Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception

Virgilio M. Villacorta[a)]
Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology, Room 36-591, 50 Vassar Street, Cambridge, Massachusetts 02139

Joseph S. Perkell[b)]
Speech Communication Group, Research Laboratory of Electronics, and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 50 Vassar Street, Cambridge, Massachusetts 02139; and Department of Cognitive and Neural Systems, Boston University, Boston, Massachusetts 02215

Frank H. Guenther
Department of Cognitive and Neural Systems, Boston University, Boston, Massachusetts 02215 and Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology, Room 36-591, 50 Vassar Street, Cambridge, Massachusetts 02139

(Received 18 January 2007; revised 25 July 2007; accepted 30 July 2007)

The role of auditory feedback in speech motor control was explored in three related experiments. Experiment 1 investigated auditory sensorimotor adaptation: the process by which speakers alter their speech production to compensate for perturbations of auditory feedback. When the first formant frequency (F1) was shifted in the feedback heard by subjects as they produced vowels in consonant-vowel-consonant (CVC) words, the subjects' vowels demonstrated compensatory formant shifts that were maintained when auditory feedback was subsequently masked by noise—evidence of adaptation. Experiment 2 investigated auditory discrimination of synthetic vowel stimuli differing in F1 frequency, using the same subjects. Those with more acute F1 discrimination had compensated more to F1 perturbation. Experiment 3 consisted of simulations with the directions into velocities of articulators model of speech motor planning, which showed that the model can account for key aspects of compensation. In the model, movement goals for vowels are regions in auditory space; perturbation of auditory feedback invokes auditory feedback control mechanisms that correct for the perturbation, which in turn causes updating of feedforward commands to incorporate these corrections. The relation between speaker acuity and amount of compensation to auditory perturbation is mediated by the size of speakers' auditory goal regions, with more acute speakers having smaller goal regions.
[DOI: 10.1121/1.2773966]

PACS number(s): 43.70.Mn, 43.70.Bk, 43.70.Fq, 43.71.Es [BHS] Pages: 2306–2319

I. INTRODUCTION

The purpose of this study is to investigate the role of sensory feedback in the motor planning of speech. Specifically, it focuses on speech sensorimotor adaptation (SA), which is an alteration of the performance of a motor task that results from the modification of sensory feedback. Such alterations can consist of "compensation"—a response to a feedback perturbation that is in the direction opposite to the perturbation, and additionally, "adaptation"—compensatory responses that persist when feedback is blocked (e.g., by masking of auditory feedback with noise) or when the perturbation is removed.

Psychophysical experiments that present human subjects with altered sensory environments have provided insight about the relationship of sensory feedback to motor control in both nonspeech and speech contexts. Experiments on limb movements have demonstrated the influence of proprioceptive feedback, i.e., feedback pertaining to limb orientation and position (Blakemore *et al.*, 1998; Bhushan and Shadmehr, 1999) and visual feedback (Welch, 1978; Bedford, 1989; Wolpert, Ghahramani and Jordan, 1995). Feedback-modification studies have also been conducted on speech production, including a number of studies that have induced compensation by altering the configuration of the vocal tract in some way (Lindblom *et al.*, 1979; Abbs and Gracco, 1984; Savariaux *et al.*, 1995; Tourville *et al.*, 2004). Other experiments have demonstrated speech compensation to novel acoustic feedback, such as delayed auditory feedback (Yates, 1963) or changes in loudness (Lane and Tranel, 1971).

Shifts of the fundamental frequency (F0) of sustained vowels have been shown to cause compensatory responses, that is, F0 modification by the speaker in the direction opposite to the shift (Kawahara, 1993; Burnett *et al.*, 1998; Jones and Munhall, 2000). Compensation for F0 shifts was especially evident when introduced during the production of tonal sequences by speakers of a tonal language (Xu *et al.*, 2004). Still others have demonstrated sensorimotor adapta-

[a)]Current address: Irvine Sensors Corporation, Costa Mesa, CA 92626.
[b)]Author to whom correspondence should be addressed. Electronic mail: perkell@speech.mit.edu

tion when vowel formants were perturbed in speakers' auditory feedback in nearly real time. For example, Houde and Jordan (1998, 2002) perturbed F1 and F2 of whispered productions of the vowel /ɛ/ along the /i/–/ɑ/ axis and found compensation that persisted in the presence of masking noise (adaptation) and generalized to other vowels. Max, Wallace and Vincent (2003) shifted all vowel formants in the same direction and showed compensation that increased with larger amounts of perturbation. Purcell and Munhall (2006) demonstrated compensation and adaptation to perturbation of F1 and F2 of voiced vowel formants. They also tracked the period following the removal of the perturbation and showed that the return to base line formant values was gradual (a "wash-out" of adaptation) and was not dependent on the number of trials during which maximal perturbation was maintained.

While introducing a vowel formant perturbation that was similar to the aforementioned paradigms, the current study builds on those earlier ones in a number of ways. The study described here: (1) utilized voiced speech (allowing for the measurement of possible fundamental frequency changes), (2) utilized a subject-dependent formant perturbation that allowed for inter-subject comparison of the degree of adaptation, (3) included female as well as male subjects, (4) measured how subjects' adaptive responses evolved over time (time-course analysis), (5) investigated the possibility of correlations between perceptual acuity and degree of adaptation, and (6) conducted simulations using a neurocomputational model of speech production that could account quantitatively for the amount and time course of compensation and adaptation. Purcell and Munhall (2006) reported results using approaches 1–4, but they did not explore the relation of compensation to auditory acuity or attempt to characterize the results with a neurocomputational model. Shifting all vowel formants in the same direction (either up or down for each subject—Max *et al.*, 2003) essentially amounts to changing the perceived length of the vocal tract (e.g., shifting the formants up corresponds to shortening the vocal tract); whereas shifting a single formant can induce the percept of a more complex change in vowel articulation (by causing the produced vowel to sound like another vowel—also see Houde and Jordan, 1998, 2002; Purcell and Munhall, 2006).

The aforementioned evidence showing specific compensatory adjustments of speech parameters in response to perturbations of sensory feedback indicates that speech movements can make use of feedback control mechanisms. A neurocomputational model of the motor planning of speech that can be used to explore these effects is the DIVA[1] model (Guenther *et al.*, 1998; Guenther *et al.*, 2006). This model postulates that speech movements are planned by combining feedforward control with feedback control (cf. Kawato and Gomi, 1992) in somatosensory and auditory dimensions. The model has been shown to account for numerous properties of speech production, including aspects of speech acquisition, speaking rate effects and coarticulation (Guenther, 1995); adaptation to developmental changes in the articulatory system (Callan *et al.*, 2000); and motor equivalence in the production of American English /r/ (Nieto-Castanon *et al.*, 2005).

According to the DIVA model, during the initial period of speech acquisition, feedforward mechanisms are not yet fully developed, so feedback control plays a large role in ongoing speech. Through training, the feedforward controller gradually improves in its ability to generate appropriate movement commands for each speech sound (phoneme or syllable); eventually, it is the dominant controller in fluent adult speech. For mature speakers, the feedback controller is always operating, but it only contributes to motor commands when sensory feedback differs from sensory expectations, e.g. in the presence of perturbations such as the auditory modification of vowel formants introduced in the current study. If such a perturbation is applied repeatedly, the model predicts that feedforward commands will be re-tuned to account for the perturbation, and that abrupt removal of the perturbation will lead to a transient after effect (evidence of adaptation) in which the speaker still shows signs of this compensation even though the perturbation is no longer present. The DIVA model also predicts that auditory perception affects motor development such that speakers with better auditory acuity will have better tuned speech production; e.g., they will produce better contrasts between sounds. Consistent with this prediction, positive correlations between auditory acuity and produced contrast in speech have been observed for both vowels and consonants (Newman, 2003; Perkell *et al.*, 2004a; Perkell *et al.*, 2004b). The model further predicts that subjects with more acute auditory perception should be able to better adapt their speech to perceived auditory errors such as those introduced by F1 perturbation. The current study addresses several of these predictions.

The study comprised three experiments. The first experiment investigated auditory sensorimotor compensation and adaptation by perturbing the first formant frequency (F1) in the feedback heard by subjects as they produced vowels in CVC words. The experimental paradigm allowed us to study the time course of formant changes throughout an experimental run in vowels produced with and without masking noise. The second experiment investigated auditory acuity, measured as discrimination of synthetic vowel stimuli differing in F1 frequency, using the same subjects; this experiment was designed to determine if individuals with more acute discrimination of vowel formants also showed greater compensation to perturbations in those formants of the first experiment. The third experiment used subject-specific versions of the DIVA model of speech motor planning to simulate the subjects' performance in the first and second experiments; it was designed to determine whether the model could account quantitatively for key aspects of sensorimotor adaptation.

II. EXPERIMENT 1

This experiment was designed to test the hypothesis that human subjects utilize auditory goals in the motor planning of speech, and should modify their vowel production to compensate for acoustic perturbations in their auditory feedback. The experiment also tested the prediction that there will be adaptation: compensation that persists in the presence of masking noise and a transient after effect in which speakers

continue to show compensation for a number of trials after the perturbation is abruptly removed.

A. Real-time formant shift in vowels

A digital signal processing (DSP) algorithm was developed for shifting the first formant frequency using a Texas Instruments (TI) C6701 Evaluation Module DSP board. The algorithm utilized linear prediction coding (LPC) analysis (Markel and Gray, 1976) and a Hessenberg QR root-finding iterative algorithm (Press *et al.*, 2002) to detect the first formant (F1) in vowels. It then utilized a direct-form transpose II filter to remove the original F1, and introduced the shifted F1. This algorithm is discussed in greater detail in Appendix I and Villacorta (2006). The overall delay introduced by the digital signal processing was 18 ms, less than the 30 ms delay at which speakers notice and are disturbed by delayed feedback (Yates, 1963).

To simplify discussion of the formant shift made by the DSP board, a unit of formant shift—*perts*—is introduced here. *Perts* simply represents a multiplier of the original formant. A formant shift of 1.3 perts increased the formant to 130% of its original value (shift up), while a 0.7 perts shift decreased the formant to 70% of its original value (shift down). A pert value of 1.0 indicates that the formant was not shifted.

