between the auditory feedback and the production were inevitable in the Accel perturbation. However, these mismatches were relatively small. The matching error for the F2 minimum was -2.65 ± 3.06 Hz for the 22 subjects, which was not significantly different from zero ($t_{21} = -0.873$, $p = 0.39$).

Data analysis. The experimenter, blinded from the perturbation status of the trials, screened all trials and excluded those containing gross formant-tracking errors from further analysis. Trials excluded due to gross formant-tracking errors amounted to 4.9% of all trials in Experiment 1 and 2.4% in Experiment 2. The formant tracks were then smoothed with a 17.3 ms Hamming window. An automated procedure was used to extract F2 values at the local extrema (peaks and valleys) of the F2 trajectory and at time midpoints between adjacent pairs of extrema. To obtain measures of articulatory timing, the time intervals between the F2 extrema were extracted automatically.

Statistical analysis involved repeated-measures ANOVA (RM-ANOVA) with subjects treated as a random factor. For each subject, each spatiotemporal measure of the F2 trajectory was averaged across all trials of the same perturbation type. The within-subject factor, namely perturbation type, took the values of noPert, Down, and Up in Experiment 1 and noPert, Accel, and Decel in Experiment 2. Correction for violation of the sphericity assumption of RM-ANOVA was performed with the Huynh–Feldt procedure. *Post hoc* comparisons with control for family-wise errors were conducted using Tukey's honestly significant differences (HSD) test.

Results

Experiment 1: Responses to spatial perturbation

In Experiment 1, 25 subjects produced the multisyllabic utterance "I owe you a yo-yo" under three different auditory perturbation conditions: (1) the baseline (noPert) condition, which involved no perturbation of auditory feedback; (2) the Down perturbation, which exaggerated the downward sweep of F2 during the word "owe"; and (3) the Up perturbation, which diminished the same downward sweep of F2 (for details, see Materials and Methods, above). The detailed timing and magnitudes of the local extrema of F2 under the noPert baseline can be found in Table 1. In Figure 3*A*, the dashed curves show the average F2 trajectories in the perturbed auditory feedback conditions; the

solid curves show the average F2 trajectories produced by the subjects under the three feedback conditions. It can be seen that these three solid curves mostly overlap from the onset to the middle of the focus interval, which is not surprising given the latency (~120–200 ms) involved in online feedback-based articulatory adjustments (Donath et al., 2002; Xu et al., 2004; Purcell and Munhall, 2006; Tourville et al., 2008). However, shortly after the F2 turning point in the syllable "owe," ~160 ms after the onset of the perturbation, the three curves begin to show a systematic pattern of divergence. Compared with the noPert (baseline) production, the average F2 trajectory produced under the Down perturbation showed increased values of F2 and the average F2 trajectory under the Up perturbation showed decreased F2 within the same time frame. The compensatory changes in the produced values of F2 can be seen more clearly in the solid curves in Figure 3*B*, which show the average change of F2 from the noPert baseline under the two types of perturbations. These compensatory changes in the magnitude of F2 were in the directions opposite to the directions of the auditory perturbation and lasted into the syllable [ju] ("you") after the end of the perturbation.

The compensatory responses reached statistical significance when the data were averaged along the unnormalized time axis and responses under the Down perturbation were compared with the noPert baseline (Fig. 3*B*, blue horizontal bar) or when the Down and Up responses were compared with each other (Fig. 3*B*, magenta horizontal bar). False discovery rate (FDR) (Benjamini and Hochberg, 1995) was used to correct for multiple comparisons. Due to slightly smaller average compensatory magnitude and greater intersubject dispersion, the compensatory response under the Up perturbation did not reach statistical significance under the corrected statistical threshold.

Figure 3*B* visualizes the F2 compensations in unnormalized (real) time. The unnormalized time axis is suitable for a first-pass examination of the data and for estimating the latency of compensation, but it suffers from two shortcomings: (1) it does not correct for the misalignment in time of the F2 extrema across trials and subjects, which may lead to unwanted smoothing of the pattern of compensation; and (2) it intermingles the F2 changes due to timing and magnitude (spatial) adjustments. To isolate spatial adjustments from timing adjustments, the time axis was normalized in a piecewise linear fashion (Fig. 3*C*). The F2 trajectories from individual trials were anchored at the set of F2 extremum landmarks ([i], $[u]_1$, $[j]_1$, $[u]_2$, $[j]_2$, $[u]_3$, and $[j]_3$; Table 1); the F2 trajectories between adjacent landmarks were computed through linear interpolation of time. Two hundred fifty uniformly spaced interpolation points were used between each pair of adjacent landmarks. This piecewise normalization isolates compensatory corrections in the magnitude of F2 from the adjustment of the timing of the F2 extremum landmarks.

The difference between the Down and Up conditions was statistically significant within a time interval between $[u]_1$ and $[u]_2$ (FDR = 0.05; Fig. 3*C*, magenta horizontal bar). Additionally, the comparisons of the individual perturbation conditions (Down and Up) with the noPert baseline reach corrected levels of significance (Fig. 3*C*, blue and red horizontal bar, respectively). Including the gradual buildup to the significant differences and the subsequent decay, the magnitude compensation spanned a longer time interval, from $[u]_1$ to $[j]_2$. The largest F2 magnitude adjustments are seen near the temporal midpoints between $[u]_1$ and $[j]_1$ and between $[j]_1$ and $[u]_2$. Interestingly, the compensation magnitude shows a dip near the $[j]_1$, an F2 maximum (Fig. 3*C*s, arrow F). The reason for this decreased F2 compensation magnitude around the semivowel is unclear, but may be related to a nonlinear saturation relation between articulatory position and formant frequency for this phoneme (Stevens, 1989).

When the F2 changes were analyzed at individual landmark points, significant compensatory changes were again observed. These landmarks included the F2 minimum at $[u]_1$, the temporal midpoint between $[u]_1$ and $[j]_1$, the F2 maximum at $[j]_1$, and the temporal midpoint between $[j]_1$ and $[u]_2$ (Fig. 3*D–G*). At each of these landmarks, RM-ANOVA indicated a significant main effect by perturbation condition (noPert, Down, and Up; $F_{(2,58)} = 4.09$, 11.4, 12.7, and 16.3; $p < 0.025$, 1×10^{-4}, 5×10^{-5}, 1×10^{-6} for the four above-mentioned landmarks, respectively). Pairwise Tukey's HSD comparisons between the Down and Up conditions reached significance for all three landmarks as well ($p < 0.05$ corrected for all landmarks). The ratio between the peak magnitudes of the compensatory response (Fig. 3*C*, thick solid curves) and the peak magnitudes of the auditory perturbation (Fig. 3*C*, dashed curves) was 18.9% for the Down perturbation and 9.7% for the Up perturbation. The magnitudes of the compensatory F2 adjustments are slightly larger under the Down perturbation than under the Up perturbation. This asymmetric pattern of compensation may be due to a greater need to avoid a predicted undershooting of the F2 target at the semivowel $[j]_1$ than to prevent a predicted overshooting, since the semivowel $[j]_1$ is associated with a local maximum of F2 that is reached from below. Despite the significance of these compensatory responses on the group level, there was considerable variability across trials and subjects. For example, for the landmark $[j]_1$, 20 of the 30 subjects showed trends consistent with the group average under the Down perturbations and 22 of the 30 showed trends consistent with the group average under the Up perturbation. This relatively high level of variability is consistent with previous findings based on real-time manipulation of formant feedback (Purcell and Munhall, 2006; Tourville et al., 2008).

In addition to these changes in the magnitude of F2, which reflected feedback-based control of the spatial parameters of articulation, we also observed significant changes in the timing measures of the F2 trajectory under the auditory perturbations. The [i]–$[u]_1$ interval, namely the interval between the F2 maximum at [i] and the F2 minimum at $[u]_1$, was affected significantly by the perturbation condition ($F_{(2,58)} = 6.6$, $p < 0.005$) and was significantly different between the Down and Up conditions ($p < 0.05$ corrected, *post hoc* Tukey's HSD). On average, this interval shortened under the Down perturbations and lengthened under the Up perturbation (Fig. 3*H*). If the F2 minimum at $[u]_1$ is defined as the end time of the syllable "owe," this observation indicates that the Down and Up perturbations led to an earlier- and later-than-baseline termination of this syllable, respectively. In other words, these perturbations altered the articulatory timing within this syllable. In comparison, the [i]–$[j]_1$ interval, namely the interval between [i] and $[j]_1$, exhibited a similar but nonsignificant trend of change ($F_{(2,58)} = 1.33$, $p > 0.25$; Fig. 3*I*). Therefore, if the F2 maximum at $[j]_1$ is regarded as the onset of the syllable [ju], it can be seen that the Down and Up perturbations did not significantly alter the onset timing of the following syllable (i.e., between-syllable timing).