B. Protocol for an experimental run

The experimental run for each subject consisted of an initial calibration phase, followed by a four-phase adaptation protocol. The purpose of the calibration phase (typically 36–54 tokens in duration) was to acclimate each subject to using visual cues (target ranges and moving displays) of loudness and duration for achieving values that were needed for successful operation of the algorithm. To help assure that the subject did not hear airborne sound, insert headphones were used (see below) and the target output sound level was set at 69 dB sound pressure level (SPL) (±2 dB), significantly less than the feedback sound level of 87 dB SPL. The target vowel duration was set at 300 ms, although the actual duration could be longer due to a reaction time delay. In this phase, subjects were also questioned about the level of masking noise (87 dB SPL); as had been found in preliminary informal testing, it was determined that the level was tolerable for them and successfully prevented them from discerning their own vowel quality.

The adaptation protocol for each presentation of a token was as follows (see Fig. 1). A monitor (1 in Fig. 1) in front of the subject displayed the token (a CVC word, such as "bet") for two seconds, and also displayed the visual cues for achieving target loudness and duration. The subject spoke into a Sony ECM-672 directional microphone placed six inches from the lips (2). The speech signal transduced by the microphone was digitized and recorded for postexperiment analysis (3). The same speech signal was sent concurrently to the TI DSP board for the synthesis of formant-shifted speech (4). The output of the DSP board (formant-shifted speech) was sent to a feedback selector switch which determined, depending on which token was presented to the subject,

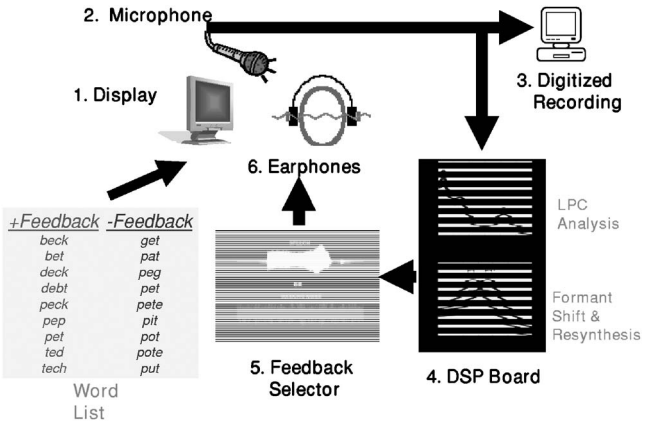

FIG. 1. Schematic diagram of the cycle that occurred during the presentation of one token during an SA experimental run. Refer to Sec. II B for a detailed description.

whether the subject heard masking noise or the perturbed speech signal (5). The appropriate signal was then presented to the subject over EarTone 3A insert earphones (Ear Auditory Systems) (6)[2]. The perturbed speech signal from the DSP board and the output signal from the selector switch were also digitized and saved for postexperimental analysis.

A total of 18 different target words ("Word List" in Fig. 1 and Table I) were used. The experiment consisted of a number of *epochs*, where each epoch contained a single repetition of each of the 18 target words. Nine of these words (*+feedback*) were presented with the subjects able to hear auditory feedback (either perturbed or unperturbed, depending on the phase of the experiment) over the earphones; all of these words contained the vowel /ɛ/ (the only vowel trained). The other nine words (*–feedback*) were presented with masking noise. Three of the *–feedback* words contained the vowel /ɛ/, one in the same phonetic context as the word presented in the *+feedback* list ("pet") and two in different phonetic contexts ("get" and "peg"). The other six *–feedback* words contained vowels different from the training vowel. The order of the *+feedback* tokens and *–feedback* tokens was randomized from epoch to epoch; however, all of the +feedback tokens were always presented before the *–feedback* tokens within an epoch.

For each subject, the adaptation protocol comprised four phases: *base line*, *ramp*, *full perturbation* and *postperturbation* (schematized in Fig. 2). Each phase consisted of a fixed number of epochs. The base line phase consisted of the first 15 epochs, and was performed with the feedback set at

TABLE I. Word list for the SA experiment.

+Feedback	–Feedback
beck	get
bet	pat
deck	peg
debt	pet
peck	pete
pep	pit
pet	pot
ted	pote
tech	put

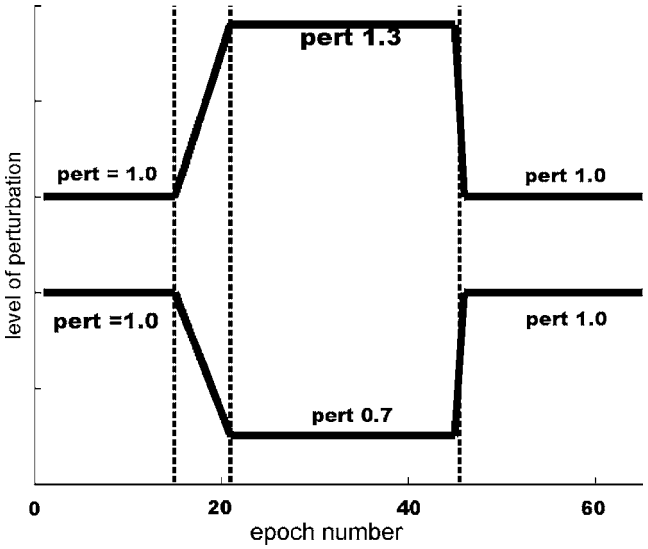

FIG. 2. Diagram of the level of F1 perturbation presented during one experimental session, as a function of epoch number (where an epoch consists of one repetition of each of the 18 words in the corpus). The 65 epochs of an experimental session are divided into four phases (demarcated by dashed vertical lines). From left to right, these phases are *base line* (epochs 1–15), *ramp* (epochs 16–20), *full perturbation* (epochs 21–45), and *postperturbation* (epochs 46–65). The protocols for two subject groups are shown: those undergoing an upward F1 shift (upper line) and those undergoing a downward F1 shift (lower line).

1.0 pert (no formant shift). The following ramp phase (epochs 16–20) was used to gradually introduce the formant shift by changing the pert level by 0.05 pert per epoch. Depending on the subject group (*shift up* or *shift down*—see below), during the full perturbation phase (epochs 21–45), the speech feedback had either a 1.3 pert shift or a 0.7 pert shift. During the entire postperturbation phase (epochs 46–65), the feedback was again set at 1.0 pert (no shift); this phase allowed for the measurement of the persistence of any adaptation learned during the full-perturbation phase. An entire experiment for one subject consisted of 65 epochs, comprising a total of 1170 tokens; the experiment lasted approximately 90–120 min.

C. Subject selection criteria and description

Subjects were 20 adult native speakers of North American English with no reported impairment of hearing or speech. Five females and five males were run with an upward F1 shift (*shift-up* subjects); another five females and five males were run with a downward F1 shift (*shift-down* subjects). The subjects had an age range from 18 to 44 with a median age of 21. Informed consent was obtained from all subjects.

D. Postexperiment spectral analysis of tokens

Following the experiment, a spectral analysis was performed on the speech signals that had been digitized directly from the microphone. Each recorded token (sampled at 16 kHz) was labeled manually at the beginning and end of the vowel on the sound-pressure wave form; then the first two formants were extracted utilizing an automated algorithm designed to minimize the occurrence of missing or spurious values. Formants were derived from an LPC spectrum taken over a sliding 30 ms window. The spectrum was measured repeatedly between 10% and 90% of the delimited vowel interval in 5% increments, and the mean formant values over these repeated measures were recorded. The analysis for a majority of the subjects used an "optimal" LPC order determined by a heuristic method that utilizes a reflection coefficient cutoff (Vallabha and Tuller, 2002). For subjects with a large number of missing or spurious formants, the analysis was repeated using LPC orders of 14–17 inclusive.

The fundamental frequency (F0) was extracted from each token using a pitch estimator that is based on a modified autocorrelation analysis (Markel *et al.*, 1976). For some tokens, F0 appeared to be underestimated, so F0 values below 50 Hz were excluded from analysis. For all but one subject, this exclusion criterion removed less than 3% of the tokens. One subject had 44% of tokens excluded by this criterion, so that subject's data were excluded from the F0 analysis.

To allow comparison among subjects with differing base line formant frequencies and F0, especially differences related to gender, each subject's formant and F0 values were normalized to his or her mean base line values, as shown in Eq. (1) for F1.

$$\text{norm\_}F1 = \frac{F1_{\text{Hertz}}}{\text{mean}(F1)_{\text{base line phase}}}. \tag{1}$$

In order to compare changes from the base line (normalized value =1.0) to the full-pert phase among all the subjects (regardless of the direction of the F1 shift), an *adaptive response index* (ARI) was calculated as shown in Eq. (2). Larger, positive ARI values indicated greater extent of adaptation for that subject, while negative ARI values (which occurred for two of the 20 subjects) indicated that those subjects produced responses that followed the perturbation, rather than compensated for it.

$$\text{ARI} = \begin{cases} \text{mean}(\text{norm\_}F1 - 1)_{\text{full pert phase}}, & \text{if } pert = 0.7 \\ \text{mean}(1 - \text{norm\_}F1)_{\text{full pert phase}}, & \text{if } pert = 1.3 \end{cases}. \tag{2}$$

E. Results

Figure 3 shows normalized F1 (solid curves) and F2 (dashed curves) values for the +*feedback* tokens averaged across all subjects in each group.[3] Data from shift-down subjects are shown with black lines; from shift-up subjects, with gray lines. The error bars show one standard error about the mean. The figure shows that subjects compensated partially for the acoustic perturbation to which they were exposed. Shift-up subjects increased vowel F1 during the experiment (black solid line), while shift-down subjects decreased F1 (gray solid line).[4] Compared to the changes in F1, F2 changed by very small amounts.