After the experiment, subjects were questioned about whether they were aware of any distortions of the auditory feedback during the experiment. Apart from the higher-than-normal loudness and the differences between hearing one's own voices through natural auditory feedback and through playback or recordings, none of the subjects reported being aware of any deviations of the auditory feedback from the natural pattern.

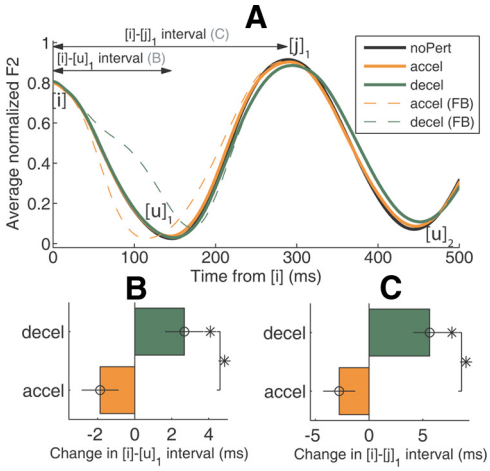

Figure 4. Articulatory adjustments under the temporal (Accel and Decel) perturbations. ***A***, Grand average (across trials and subjects) of F2 trajectories aligned at the F2 maximum at [i]. The format is the same as Figure 3*A*. The solid curves show production; the dashed curves show auditory feedback. The magnitude of the F2 at the $[u]_1$ minimum under the Decel perturbation (dashed green curve) is apparently altered from the value in the production because the timing of the $[u]_1$ minimum varies across different trials and different subjects. In individual trials, the F2 magnitudes at this minimum were always preserved by the Decel perturbation (see ***B***). ***B***, ***C***, Articulatory timing changes under the perturbations. ***B***, Change in the [i]–$[u]_1$ interval (error bars are +1 SEM). ***C***, Change in the interval between the [i]–$[j]_1$. Asterisks, Significant difference at $p < 0.05$ (*post hoc* Tukey's HSD following RM-ANOVA). FB, Feedback.

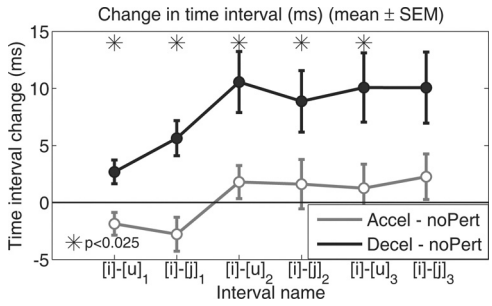

Figure 5. Changes in articulatory timing beyond the vicinity of the focus interval. Changes in the timing of the six major F2 landmarks ($[u]_1$, $[j]_1$, $[u]_2$, $[j]_2$, $[u]_3$, and $[j]_3$; see Table 1) under the Accel and Decel perturbations. The filled symbols represent significant difference from the baseline (*t* test, $p < 0.025$); the asterisks indicate significant difference between the Accel and Decel conditions (paired *t* test, $p < 0.025$).

Experiment 2: Articulatory timing adjustments under the temporal perturbations

Experiment 1 provided evidence for the involvement of auditory feedback in the online feedback-based guidance of the spatial aspect of multisyllabic articulation. As for the role of auditory feedback in controlling syllable timing, such a role was observed only in the control of within-syllable timing (Fig. 3*H*), and not in the control of between-onset timing (Fig. 3*I*). There are two alternative explanations for this pattern: (1) the syllable onset times may be completely preprogrammed, so that changes in auditory or other sensory feedback cannot affect the syllable-onset times; and (2) auditory feedback is used by the speech motor system in the online control of syllable timing, but the Down and Up perturbations used in Experiment 1 are not suitable types of perturbation to demonstrate such a role of auditory feedback.

To distinguish between these two possibilities, we used two novel types of perturbations of F2 trajectories, namely Accel and Decel temporal perturbations. Unlike the spatial perturbations used in Experiment 1, these temporal perturbations alter the timing of the F2 minimum associated with $[u]_1$. We hypothesized that with these new perturbations, significant changes in the subjects' articulatory timing would be observed, which would support a role of auditory feedback in the online control of both within-syllable and between-syllable timing.