Generally, subjects responded with only a short delay to the acoustic perturbation: the first significant change in normalized F1 occurred during the second epoch in the ramp phase (epoch 17). This compensation was retained for some time after the perturbation was turned off at epoch 45 (i.e., during the *postpert* phase), indicating that subjects had

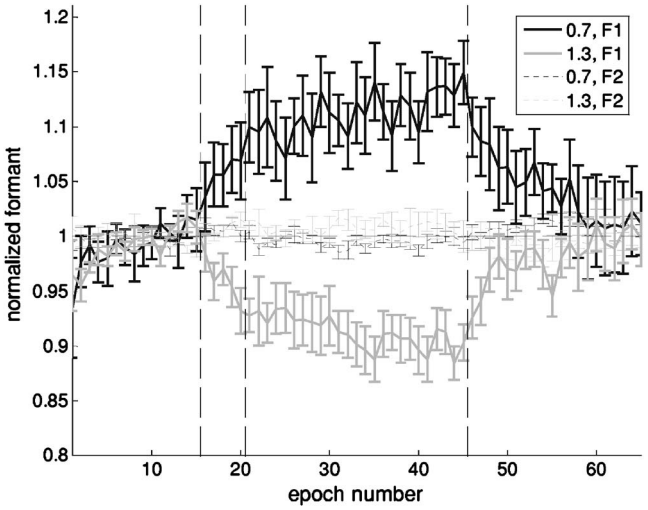

FIG. 3. Produced first and second formant frequencies, normalized to the adjusted base line, as a function of epoch number in *+feedback* words for all subjects. The upper curve corresponds to the normalized F1 for the ten subjects run on the shift-down protocol; the lower curve corresponds to the shift-up protocol. Each data point is the mean value of the nine *+feedback* words across ten subjects (five male, five female). The dashed vertical lines demarcate the phases of the protocol; the dashed horizontal line corresponds to base line values. Normalized F2 values are shown as the dashed curves, which remain close to the base line value of 1.0. The error bars depict the standard error of the mean among ten subjects.

adapted to the perturbation. Normalized F1 consistently returned to base line within the standard error after epoch 55, approximately 15–20 min into the postpert phase. This finding is consistent with those of Purcell and Munhall (2006), who also showed that recovery to base line formant frequencies was not immediate when the formant perturbation was removed.

The extent of adaptation was less than the amount required to fully compensate for the acoustic perturbation. For shift-down subjects, full compensation (i.e., the inverse of 0.7) would be represented by a normalized F1 value of 1.429; the greatest actual change for the shift-down subjects had a mean normalized value of 1.149 (i.e., approximately 35% compensation), which occurred in epoch 45. Similarly, full compensation for the shift-up subjects (1.3 pert shift) would be represented by a normalized F1 value of 0.769. Their greatest change had a mean normalized value of 0.884 (approximately 50% compensation), which occurred in epoch 44.

The *–feedback* tokens were analyzed in the same way to determine the extent to which adaptation would occur for the same vowel with auditory feedback masked (that is, without perception of the perturbed signal). As mentioned above, the word list contained tokens that were uttered with auditory feedback masked, but which contained the same vowel the subjects had heard with full perturbation (/ɛ/). The DIVA model predicts that adaptation learned for /ɛ/ with feedback perturbed should be maintained even without acoustic feedback. Indeed, in their SA study of with whispered vowels, Houde and Jordan (1998, 2002) demonstrated that such adaptation was maintained in the absence of acoustic feedback and also that it generalized to productions of the same vowel in different phonetic contexts. The current *–feedback* adaptation results for /ɛ/ are divided into two groups: *–feedback* adaptation for the *same context* token, and *–feedback* adaptation for *different context tokens*. The *same context* token—referring to the fact that this token is also contained in the *+feedback* word list—is the token "pet." The *different context* tokens are the tokens "get" and "peg," which were not present in the *+feedback*, word list.

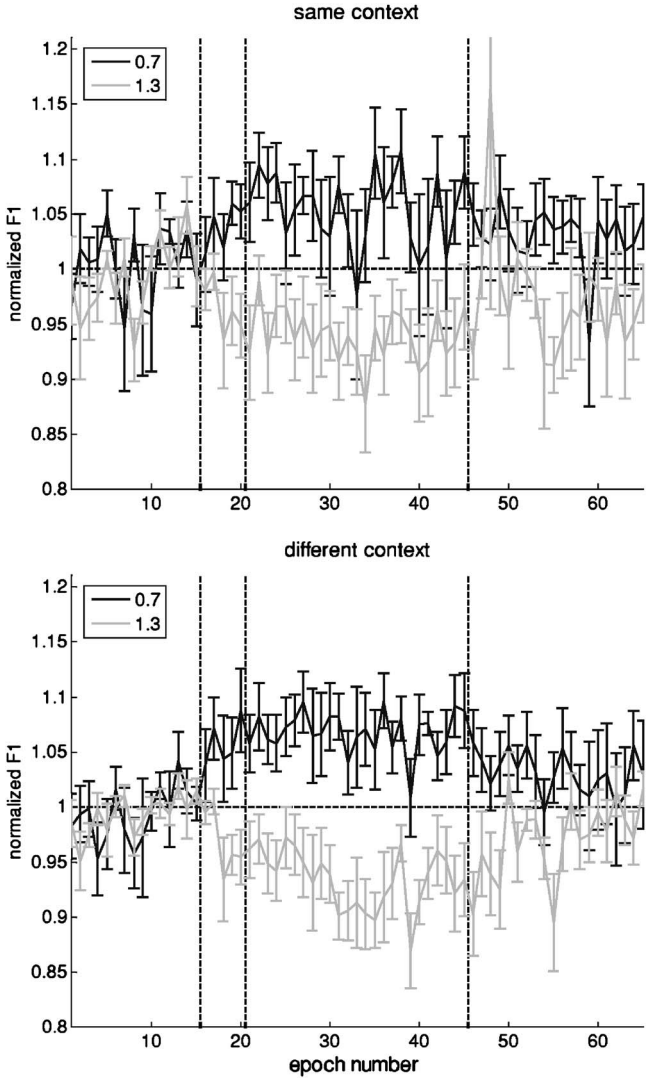

FIG. 4. Produced first formant frequency, normalized to the base line, in the *–feedback* words containing the vowel /ɛ/. The top plot shows normalized F1 for the *same context*, *–feedback* token ("pet"), while the bottom figure shows normalized F1 for the *different context*, *–feedback* tokens ("get" and "peg"). The axes, data labels and vertical markers are the same as in Fig. 3, except that normalized F2 is not shown.

Figure 4 shows that the adaptation to perturbation of *+feedback* /ɛ/ tokens does indeed occur for the *same context*, *–feedback* token. However, adaptation in the *–feedback* tokens occurred to a lesser extent than in the *+feedback* tokens (compare with Fig. 3). This finding is confirmed by comparing ARI values (Eq. (2)) between *+feedback* tokens and *–feedback* tokens. The *ARI* for the *–feedback, same context* condition was 58% of the ARI for the *+feedback* tokens, which is a significant difference ($t[198]=2.3, p<0.05$). Additionally, the ARI in the *–feedback, different context* condition was 67% of the *+feedback* (ARI) condition, which is also a significant difference ($t[218]=2.47$, $p<0.05$). While

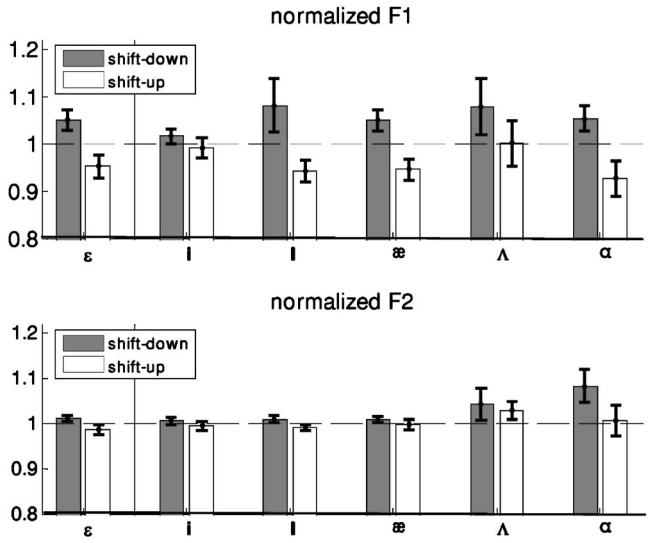

FIG. 5. Full-pert phase formants normalized to base line for all *–feedback* token vowels. Mean first formant values are shown in the upper plot; second formant values, in the lower plot. Values from shift-down subjects are represented by dark bars and from shift-up subjects, by light bars. Error bars show standard error about the mean.

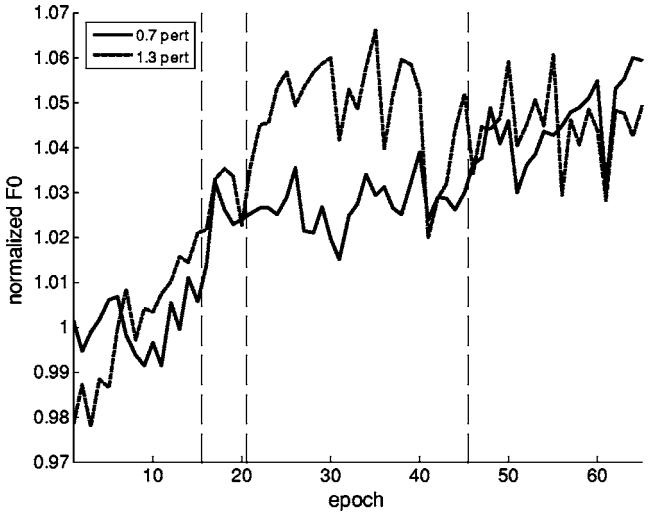

FIG. 6. Normalized F0 as a function of epoch number. To maintain consistency with Fig. 3, only *+feedback* vowels are shown. The solid line represents the mean values from the shift-down subjects; the dashed line represents the mean values from the shift-up subjects. The vertical lines demarcate the phases of the experiment.

the mean changes in F1 for the *different context* tokens appear to be greater than the F1 changes for the *same context* tokens, there were no significant differences between the two context conditions for both the shift-down and shift-up subjects.

Several *–feedback* tokens contained vowels different from the one subjects produced with feedback perturbed (/ɛ/). These tokens were included in the protocol to establish the degree to which adaptation would generalize to unperturbed vowels. The bar plots in Fig. 5 display the amount of adaptation found for the following vowels: /ɪ/ ("pit"), /i/ ("pete"), /æ/ ("pat"), /ʌ/ ("put"), and /ɑ/ ("pot").[5] The *–feedback* token /ɛ/ is also displayed for comparison. Shown are the mean F1 (upper plot) and F2 (lower plot) of these vowels, normalized with respect to each vowel's base line formant values.