The baseline values of the time intervals can be found in Table 1. Unlike in Experiment 1, no significant change in the magnitude of F2 was observed in response to the Accel and Decel perturbations (data not shown). However, in the temporal domain, the subjects' articulation showed an asymmetric pattern of temporal changes under the Accel and Decel perturbations. Significant articulatory timing changes were observed only under the Decel perturbation, which resulted in increases in both measured intervals. This can be seen from the slightly delayed F2 minimum at $[u]_1$ and F2 maximum at $[j]_1$ in the average Decel curve compared with those in the average noPert curve (Fig. 4*A*). As Figure 4, *B* and *C*, shows, the changes in the [i]–$[u]_1$ and [i]–$[j]_1$ intervals were quite small under the Accel perturbation, but were much greater and statistically significant under the Decel perturbation. The main effect of perturbation condition was significant for both intervals ([i]–$[u]_1$: $F_{(2,42)} = 9.08$, $p < 0.001$; [i]–$[j]_1$: $F_{(2,42)} = 13.7$, $p < 0.0001$); the changes of both intervals under the Decel perturbation from the noPert baseline were statistically significant (Fig. 4*B*,*C*). On the individual-subject level, 16 of the 22 subjects showed timing-correction trends consistent with the group average.

These temporal adjustments were qualitatively different from the spatial compensation observed in Experiment 1. The timing adjustments in this experiment were in the same direction as the temporal perturbations in the auditory feedback, whereas the spatial corrections in Experiment 1 opposed the feedback perturbation. Across the 22 subjects in Experiment 2, the ratio between the change in the [i]–$[u]_1$ interval in the subjects' production under the Decel perturbation and the perturbation of that interval in the auditory feedback was 12.6 ± 4.8% (mean ± 1 SEM). The change in the [i]–$[j]_1$-produced interval amounted to 26.1 ± 6.6% of the perturbation of the [i]–$[u]_1$ interval in the auditory feedback. These ratios of temporal adjustments were somewhat greater than the ratios of compensation under spatial perturbation observed in Experiment 1 and in previous studies that concentrated on quasistatic articulatory gestures (Purcell and Munhall, 2006; Tourville et al., 2008).

In addition to the effects on the [i]–$[u]_1$ and [i]–$[j]_1$ intervals, which were relatively close in time to the perturbation, the Decel perturbation also caused timing alterations in later parts of the utterance. As Figure 5 shows, the timing of the six major F2 landmarks (the minima of $[u]_1$, $[u]_2$, and $[u]_3$; and the maxima of $[j]_1$, $[j]_2$, and $[j]_3$) all showed significant lengthening under the Decel perturbation. These results indicate that, although the manipulation of auditory feedback was applied locally on an early part of the sentence, the Decel had global effects on syllable timing within this utterance. This timing change beyond the perturbed section of the utterance was a consequence of the delaying in the earlier syllables and a lack of subsequent efforts of the speech motor system to catch up (for implications, see Discussion, below). By contrast, the Accel perturbation caused no significant change in any of the three time intervals. After the completion of the session, subjects were asked whether they were aware of any distortion of the auditory feedback. Six of the 22 subjects (27%, higher compared with the 0% ratio in Experiment 1) reported becoming aware of the temporal distortions during the experiment. The words they used to describe their subjective perceptions of the perturbations included "echo," "out of sync," and "garbled." However, there was no evidence that these six subjects'

showed timing adjustment responses that were different from the other subjects.

Discussion

We performed two experiments using perturbations to speakers' auditory feedback of formant trajectories during the articulation of a multisyllabic utterance. From perturbing and measuring the trajectory of F2 produced by subjects under two types of perturbations, we observed significant and specific acoustic adjustments (which reflect articulatory adjustments) in response to these perturbations. To our knowledge, this is the first study to provide evidence indicating that the speech motor system uses auditory feedback to fine-tune spatiotemporal parameters of multisyllabic articulation in an online, moment-by-moment basis during multisyllabic articulation and to characterize the spatiotemporal details of this online feedback-based control.