For most vowels, the mean normalized F1 was significantly above the base line in shift-down subjects, and was significantly below the base line in shift-up subjects ($p<0.01$). However, the vowels /i/ and /ʌ/ did not show consistent F1 generalization. The shift-down subjects demonstrated vowel a small significant increase of F1 for /i/ ($t[249]=2.33, p<0.05$); the shift-up subjects showed a small decrease in F1 for /i/ that was not significant. For the vowel /ʌ/, the shift-down subjects demonstrated a significant upward F1 shift, but the shift-up subjects failed to demonstrate a significant decrease. (Villacorta, 2006, shows that this lack of generalization for /ʌ/ is due to the male, shift-up subgroup.) As seen in the lower plot, changes in F2 were considerably smaller in magnitude than in F1 ($p<0.05$), demonstrating formant specificity of the generalization for most of the vowels. The vowel /i/ did not show a significant F2 change for either shift-down or shift-up subjects—likely due to the fact that the F1 changes for /i/ were also relatively small. The vowel /ɑ/ did not show significantly smaller F2 changes (compared to F1 changes) in the shift-down subjects. Anomalously, the vowel /ʌ/ showed F2 increases for the shift-down as well as the shift-up subjects (possibly related to the above-mentioned outlying F1 responses of the male, shift-up subgroup).

Figure 6 shows F0 as a function of epoch number, averaged across shift-up (dashed line) and across shift-down subjects (solid line) and normalized to the mean of the base line epochs. The figure shows that both shift-down and shift-up subjects demonstrated a general trend of increasing F0 throughout the experiment. The relation between changes in F0 and F1 (factoring out the common upward trend in F0) was investigated by calculating the difference between subject F0 value and the mean F0 and the difference between subject F1 and mean F1 across all subjects at each epoch. It was found that subjects modified F0 in a direction opposite to the compensatory F1 shift they produced; this relation was highly significant ($r=-0.74$, $p<0.001$). It is possible that the duration of each utterance (300 ms), the large number of utterances produced by each subject (approximately 1170 tokens), and the overall duration of the experiment (90–120 min) all combined to cause fatigue that led to an upward drift in F0. Some support for this claim can be inferred from a similar upward F0 drift observed by Jones and Munhall (2000).[6]

Analysis of the adaptive response index values for F1 and F2 showed that, from the ramp phase through the post-pert phase, the direction of the small AR_{F2} change appears to be opposite to AR_{F1} changes. The mean AR values across all subjects for this subset of epochs (ramp phase through post-pert phase), showed a significant inverse relation between AR_{F1} and AR_{F2} ($r=-0.78$, $p<0.001$). Thus the observed changes in F0, F1 and F2 lead to the inference that the auditory space in which subjects adapt is characterized by dimensions that depend on multiple formants and F0.

III. EXPERIMENT 2

To investigate whether subjects' auditory acuity was related to the amount of their adaptation, a second experiment

was conducted to measure auditory acuity of F1 variation with the same subjects who served in Experiment 1. This experiment consisted of three parts: (1) a recording of the subject's "base" tokens, (2) an adaptive staircase discrimination task and (3) a second, more finely tuned discrimination task. The end result was a measure of each subject's auditory acuity. The use of a two-stage protocol for obtaining an accurate estimate of auditory acuity was based on prior work (Guenther *et al.*, 1999a; Guenther *et al.*, 2004).[7]

A. Participating subjects

The subjects were a subset of those who participated in Experiment 1. Seven out of the original 20 subjects were no longer available at the time Experiment 2 was conducted, so the results from the acuity experiment were based on the 13 subjects who could be recalled. Informed consent for the auditory acuity experiments was obtained from all subjects.

B. Recording of the subject's speech

Subject-specific synthetic stimuli were used for the acuity tests. For this purpose, each subject was recorded while speaking ten tokens each of the words "bet," "bit" and "bat." The recordings were conducted in a sound attenuating room using a head-mounted piezo-electric microphone (Audio-Technica, model AT803B) placed at a fixed distance of 20 cm from the speaker's lips. Elicited utterances were presented on a monitor. As in Experiment 1, the monitor also displayed cues that induced the subject to speak at a target loudness (85±2 dB SPL) and word duration (300 ms). Subjects were allowed to practice to achieve these targets. The F1 frequency for each "bet" token was measured, and the "bet" token with the median F1 value was used to determine the F1 of a *base token*. Synthetic vowels varying in F1 were generated offline using a MATLAB program that ran a formant perturbation algorithm identical to what was run on the TI DSP board.

The acuity tests were carried out in the same sound attenuating room in which the recordings were made, though not always on the same day. Subjects heard stimuli over closed-back headphones (Sennheiser EH2200), played on a computer controlled by a MATLAB script.

C. Staircase protocol for estimation of *jnd*

In an initial stage of acuity testing, a staircase protocol was used to rapidly obtain an approximate estimate of the just noticeable difference (*jnd*) in F1 for each subject. This estimate was then used to determine a narrower range of tokens for the second stage, which utilized a larger number of trials with token pairs that were chosen to fall near the subject's initial *jnd*, in order to produce a more accurate estimate of auditory acuity.

An adaptive, one-up, two-down staircase protocol was run to estimate the *jnd* for F1 around the *base token* obtained from the subject's speech recording (as illustrated in Fig. 7). In this procedure, pairs of tokens that were either the *same* or *different* from each other were presented to the subject with equal probability. The *same* pairs consisted of repetitions of the *base token*, while the *different* pairs consisted of tokens

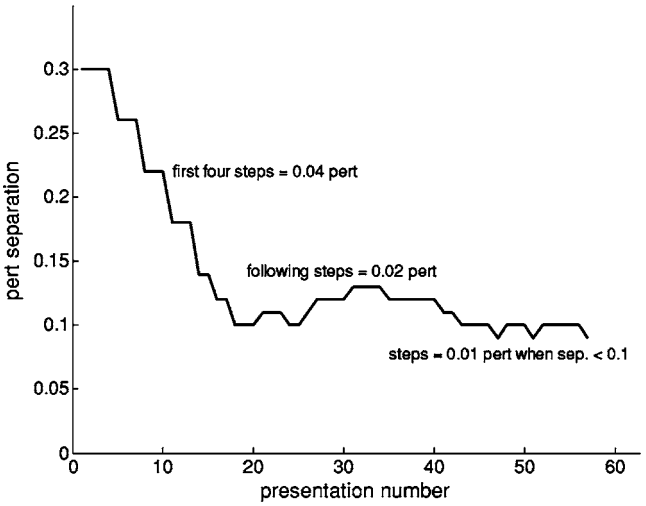

FIG. 7. Example of the adaptive procedure used to estimate *jnd*. The abscissa shows the presentation number of the given pair, and the ordinate depicts the separation of the *different* pairs in pert. The text within the figure gives conditions for changes in step size. The staircase terminated after eight reversals.

with F1 values greater or lesser than that of the *base token*, equally spaced in pert. For example, the *different* pair separated by 0.3 pert consisted of the 0.85 pert and the 1.15 pert tokens. Whenever the subject responded incorrectly to either the *same* or *different pairs*, the distance between the members in the *different* pairs increased. Whenever the subject responded correctly to two presentations of a given different pair, the distance between the members of the different pairs decreased. The separation was unchanged when the subject responded correctly to a *same*, pair presentation.[8]

D. Determining auditory acuity

A more precise protocol involving many more *same-different* judgments was then run on each subject. In the *jnd* protocol, presented tokens were either the *same* (with both tokens equal to the base token) or *different* (straddling the base token). The *different* pairs were spaced by the following multiples of the jnd_{est}: ±0.25, ±0.5, ±0.75, ±1.0 and ±1.4. The +multiple of the jnd_{est} pair (e.g., +0.25, +0.5) was always presented with the corresponding multiple (e.g., −0.25, −0.5) for a *different* pair presentation, though the order of the tokens within a pair was randomized (e.g., +0.25 followed by −0.25 or −0.25 followed by +0.25). Each unique pair (the single *same* and each of the five *different* pairs) was presented to the subject 50 times, for a total of 300 presentations per block. Subjects were given feedback consisting of the correct response to the pair just presented.

A d' score for each pair was calculated using a standard signal detection theory formula (Macmillan and Creelman, 2005) shown in Eq. (3), where z is the normal inverse function, H is the hit rate (responds *different*|different) and F is the false alarm rate (responds *different*|same). Note that all rates were calculated as a fraction of a total of 50.5 presentations (rather than 50 presentations) to avoid undefined z scores.

$$d' = z(H) - z(F) \tag{3}$$

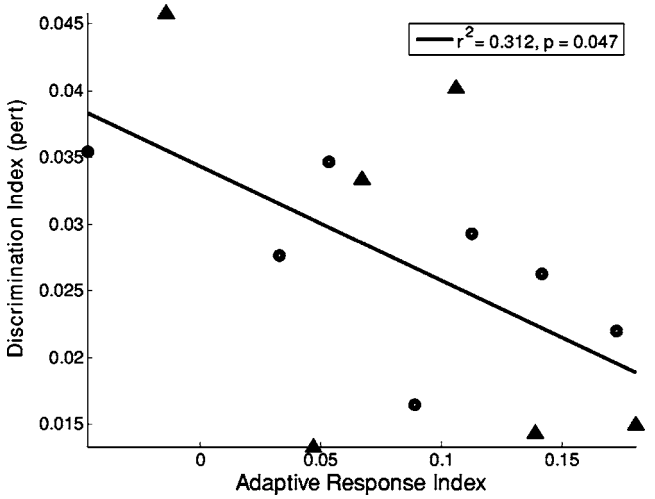

FIG. 8. The adaptive response index is correlated with the *jnd* score of the base token. The ordinate shows the *jnd* score (*Discrimination Index*), while the abscissa shows the *adaptive response index*. The open circles represent shift-down subjects, while the triangles represent shift-up subjects. Statistics for the regression line are shown in the legend.