The effects of noise masking (Van Summers et al., 1988) and delayed auditory feedback (DAF) (Zimmermann et al., 1988) on temporal parameters of connected speech have long been known. Both manipulations lead to slowing down of speaking rate; DAF can lead to breakdowns of speech fluency. However, arguments against interpreting those data as supporting a role of auditory feedback in multisyllabic articulation have been based mainly on the unnaturalness of the noise-masking and DAF conditions, which can generate results that may not reflect the control strategy used under normal (unperturbed) speaking conditions (Lane and Tranel, 1971; Borden, 1979). The perturbations used in the current study were more natural and subliminal compared with the readily perceived traditional manipulations of auditory feedback. Most subjects in the current study reported being unaware of the perturbations. Therefore, the patterns of compensation observed under the perturbations of this study can be more readily interpreted as reflecting mechanisms used in unperturbed speech production.

Experiment 1: Responses to spatial perturbations

The compensatory adjustments in the magnitude of F2 observed in Experiment 1 are qualitatively similar to the previously observed online compensation during the monophthong [ɛ] (Purcell and Munhall, 2006; Tourville et al., 2008), timing-varying vowels (Cai et al., 2010), and pitch production (Burnett et al., 1998; Donath et al., 2002). However, since the compensatory responses in the current study were observed during multisyllabic articulation, they indicate that the role of auditory feedback in online articulatory control extends beyond quasistatic or single time-varying gestures, and to the control of articulatory trajectories that connect phonemes in a sequence beyond phonemic and syllabic boundaries.

The observed patterns of change are consistent with a Smith-predictor control system that uses internal forward models (Miall and Wolpert, 1996; Wolpert et al., 1998; Kawato, 1999) and detects mismatches between expected and actual sensory feedback to generate ongoing commands to the vocal tract. These forward models integrate sensory feedback with motor efference copies to predict the consequences of the motor programs that are about to be issued (Hickok et al., 2011). These predictions are compared with the auditory targets for the phonemes to be produced (Guenther et al., 1998). If a mismatch arises between the predicted feedback and the auditory target, the control system will modify the motor programs issuing them to the articulators, so as to preemptively minimize the errors. Tourville et al. (2008) observed that the bilateral posterior superior temporal cortex, right motor and premotor cortices, and inferior cerebellum are involved in the online auditory feedback-based control of a static articulatory gesture. We postulate that the online control of multisyllabic articulation involves a similar neural substrate, possibly with the additional role played by cerebellum in internal modeling and state estimation (Miall et al., 2007), which are necessary for forming sensory expectations during sequential movements.

The magnitude of the compensatory adjustment in the produced F2 was ~11% of the magnitude of the perturbation in the auditory feedback. This ratio of compensation appears to be larger than the ratio of compensation observed in the prior studies of the monophthong [ɛ], which was shown to be ~3–6% at 250 ms after perturbation onset (Purcell and Munhall, 2006; Tourville et al., 2008). This result may reflect a greater role of auditory feedback during phoneme-to-phoneme transitions than during within-phoneme gesture stabilization, and appear to be consistent with the finding of greater compensations to perturbations of pitch feedback in time-varying pitch sequences than in quasistatic (repeating) ones (Xu et al., 2004; Chen et al., 2007). Therefore, there seems to be converging evidence for a greater role of auditory feedback-based control during the production of sequential, time-varying gestures than during quasistatic articulatory or phonatory gestures, perhaps due to the fact that time-varying gestures are more natural aspects of normal speech than the quasistatic gestures.

Experiment 2: Response to temporal perturbations

The temporal adjustments observed under the Accel and Decel perturbations of Experiment 2 altered the timing of the local F2 minimum that corresponds to $[u]_1$ in the word "owe" in the auditory feedback. One type of perturbation, Decel, led to not only a significant lengthening of the syllable [ou] ("owe") in the subjects' production, but also delayed initiation of the following syllable [ju] ("you"). These temporal corrections accounted for considerable fractions (~14–27%) of the timing perturbations in the auditory feedback. In addition, the timing of the syllables subsequent to the cessation of the Decel perturbation was also altered. These findings argue against the notion that the syllable timing in an utterance is immutable and inherent to the preplanned speech motor program (cf. Fowler, 1980), which is implicitly embodied by the task dynamic model of speech production (Saltzman and Munhall, 1989; Saltzman et al., 2006). Contradictory to this concept of a timing score that determines the timing of syllables, our findings provide evidence supporting the notion that articulatory timing can be adjusted dynamically as the sensorimotor process of articulation unfolds. In particular, the speech motor system processes the auditory feedback from earlier segments of an utterance in a way that allows for adjustment of articulatory timing in ensuing parts of the utterance. Future studies are required to test the generality of the current findings to other phoneme sequences, especially to feedback-based timing adjustments during consonants other than semivowels. In addition, examining the effects of speaking rate on timing control may elucidate whether the timing adjustments are constrained by phonemic or syllabic boundaries or by neural latencies of the response. The neural substrates of the feedback-based timing adjustments may include the basal ganglia and cerebellum, which both have been shown to play roles in speech motor timing (Wildgruber et al., 2001; Ackermann, 2008).