Data consisting of d' score as a function of pair separation (in perts) were then fitted with a sigmoid function. A sigmoid function was used in this case because it is monotonic and best captures the sharp rise of d' in the sensitive region, while also capturing ceiling and floor effects observed in the data. To estimate perceptual acuity, a "discrimination index" (DI) was calculated from the sigmoid fit to the d' function. We defined the DI as the separation (in perts) that corresponds to a d' of 0.7. (A d' of 0.7 was used here because it was the maximum d' value common to all subjects run on the perceptual acuity protocol.) Note that the larger the DI, the worse the subject's acuity (i.e., the further apart two stimuli need to be for detection by the subject).

E. Results

The subjects' DIs were significantly correlated with their *adaptive response indices*, as shown in Fig. 8. This figure shows DI as a function of ARI for the shift-down subjects (open circles) and the shift-up subjects (triangles), along with a regression line. The line demonstrates the predicted trend: subjects with smaller *jnds* tend to adapt to a greater extent. The relation between *jnd* and adaptive response was significant ($r=0.56$, $p<0.047$), accounting for 31% of the variance.

It was observed that the produced F1 separation between neighboring vowels varied from subject to subject, which could have a confounding influence on the extent of adaptation measured during the SA experiment and therefore on the correlation with *jnd* values. Since the SA experiment included base line (epochs 1–15) tokens of the vowels /æ/, /ɛ/, and /ɪ/ ("pat," "pet," and "pit" used as −feedback tokens), it was possible to measure the F1 separation in neighboring vowels and subsequently control for it. Equation (4) shows how normalized vowel separation in F1 was calculated. Note that the $F1\_separation$ values are normalized by the base line F1 from the word "pet," and that only *−feedback* base line tokens were used for this measurement.

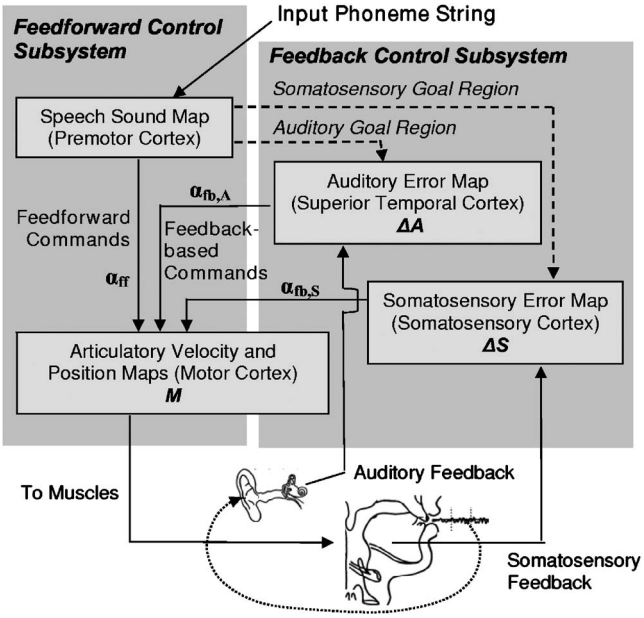

FIG. 9. A functional diagram of the DIVA model of speech motor control. The feedforward component projects from the speech sound map $\boldsymbol{P}$, and is scaled by weight $\boldsymbol{\alpha_{ff}}$. The feedback component consists of projections from the auditory ($\boldsymbol{\Delta A}$) and somatosensory ($\boldsymbol{\Delta S}$) error maps, and are scaled by weights $\boldsymbol{\alpha_{fb,A}}$ and $\boldsymbol{\alpha_{fb,S}}$, respectively. The feedforward and feedback projections are integrated by the speech motor cortex $\boldsymbol{M}$ to yield the appropriate speech motor commands, which drive the vocal-tract articulators to generate speech sounds.

$$F1\_separation_{\text{pet–pit}} = \frac{\text{pet}\_F1_{\text{median}} - \text{pit}\_F1_{\text{median}}}{\text{pet}\_F1_{\text{median}}}$$

$$F1\_separation_{\text{pat–pet}} = \frac{\text{pat}\_F1_{\text{median}} - \text{pet}\_F1_{\text{median}}}{\text{pet}\_F1_{\text{median}}}. \quad (4)$$

For a given subject, the relevant $F1\_separation$ value was the one characterizing the separation between the two neighboring vowels corresponding to the direction of perturbation used in the SA protocol. Therefore, $F1\_separation_{\text{pet–pit}}$ was used for the shift-down subjects and $F1\_separation_{\text{pat–pet}}$ was used for the shift-up subjects.

The partial correlation coefficient ($r_{x,y\|z}$) represents the correlation between two measures (DI) and ARI when controlling for *normalized* $F1\_separation$. This statistic, $r_{\text{acuity\_index ARI}\|\text{norm\_F1\_separation}}$ had a highly significant value ($r=0.79$; $p<0.001$), accounting for over 62% of the variance and indicating that smaller *jnd* values (i.e., greater perceptual acuity) are associated with larger adaptation scores.[9]

IV. EXPERIMENT 3

This experiment was designed to compare simulations using the DIVA model of speech motor planning to the human subject results from the SA and auditory acuity studies. Figure 9 shows a simplified schematic diagram of the DIVA model, indicating the relation between feedback and feedforward control of speech movements in the cerebral cortex. The model is described here briefly; it is discussed in depth in Guenther *et al.* (2006).

The *speech sound map* (hypothesized to lie in left pre-

motor cortex) projects sensory expectations associated with the current speech sound to auditory (ΔA) and somatosensory (ΔS) error cells, where these expectations (or *goals*) are compared to the actual sensory feedback. The projections of sensory expectations are learned and improve with practice. The output from the sensory error cells projects to an articulatory velocity map, resulting in the feedback-based component of the motor command; the gains $\alpha_{fb,A}$ and $\alpha_{fb,S}$ control how much each feedback source contributes to the overall motor command.

The speech sound map—aside from giving rise to the sensory expectations projecting to the sensory error cells—also projects directly to motor cortex, giving rise to a feedforward component of the motor command. By incorporating the results of previous attempts to produce the given speech sound with auditory feedback available, this motor command improves over time.

The feedforward and the two feedback components of the motor command are integrated to form the overall motor command M, which determines the desired positions of the speech articulators. The motor command M in turn drives the articulators of the vocal tract, producing the speech sound; this production provides sensory feedback to the motor control system. For use in simulations, the DIVA model's motor commands, M, are sent to an articulatory based speech synthesizer (Maeda, 1990) to produce an acoustic output.

When the model is first learning to speak (corresponding to infant babbling and early word production), the feedback component of speech control plays a large role, since the model has not yet learned feedforward commands for different speech sounds. With continued speech training, the feedforward projections from the speech sound map improve in their ability to predict the correct feedforward commands. In trained fluent (e.g., adult) speech in normal conditions, feedforward control dominates the motor command signal since the error signals resulting from the auditory and somatosensory error cells are small due to accurate feedforward commands. Alterations in auditory feedback—as introduced by the SA protocol—produce mismatches between expected and actual auditory consequences, which results in an auditory error signal. This causes the feedback control signal (specifically the auditory component) to increase and significantly influence the output motor commands. Adaptation occurs in this model as the feedforward projections are adjusted to account for the acoustic perturbation.

In the SA protocol, only the auditory component of the sensory feedback is perturbed; the somatosensory feedback is left unperturbed. The model predicts that adaptation should not fully compensate for purely auditory perturbations due to the influence of somatosensory feedback control. That is, as the feedforward commands change to compensate for the auditory perturbation, somatosensory errors begin to arise and result in corrective motor commands that resist changes in the feedforward command. As observed above, analyses from the +*feedback* tokens of the SA subjects also demonstrated only partial compensation (refer to Sec. I B), supporting the model's prediction.

A. Modeling variation in auditory acuity

One important property of the DIVA model is its reliance on sensory goal *regions*, rather than *points* (Guenther *et al.*, 1998; Guenther, 1995). The notion of sensory goal regions explains a number of phenomena related to speech production. These observed behaviors include motor equivalent articulatory configurations (Guenther, 1995; Guenther *et al.*, 1998) and their use in reducing acoustic variability (Guenther *et al.*, 1998; Guenther *et al.*, 1999b; Nieto-Castanon *et al.*, 2005), as well as anticipatory coarticulation, carryover coarticulation, and effects related to speaking rate (Guenther, 1995).

Prior studies have demonstrated that speakers with greater auditory acuity produce more distinct contrasts between two phonemes (Newman, 2003; Perkell *et al.*, 2004a, b). According to the DIVA model, these larger contrasts result from the use of smaller auditory goal regions by speakers with better acuity; this may occur because these speakers are more likely to notice poor productions of a sound and thus not include them as part of the sound's target region. In keeping with this view, we created a version of the model for each individual subject by using an auditory target region size for the vowel /ɛ/ that was proportional to the subject's discrimination index. The details of this process are described in Appendix II. In short, subjects with a larger discrimination index (reflecting poorer acuity) were modeled by training the DIVA model with large target regions, while subjects with better acuity were modeled by training on smaller target regions. These varying trained models were then used in a simulation experiment that replicated the sensorimotor adaptation paradigm of Experiment 1.

B. Design of the SA simulations within the DIVA model

Twenty simulations were performed, using subject-specific versions of the DIVA model; each simulation corresponded to a particular subject's SA run, with the model's target region size adjusted using the relation between acuity and adaptive response described in Appendix II. Each simulation consisted of the same four phases as the human subject SMA experiment: *base line, ramp, full pert*, and *post pert*. During these phases, auditory feedback to the model was turned on and off to replicate the +*feedback* and −*feedback* SA results. Like the human subject experiment, the perturbation to F1 in the model's auditory feedback during the *full-pert* phase was either 0.7 or 1.3 pert (depending on the subject being simulated), and the perturbation was ramped up during the ramp phase as in the experiment.

In the SA experiment with human subjects, each epoch contained nine +*feedback tokens* and three −*feedback* tokens that contained the vowel /ɛ/. To maintain this ratio while simplifying the simulations, one epoch in the simulation was composed of four trials: three trials with feedback turned on, followed by one trial with feedback turned off.

C. Results

Figure 10 compares the results from +*feedback* trials in the DIVA simulations to the corresponding human subject

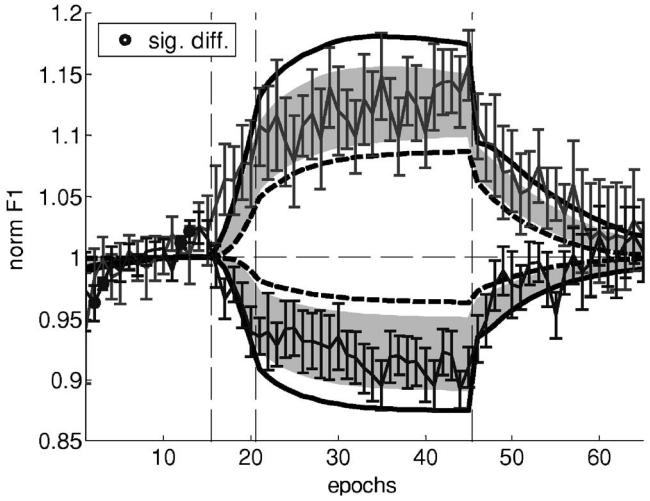

FIG. 10. Normalized F1 as a function of epoch number during the SA protocol in *+feedback* trials: DIVA simulations compared to human subject results. The thin lines shown with standard error bars correspond to the subject SA data (20 subjects). The shaded region corresponds to the DIVA simulations, and represents the 95% confidence interval about the mean. The vertical dashed lines show the experiment phase transitions; the horizontal dashed line indicates base line. The open circles indicate epochs in which the data and the simulation results were significantly different. The black solid curves correspond to high-acuity simulations, while the black dashed curves correspond to low-acuity simulations.

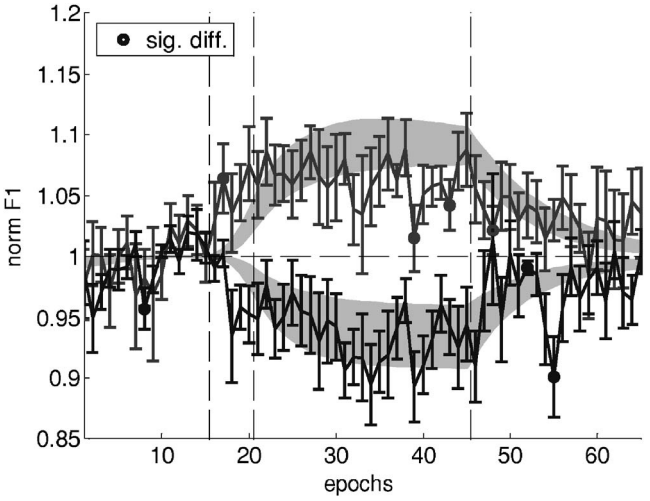

FIG. 11. Normalized F1 during the SA protocol in *–feedback* trials: DIVA simulations compared to subject results.

data. These results demonstrate that the SA simulations account for the main trends found in the human SA data: (1) a compensatory change in F1 that builds gradually over the *ramp* and *full pert* phases, (2) a maximum F1 deviation that only partially compensates for the perturbation, and (3) a gradual return to the base line F1 value in the *postpert* phase. Furthermore, acuity and the extent of F1 deviation are positively related in the model, evident by comparing the high acuity (solid lines) to the low acuity (dashed lines) simulations, as in the human subject data (not shown in Fig. 10). Finally, there is a slight asymmetry between the shift-up group and shift-down group, seen in both the simulations and the human subject results. This is not surprising, given that the inverse of the perturbation—which represents the maximal response expected—is a larger change from base line for the shift-down condition than for the shift up.

To determine if the simulation results were significantly different from the human subject results, a pooled, two-tail t test was performed on an epoch-by-epoch basis between the two sets of results; differences statistically significant at a level of $p=0.05$ are indicated in Fig. 10 by the open circles. The simulation results differed significantly only during four epochs, all of which were in the *base line* phase, where the experimental subjects showed considerable drift in F1 compared to the constant F1 of the model's productions. During the ramp phase, the human SA results seem to show a faster adaptive response than the simulation results, but this difference is not statistically significant.

Like the human subject results, the DIVA simulations produced very little change in the second formant (not shown): the normalized F2 during the full-pert phase had a mean value of $1.0135+/-0.0035$ for the shift-down simulations, and a mean value of $0.9975+/-0.0004$ for the shift-up simulations. It should be noted that the simulations and the humans subject results differed in the direction of the F2 changes; unlike the human subjects, who showed F1 and F2 shifting in opposite directions, the simulations showed changes in F1 and F2 occurring in the same direction. As described earlier, the shifting of F1 and F2 in opposite directions by the experimental subjects may indicate the use of an auditory planning frame that is not strictly formant based as implemented in the model simulations, but rather is better characterized by relative values of the formants and F0.

Figure 11 compares the results from *–feedback* trials in the DIVA simulations to the corresponding human subject data. The simulations exhibit adaptive responses that are similar in extent to those seen in human data in *–feedback* tokens. Excluding differences in the base line phase, the *–feedback* simulations differed from the human subject data in four epochs for the shift-down condition (one epoch in the ramp phase, two in the full-pert phase and one in the postpert phase), and in two epochs for the shift-up condition (both in the postpert phase). It should be noted that, because corrections for multiple comparisons were not done in order to make the test of the model more stringent, one would expect 2–3 epochs out of 50 to show "false" significant differences (for a significance threshold of $p=0.05$) even if the statistical distributions of model and subject productions were identical.

V. DISCUSSION

The studies presented in this article reveal several details of the process by which individuals modify speech in order to compensate for altered acoustic feedback. The results from Experiment 1 indicate that, in response to perturbations of the first formant (F1) in the acoustic feedback of vowel productions, subjects compensate by producing vowels with F1 shifted in a direction opposite to the perturbation. Specifically, shift-down subjects exhibited 35% compensation, and shift-up subjects exhibited 50% compensation. This range of compensation is similar to other experiments in vowel formant manipulation (Houde *et al.*, 1998, 2002; Max *et al.*, 2003). Although we observed an asymmetry in compensation relative to the direction of shift, this asymmetry arises from

how we defined complete compensation (see Sec. I B). Compensation was present even when the subjects' auditory feedback was blocked by masking noise—defined as true adaptation in Houde and Jordan (2002)—and it generalized to most other vowels that were not perturbed. This adaptation persisted for a period (roughly 10 epochs) after the perturbation was removed, decaying in a similar manner to that shown in Purcell and Munhall (2006).

Previous vowel formant SA experiments also exhibited inter-subject variation in the extent of adaptation, which was found in our experiment. Unlike those aforementioned experiments, we also measured auditory acuity to formant differences in a majority of our subjects (Experiment 2), allowing us to study a possible source of this inter-subject variation. Indeed, the results of Experiment 2 show that auditory acuity and compensatory responses to perturbations of F1 are significantly correlated: subjects with greater acuity demonstrated greater responses. This correlation increased when factoring out inherent speaker differences in the separation in F1 for neighboring vowels. This finding is similar to previous evidence of a linkage between the discrimination of vowel and sibilant acoustics and the production of those sounds with greater acoustic contrast (Perkell *et al.*, 2004b; Newman, 2003; Perkell *et al.*, 2004a); in the current study, "production contrast" is reflected in greater compensatory change in the first formant frequency. The strength of the correlation between subject acuity and amount of compensation when individual vowel spacing is factored out ($r=-0.79$; $p<0.001$) is noteworthy, especially considering possible confounds such as imperfect performance of the SA signal processing algorithm and variability inherent in the methods used for the perceptual testing.

In the simulations, the DIVA model was able to account quantitatively for several key characteristics of the human subject results. Similar to the human subject results, the DIVA simulations demonstrated formant-specific compensation to acoustic perturbation, adaptation in the absence of auditory feedback and persistence for a short period after the perturbation was removed. Such adaptation can be attributed to modification of feedforward commands that washes out once the source of auditory errors has been removed.

Individual auditory acuity for vowels is reflected in the DIVA model by auditory goal regions of varying size, which are smaller in individuals with greater acuity. The simulations also demonstrated that smaller auditory goal regions—hypothetically determined by the subject's acuity—were associated with greater compensatory changes in the first formant. Moreover, the DIVA simulations provide an additional explanation for the observation that partial compensation was measured in Experiment 1: feedback control utilizes both auditory inputs (perturbed by the SA algorithm) and somatosensory inputs (unperturbed). That is, the presence of somatosensory feedback in the model also acts to resist changes in the feedforward commands, limiting the extent of adaptation. Thus, the DIVA model provides a plausible account of the inter-subject variation and incomplete adaptation found in this and other (Houde and Jordan, 2002; Purcell and Munhall, 2006) speech SA experiments.

Even though only tokens containing the vowel /ɛ/ received perturbed feedback, the adaptation of F1 generalized to tokens containing other vowels as well, which suggests that the subjects are not learning to modify motor commands that are specific to just the vowel that was perturbed in the SA feedback. Rather, subjects appear to have learned to modify the articulations in a way so that the adapted response can be applied globally to other vowels. Generalization is an advantageous property for speech motor planning, since speech production is normally a generative process, with each utterance being unique. Generalizing adaptation learned for one specific context to other contexts also enhances an individual's ability to rapidly modify their spoken clarity and maintain intelligibility in the face of variations in acoustic transmission conditions.[10]

At the same time, the vowels exhibited different degrees of generalization, and the vowel /i/, was the most consistently unchanged (Fig. 5). It may be necessary to hear more than a single vowel being perturbed to completely and uniformly update the vowel formant map; however, the lack of generalization for /i/ may also be explained by the possibility that, unlike other vowels, /i/ has a well-defined somatosensory target (in addition to an auditory one) that is characterized by a "saturation effect"—pressing the sides of the tongue blade against the lateral aspects of the hard palate (Perkell, 1996; Fujimura and Kakita, 1979). In DIVA, this strong somatosensory target would resist compensation to auditory perturbations since the somatosensory feedback would counteract the auditory-based compensation as soon as it tried to move the production away from the somatosensory target. Because the DIVA model learns a feedforward command for each speech sound independently, it cannot in its current form account for these generalization results, which will be used to guide future modification of the model.