The response to these temporal perturbations showed an asymmetric pattern: the Decel perturbation led to significant delays in the termination of the perturbed syllable and the initiation of the following syllables; in comparison, the Accel perturbation caused much smaller timing adjustments. This asymmetric pattern is consistent with the observation by Perkell et al. (2007): whereas sudden loss of auditory feedback (by switching off the cochlear implants of implant recipients) during production of a

vowel led to significant lengthening of the duration of the vowel, sudden restoration of auditory feedback (by switching on the implants) caused no significant changes in vowel duration. The sudden loss of auditory feedback in Perkell et al. (2007) is somewhat analogous to the Decel perturbation in the current study in that they both involve a temporarily belated arrival of expected auditory feedback pattern. A recent study (Mochida et al., 2010) also observed asymmetric compensation to temporal manipulations of auditory feedback, but they observed significantly earlier-than-baseline initiation of syllables in response to auditory feedback advanced in time and no significant change in production timing when auditory feedback was delayed. Mochida and colleagues (2010) used a nonsense stimulus utterance consisting of a single repeated syllable spoken under external rhythmic pacing, whereas the current study used a semantically valid multisyllabic utterance with variegated syllables produced with natural speech timing. These methodological differences may underlie the different patterns in timing adjustment patterns found in the two studies.

The current findings demonstrate that the normal process of speech motor control makes use of auditory feedback to adjust articulatory processes, with the aim of minimizing the amount of error in reaching spatial auditory goal regions (Guenther et al., 1998; Guenther, 2006) for successive syllables. Previous studies based on mechanical perturbation of the lips and jaw during speech demonstrated short-latency, task-specific compensations to mechanical perturbations of the articulators (Gracco and Abbs, 1989; Munhall et al., 1994). When viewed in light of those previous results, the results of the current study indicate that the speech motor system makes use of both somatosensory and auditory feedback to control articulatory movements online.

In the current study, we discovered that the speech motor system monitors the spatiotemporal details of auditory feedback, extracts relevant information from them during rapid sequencing of phonemes and syllables, and then uses such information to fine-tune both the spatial and temporal parameters of the ensuing speech movements with a short latency (~150 ms). These findings raise the question of whether sensory feedback control may be operating during the production of other types of highly skilled rapid sequential movement, such as writing, typing, and performing music (Pfordresher and Palmer, 2002). As in the case of speech production, devising novel paradigms to investigate this issue offers interesting theoretical and experimental challenges.

References

Ackermann H (2008) Cerebellar contributions to speech production and speech perception: psycholinguistic and neurobiological perspectives. Trends Neurosci 31:265–272.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B 57:289–300.

Borden GJ (1979) Interpretation of research on feedback interruption in speech. Brain Lang 7:307–319.

Burnett TA, Freedland MB, Larson CR, Hain TC (1998) Voice F0 responses to manipulations in pitch feedback. J Acoust Soc Am 103:3153–3161.

Cai S, Ghosh SS, Guenther FH, Perkell JS (2010) Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong /iau/ and its pattern of generalization. J Acoust Soc Am 128:2033–2048.

Chen SH, Liu H, Xu Y, Larson CR (2007) Voice F0 responses to pitch-shifted voice feedback during English speech. J Acoust Soc Am 121:1157–1163.

Chen Z, Liu P, Jones JA, Huang D, Liu H (2011) Sex-related differences in vocal responses to pitch feedback perturbations during sustained vocalization. J Acoust Soc Am 128:EL355–EL360.

Crystal TH, House AS (1988) Segmental durations in connected speech signals: current results. J Acoust Soc Am 83:1553–1573.

Desmurget M, Grafton S (2000) Forward modeling allows feedback control for fast reaching movements. Trends Cogn Sci 4:423–431.