While subjects' adaptive responses were expressed mainly by adjustments of the formant that was perturbed (F1), they also exhibited small changes in the second formant (F2) and even the fundamental frequency (F0), in response to the F1 perturbation. These small changes in F2 and F0 occurred in directions that were opposite of the direction of F1 changes, and would be compensatory responses if subjects perceived vowel auditory dimensions in a normalized space sensitive to relative values of formants and F0, such as the perceptual space described by Miller (1989). Simulations using the Miller space were also run (Villacorta, 2006); however, with the exception of the very small changes in F2 and F0 seen in the experimental subjects, the experimental data were better fit by versions of the model utilizing straight formant frequencies rather than the Miller space.

The quantitative simulation of human subject SA results and their relation to speaker acuity by the DIVA model adds to the list of measured speech phenomena accounted for by the model (see also Callan *et al.*, 2000, Guenther, 1994, 1995; Guenther *et al.*, 1998, 2006; Nieto-Castanon *et al.*, 2005; Perkell *et al.*, 2004a, b). The notion of tightly coupled feedforward and feedback controllers responsible for the model's ability to account for the experimental data is also a feature of other sensorimotor control CHI architectures (Wolpert and Kawato, 1998), although it is not clear if the

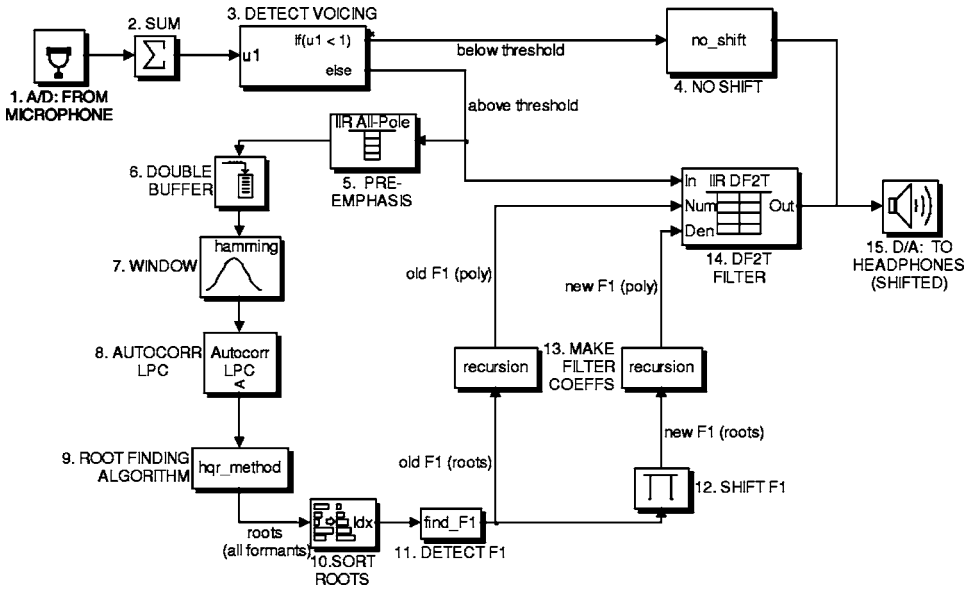

FIG. 12. Block diagram of the formant-shifting algorithm used to introduce acoustic perturbations in the SA experiment.

details of these architectures (which differ somewhat from the DIVA model) would allow quantitative fits of the current data. Specifically, we have attributed the finding that subjects exhibited a lower adaptive response in their *–feedback* vowels when compared to their *+feedback* vowels (also seen by Houde and Jordan, 2002) to the countervailing influence of unperturbed somatosensory feedback, which would be more dominant in the absence of auditory feedback. These hypotheses would be put to a more stringent test in an experiment that incorporated both auditory perturbations (Tourville *et al.*, 2004; Purcell and Munhall, 2006)—as in this study—and somatosensory perturbations (see Honda and Murano, 2003; Tourville *et al.*, 2004)—in the same experimental run.

ACKNOWLEDGMENTS

This research was supported by Grant No. DC01925 and DC02852 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health. We are grateful to Oren Civier for his help with the DIVA code, to Harlan Lane for his advice on various aspects of the study and to two thoughtful reviewers for their constructive comments.

APPENDIX I. USE OF LPC COEFFICIENTS TO DETERMINE AND SHIFT F1

The following describes the signal processing that was utilized to introduce F1 shifts in nearly real time, as illustrated in Fig. 12. The DSP board receives an analog speech signal from the microphone and converts it to a digital signal, which is sent to the receiving (Rx) buffer. One of the algorithm's first functions is to calculate the sum of all values within the Rx buffer to determine its amplitude, and then determine if this value is above or below a threshold value (with signals above threshold indicating vowel production). The threshold value is set so that values below it are not sent through the formant shifting algorithm, while values above it—those within a vowel—are.

If the Rx buffer contains values from within a vowel, the signal is then pre-emphasized to improve LPC formant analysis, and coupled with the previous buffer to improve frequency resolution. A Hamming window is then applied to this double-buffered frame, followed by autocorrelation LPC analysis. The output of this operation is an 8th order polynomial (see Eq. (A1)), which can characterize up to four poles of the spectrum resulting from the analyzed speech buffer

$$A(z) = 1 + \sum_{i=1}^{8} a_i z^{-i}. \tag{A1}$$

Equation (A1) can alternatively be written with the complex roots of $A(z)$ stated explicitly, as in Eq. (A2). Written in this way, the formants of the analysis buffer are directly related to the angle θ of the complex roots. Thus, to pick out individual formants from the LPC polynomial, it is necessary to determine its complex roots, utilizing an iterative rootfinding algorithm based on the Hessenberg QR method (Press *et al.*, 2002)

$$A(z) = \prod_{i=1}^{4} (1 - c_i z^{-1})(1 - c_i^* z^{-1}),$$

where

$$c_i = r_i \cos\,\theta_i + j r_i \sin\,\theta_i,$$

$$c_i^* = r_i \cos\,\theta_i - j r_i \sin\,\theta_i. \tag{A2}$$

The roots are then sorted based on angle of the complex roots, and F1 is determined from the sorted array of roots as the lowest non-negative, nonzero root. The root related to the shifted F1 is calculated by rotating the angle of the complex root representing the original F1. A simple recursion formula is used to convert the roots of the original and shifted F1 values to polynomial coefficients. Using the new coefficients, the perturbation algorithm regenerates the speech signal with the shifted F1 value by implementing a direct-form II transposed filter (Oppenheim and Schafer, 1999) to filter data within the Rx current buffer; the original F1 is zeroed out while the perturbed F1 value is introduced simultaneously. The time-domain difference equation corresponding

to this filtering is described in Eq. (A3), with r and θ representing the magnitude and angle resulting from the complex root corresponding to the first formant

$$y[n] = x[n] - (2r\cos\theta)x[n-1] + r^2x[n-2] + (2r\cos\theta')y[n-1] - r^2y[n-2]. \quad \text{(A3)}$$

Regardless of whether the current buffer is shifted or not, the resulting speech is moved into the transfer (Tx) buffer, which is then converted back to an analog signal and sent to the output of the DSP board. The delay from the algorithm is 128 samples with sampling rate of 8000 Hz. Including a 2 ms delay resulting from anti-aliasing filtering from the analog-to-digital conversion, the overall delay in the board is 18 ms.

Aside from the threshold criterion, another criterion for shifting F1 is that the detected value had to fall within the following ranges of frequencies:

$$250\ \text{Hz} < F1 < 950\ \text{Hz}\ (\textit{male subjects})$$
$$400\ \text{Hz} < F1 < 950\ \text{Hz}\ (\textit{female subjects}). \quad \text{(A4)}$$

F1 values below the lower limit of the window tended to be near the value of the fundamental frequency, while F1 values above the window's upper limit tended to be very close to the value of the second formant. That is, a formant value detected outside the window is likely to not be the actual F1, indicating that it should be excluded. Valid F1 values could occur outside of this range, which was a basis for rejecting the data from a preliminary subject (Villacorta, 2006).

As mentioned above, the experimental setup allowed for simultaneous recording of the input to the DSP board (no perturbation) and output of the board (with perturbation) for off-line analysis. As a validation of the effectiveness of the perturbation procedure, unperturbed and perturbed vowel formants from the same recording compared for preliminary subjects who did not hear feedback. These results showed that most of the shifted values were within 2.5% of the expected shift regardless of direction.

APPENDIX II. IMPLEMENTATION OF SUBJECT AUDITORY ACUITY IN THE DIVA MODEL

Previous versions of the DIVA model implemented a conceptualization of auditory goal regions in which auditory feedback error resulted only if the actual feedback fell outside the auditory goal region, while no feedback error resulted if the actual feedback fell within the auditory goal region (Guenther *et al.*, 2006; Guenther, 1995; Guenther *et al.*, 1998). Here, the discontinuity in the feedback error signal was removed by representing the actual feedback and the goal region as Gaussian distributions. The magnitude of the feedback error signal was proportional to the rectified difference between these distributions (ensuring that the feedback error signal is smaller when the actual feedback is closer to the goal), while the direction of the feedback error signal was determined by the position of the actual feedback relative to the center of the goal region (ensuring that the feedback signal compensates for the perturbation).

Speech motor planning systems of subjects of differing auditory acuity were simulated by changing the variance of the Gaussian distributions, with subjects having greater acuity possessing smaller variance values (and thus narrower goal regions). To do this, the regression line between the discrimination indices and the adaptive response indices shown in Fig. 8 was used to determine the subject's auditory region boundary size from the subject's measured ARI score. While only 13 subjects had measured discrimination index scores, all 20 subjects were simulated in the DIVA model utilizing their ARI scores and the linear relation. Had the simulation results been limited to the 13 subjects whose acuity was measured, the results would have been essentially the same, although the estimate of the true distribution of the model's productions would not have been as good.

[1]"DIVA" is an acronym for *Directions into Velocities of Articulators*; the model is so named because of its reliance on mappings that transform sensory errors into corrective motor commands.

[2]The use of insert headphones could have enhanced bone conduction of low-frequency energy, including the frequency region of F1 of the unperturbed speech signal (see Porschmann, 2000), making it possible that the subject heard a mixture of perturbed and unperturbed signals. As mentioned previously, the likelihood of the occurrence of such a confound was minimized by determining informally that the ratio of the level presented by the headphones to that of the subject's sound output, approximately 18 dB, was sufficient to mask produced vowel quality.

[3]Presentation and discussion of results from individual subjects are beyond the scope of the current report; for such details, see Villacorta (2006).

[4]Figure 3 shows a gradual increase in F1 values during the base line phase in both shift-up and shift-down subjects; this increase was especially notable in the first five epochs of the base line phase. To exclude low F1 values observed in the early part of the base line phase, the normalization shown in Eq. (1) used epochs 6–15 (an adjusted base line phase). This gradual F1 increase is discussed further in Villacorta (2006).

[5]The vowel /o/ in "pote" was also on the word list; as a diphthong, this vowel had large variations in formant values vs. time and was not included in the analysis.

[6]Subjects in Jones and Munhall (2000) produced an upward F0 shift in response to F0 shifts in their acoustic feedback, regardless of whether they were exposed to shift-up, shift-down or control protocols. Subjects exposed to the shift-down protocol increased F0 to a greater degree than the controls, while those exposed to the shift-up protocol increased in F0 to a lesser degree than the controls. When the overall increase in F0 was factored out, the subjects produced a shift in F0 that was opposite to the F0 perturbation in their auditory feedback.

[7]Initially, auditory acuity, in the form of *jnds*, were determined for each subject at three milestones: lower (at 0.85 pert), center (at 1.0 pert), and upper (at 1.15 pert). Only the results of the center milestone *jnd* determination are discussed here, as no significant findings resulted from cross-subject correlations involving the *jnd* from the two other milestones. A goodness-rating task, in which subjects were instructed to rate 41 tokens ranging from 0.7 to 1.3 pert for how well they sounded like the vowel /ɛ/, was also run as part of the perceptual experiment. Results from the goodness-rating task and the noncenter milestone *jnd* measurements are discussed in depth in Villacorta (2006).

[8]The stimulus pairs initially were separated from each other by 0.30 pert. The first four changes in separation were 0.04 pert, then subsequently by changes in separation of 0.02 pert. Once the tokens were within 0.10 pert from each other, the separation was only changed by 0.01 pert. After eight reversals (changes in direction of the staircase), the protocol terminated, and the jnd_{est} was calculated as the median value of the last four reversals on the staircase. Two of the subjects had jnd_{est} that were higher than the initial value set at the beginning of the staircase protocol; that is, the staircase "climbed" rather than "descended." We assumed that this was due to an initial misunderstanding of the protocol instructions, and re-ran the entire acuity experiment, including regeneration of the stimulus vowel continuum.

[9]Two subjects had adaptive response indices that were negative, indicating that they changed their productions in the same direction as the perturbation

rather than in the opposite direction. This result has been found in other auditory perturbation experiments (e.g., Burnett *et al.*, 1998). A speaker with poor acuity, who does not detect a change at all, is equally likely to have a small positive or small negative ARI. Those with good acuity have a large compensation component so their ARI is never negative. For this reason the negative adaptive responses were included in the correlation analyses rather than being removed as outliers.

[10]There is another possible explanation for generalization: vowels are learned and perhaps controlled as part of a paradigm or system. If the feature values of one vowel are modified, the system is changed in a way that could lead to changes in other vowels (Harlan Lane, personal communication).

Abbs, J. H., and Gracco, V. L. (**1984**). "Control of complex motor gestures: Orofacial muscle responses to load perturbations of lip during speech," J. Neurophysiol. **51**, 705–723.

Bedford, F. L. (**1989**). "Constraints on learning new mappings between perceptual dimensions," J. Exp. Psychol. Hum. Percept. Perform. **15**, 232–248.

Bhushan, N., and Shadmehr, R. (**1999**). "Computational nature of human adaptive control during learning of reaching movements in force fields," Biol. Cybern. **81**, 39–60.

Blakemore, S. J., Goodbody, S. J., and Wolpert, D. M. (**1998**). "Predicting the consequences of our own actions: The role of sensorimotor context estimation," J. Neurosci. **18**, 7511–7518.

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.

Callan, D. E., Kent, R. D., Guenther, F. H., and Vorperian, H. K. (**2000**). "An auditory–feedback based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system," J. Speech Lang. Hear. Res. **43**, 721–736.

Fujimura, O., and Kakita, Y. (**1979**). "Remarks on quantitative description of lingual articulation," in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Öhman (Academic, San Diego), pp. 17–24.

Guenther, F. H. (**1995**). "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production," Psychol. Rev. **102**, 594–621.

Guenther, F. H. (**1994**). "A neural network model of speech acquisition and motor equivalent speech production," Biol. Cybern. **72**, 43–53

Guenther, F. H., Espy-Wilson, C. Y., Boyce, S. E., Matthies, M. L., Zandipour, M., and Perkell, J. S. (**1999b**). "Articulatory tradeoffs reduce acoustic variability during American English /r/ production," J. Acoust. Soc. Am. **105**, 2854–2865.

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (**2006**). "Neural modeling and imaging of the cortical interactions underlying syllable production," Brain Lang **96**, 280–301.

Guenther, F. H., Hampson, M., and Johnson, D. (**1998**). "A theoretical investigation of reference frames for the planning of speech movements," Psychol. Rev. **105**, 611–633.

Guenther, F. H., Husain, F. T., Cohen, M. A., and Shinn-Cunningham, B. G. (**1999a**). "Effects of categorization and discrimination training on auditory perceptual space," J. Acoust. Soc. Am. **106**, 2900–2912.

Guenther, F. H., Nieto-Castanon, A., Ghosh, S. S., and Tourville, J. A. (**2004**). "Representation of sound categories in auditory cortical maps," J. Speech Lang. Hear. Res. **47**, 46–57.

Honda, M., and Murano, E. (**2003**). "Effects of tactile and auditory feedback on compensatory articulatory response to an unexpected palatal perturbation," Proceedings of the Sixth Speech Production Seminar, Sydney, Australia, Dec. 7–10, 2003.

Houde, J. F., and Jordan, M. I. (**1998**). "Sensorimotor adaptation in speech production," Science **279**, 1213–l216.

Houde, J. F., and Jordan, M. I. (**2002**). "Sensorimotor adaptation of speech I: Compensation and adaptation," J. Speech Lang. Hear. Res. **45**, 295–310.

Jones, J. A., and Munhall, K. G. (**2000**). "Perceptual calibration of F0 production: Evidence from feedback perturbation," J. Acoust. Soc. Am. **108**, 1246–1251.

Kawahara, H. (**1993**). "Transformed auditory feedback: Effects of fundamental frequency perturbation," J. Acoust. Soc. Am. **94**, 1883–1884.

Kawato, M., and Gomi, H. (**1992**). "The cerebellum and VOR/OKR learning models," Trends Neurosci. **15**, 445–453.

Lane, H. L., and Tranel, B. W. (**1971**). "The Lombard sign and the role of hearing in speech," J. Speech Lang. Hear. Res. **14**, 677–709.

Lindblom, B. E. F., Lubker, J. F., and Gay, T. (**1979**). "Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation," J. Phonetics **7**, 147–161.

Macmillan, N. A., and Creelman, C. D. (**2005**). *Detection Theory: A User's Guide*, 2nd ed. (Lawrence Erlbaum, Mahwah, NJ).

Maeda, S. (**1990**). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modeling*, edited by W. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp. 131–149.

Markel, J. D., and Gray, A. H. (**1976**). *Linear Prediction of Speech* (Springer-Verlag, New York).

Max, L., Wallace, M. E., and Vincent, I. (**2003**). "Sensorimotor adaptation to auditory perturbations during speech: Acoustic and kinematic experiments," *Proceedings of the 15th Internatiional Congress of Phonetic Sciences*, Barcelona, pp. 1053–1056.

Miller, J. D. (**1989**). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am. **85**, pp. 2114–2134.

Newman, R. S. (**2003**). "Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report," J. Acoust. Soc. Am. **113**, 2850–2860.

Nieto-Castanon, A., Guenther, F. H., Perkell, J., and Curtin, H. D. (**2005**). "A modeling investigation of articulatory variability and acoustic stability during American English /r/ production," J. Acoust. Soc. Am. **117**, 3196–3212.

Oppenheim, A. V., and Schafer, R. W. (**1999**). *Discrete-Time Signal Processing*, 2nd ed. (Prentice–Hall, Upper Saddle River, NJ).

Perkell, J. S. (**1996**). "Properties of the tongue help to define vowel categories: Hypotheses based on physiologically oriented modeling," J. Phonetics **24**, 3–22.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E. S., and Tiede, M. (**2004a**). "The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts," J. Acoust. Soc. Am. **116**, 2338–2344.

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., and Marrone, N. (**2004b**). "The distinctness of speakers' /s/–/ʃ/ contrast is related to their auditory discrimination and use of an articulatory saturation effect," J. Speech Lang. Hear. Res. **47**, 1259–1269.

Porschmann, C. (**2000**). "Influences of bone conduction and air conduction on the sound of one's own voice," Acta Acust. (Beijing) **86**, 1038–1045.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (**2002**). *Numerical Recipes in C*, 2nd ed. (Cambridge University Press, Cambridge).

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.

Savariaux, C., Perrier, P., and Orliaguet, J. P. (**1995**). "Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production," J. Acoust. Soc. Am. **98**, 2428–2842.

Tourville, J. A., Guenther, F. H., Ghosh, S. S., and Bohland, J. W. (**2004**). "Effects of jaw perturbation on cortical activity during speech production," J. Acoust. Soc. Am. **116**, 2631(A).

Vallabha, G. K., and Tuller, B. (**2002**). "Systematic errors in the formant analysis of steady-state vowels," Speech Commun. **38**, 141–160.

Villacorta, V. M. (**2006**). "Sensorimotor adaptation to perturbations of vowel acoustics and its relation to perception," Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Welch, R. B. (**1978**). *Perceptual Modification: Adapting to Altered Sensory Environments* (Academic, New York).

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (**1995**). "Are arm trajectories planned in kinematic or dynamic coordinates? An adaptation study," Exp. Brain Res. **103**, 460–470.

Wolpert, D. M., and Kawato, M. (**1998**). "Multiple paired forward and inverse models for motor control," Neural Networks **1217**, 1–13.

Xu, Y., Larson, C. R., Bauer, J. J., and Hain, T. C. (**2004**). "Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences," J. Acoust. Soc. Am. **116**, 1168–1178.

Yates, A. J. (**1963**). "Delayed auditory feedback," Psychol. Bull. **60**, 213–232.