Donath TM, Natke U, Kalveram KT (2002) Effects of frequency-shifted auditory feedback on voice F0 contours in syllables. J Acoust Soc Am 111:357–366.

Fairbanks G (1955) Selective vocal effects of delayed auditory feedback. J Speech Hear Disord 20:333–345.

Fowler CA (1980) Coarticulation and theories of extrinsic timing. J Phonetics 8:113–133.

Gracco VL, Abbs JH (1989) Sensorimotor characteristics of speech motor sequences. Exp Brain Res 75:586–598.

Guenther FH (2006) Cortical interactions underlying the production of speech sounds. J Commun Disord 39:350–365.

Guenther FH, Hampson M, Johnson D (1998) A theoretical investigation of reference frames for the planning of speech movements. Psychol Rev 105:611–633.

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

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

Howell P, Sackin S (2002) Timing interference to speech in altered listening conditions. J Acoust Soc Am 111:2842–2852.

Kawato M (1999) Internal models for motor control and trajectory planning. Curr Opin Neurobiol 9:718–727.

Lane H, Tranel B (1971) Lombard sign and role of hearing in speech. J Speech Hear Res 14:677–709.

Miall RC, Wolpert DM (1996) Forward models for physiological motor control. Neural Netw 9:1265–1279.

Miall RC, Christensen LO, Cain O, Stanley J (2007) Disruption of state estimation in the human lateral cerebellum. PLoS Biol 5:e316.

Mochida T, Gomi H, Kashino M (2010) Rapid change in articulatory lip movement induced by preceding auditory feedback during production of bilabial plosives. PLoS One 5:e13866.

Munhall KG, Löfqvist A, Kelso JA (1994) Lip-larynx coordination in speech: effects of mechanical perturbations to the lower lip. J Acoust Soc Am 95:3605–3616.

Perkell JS, Lane H, Denny M, Matthies ML, Tiede M, Zandipour M, Vick J, Burton E (2007) Time course of speech changes in response to unanticipated short-term changes in hearing state. J Acoust Soc Am 121:2296–2311.

Pfordresher PQ, Palmer C (2002) Effects of delayed auditory feedback on timing of music performance. Psychol Res 66:71–79.

Purcell DW, Munhall KG (2006) Compensation following real-time manipulation of formants in isolated vowels. J Acoust Soc Am 119:2288–2297.

Rothauser EH, Chapman WD, Guttman N, Nordby KS, Silbiger HR, Urbanek GE, Weinstock M (1969) IEEE recommended practice for speech quality measurements. IEEE Trans Audio Electroacoust 17:225–246.

Saltzman EL, Munhall KG (1989) A dynamical approach to gestural patterning in speech production. Ecol Psychol 1:333–382.

Saltzman E, Nam H, Goldstein L, Byrd D (2006) The distinctions between state, parameter and graph dynamics in sensorimotor control and coordination. In: Progress in motor control: motor control and learning over the life span (Latash ML, Lestienne F, eds), pp 63–73. New York: Springer.

Saunders JA, Knill DC (2003) Humans use continuous visual feedback from the hand to control fast reaching movements. Exp Brain Res 152:341–352.

Stevens KN (1989) On the quantal nature of speech. J Phonetics 17:3–46.

Tourville JA, Reilly KJ, Guenther FH (2008) Neural mechanisms underlying auditory feedback control of speech. Neuroimage 39:1429–1443.

Van Summers W, Pisoni DB, Bernacki RH, Pedlow RI, Stokes MA (1988) Effects of noise on speech production: acoustic and perceptual analyses. J Acoust Soc Am 84:917–928.

Wildgruber D, Ackermann H, Grodd W (2001) Differential contributions of motor cortex, basal ganglia, and cerebellum to speech motor control: effects of syllable repetition rate evaluated by fMRI. Neuroimage 13:101–109.

Wolpert DM, Miall RC, Kawato M (1998) Internal models in the cerebellum. Trends Cogn Sci 2:338–347.

Xu Y, Larson CR, Bauer JJ, Hain TC (2004) Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. J Acoust Soc Am 116:1168–1178.

Zimmermann G, Brown C, Kelso JAS, Hurtig R, Forrest K (1988) The association between acoustic and articulatory events in a delayed auditory-feedback paradigm. J Phonetics 16:437–451.