SENSORY GOALS AND CONTROL MECHANISMS FOR PHONEMIC ARTICULATIONS

Joseph S. Perkell

Speech Communication Group, Research Laboratory of Electronics,
Massachusetts Institute of Technology, Cambridge, MA, 02139 U.S.A.
perkell@speech.mit.edu

ABSTRACT

An overview of speech production is described in which the goals of phonemic speech movements are implemented in auditory and somatosensory domains and the movements are controlled by a combination of feedback and feedforward mechanisms. Findings of motor-equivalent trading relations in producing /u/ and /r/, cross-speaker relations between vowel and consonant production and perception, and speakers' use of a "saturation effect" in producing /s/ support the idea that the goals are in sensory domains. Results of production experiments in which auditory feedback was modified and interrupted provide insight into the nature of feedback and feedforward control mechanisms. The findings are all compatible with the DIVA model of speech motor planning [3], which makes it possible to quantify relations among phonemic specifications of utterances, brain activity, articulatory movements and the speech sound output.

Keywords: phonemic goals; auditory feedback; feedback control; feedforward control.

1. INTRODUCTION

This paper is concerned with the nature of motor programming goals for phonemic speech articulations and how feedforward and feedback mechanisms are used to produce the movements that achieve those goals.

It is widely acknowledged that properties of the speech production mechanism have had major influences on the inventories of sounds or phonemes that languages employ, and also on some of the strategies that languages adopt for concatenating phonemes into meaningful sequences. A great deal of research on speech motor control and the mechanisms that underlie sound categories has been directed at identifying the *controlled variables*, that is, the basic units of speech motor programming. To address this issue, investigators have asked, "What is the *task space*, or the domain of the fundamental control parameters?"

Our approach to this question is motivated by observing that the objective of the speaker is to produce sound strings with acoustic cues that can be transformed into intelligible patterns of auditory sensations in the listener. These acoustic cues consist mainly of time-varying patterns of formant frequencies for vowels and glides, and noise bursts, silent intervals, aspiration and frication noises, and rapid formant transitions for consonants. The properties of such cues are determined by parameters that can be observed in several domains, including: levels of muscle tension, movements of articulators, changes in the vocal-tract area function and aerodynamic events. Hypothetically, motor control variables could consist of any combination of these parameters.

Several recent lines of evidence have supported the view that goals for phonemic articulations are in sensory domains – auditory and somatosensory. Such findings are compatible with the function of the DIVA model of speech motor planning [3]. DIVA is a neurocomputational model of relations among cortical activity for producing speech sounds, the motor output and the resulting sensory consequences. In the model, phonemic goals are encoded in neural projections (mappings) from premotor cortex to sensory cortex, mappings that describe *regions in multidimensional auditory-temporal and somatosensory-temporal spaces*. The model has two control subsystems, a feedback subsystem and a feedforward subsystem. Feedback control employs error detection and correction to teach, refine and update feedforward control mechanisms. As speech is acquired and becomes fluent, speech sounds, syllables and words become encoded as sequences of feedforward commands that no longer rely on auditory feedback.

The following sections summarize several results from our laboratory that support this view.

2. PHONEMIC GOALS

How are phonemic goal regions determined? One influence is from properties of speakers' production mechanisms that are characterized by quantal relations between articulation and acoustics [12]. There are a number of examples in which a continuous change in an articulatory parameter produces discontinuous changes in a salient acoustic parameter, resulting in regions of relative acoustic stability and regions of rapid change. Modeling and experimental results support the idea that such regions of stability help to define phonemic goals and sound categories [12, 13, 10, 5].

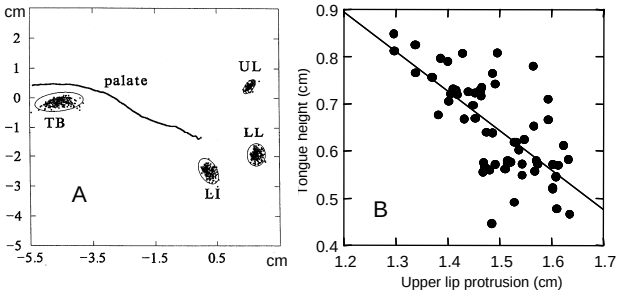
There are also quantal relations between articulatory movements and the area function, which are expressed when two articulators come into contact with one another. Fujimura and Kakita [1] have modeled such a "saturation effect" for the vowel /i/ by showing how the vocal-tract cross-sectional area at the acoustically sensitive place of maximum constriction can be stabilized by pressing the lateral edges of a stiffened tongue blade against the sides of the hard palate with co-contraction of the anterior and posterior portions of the genioglossus muscle.

Such mechanisms can provide a general framework for the determination of sound patterns, and more specific implementations of the mechanisms can be utilized by individual speakers. One such example is shown below for a saturation effect in the production of the sound /s/. Other examples below provide support for some of the features of the DIVA model, including the use of sensory goals regions and feedback and feedforward control.

2.1. Auditory goals for /u/ and /r/: Motor equivalence

The use of auditory goals is consistent with findings of articulatory-to-acoustic motor equivalence for the vowel /u/ [8] and the semivowel /r/ [2]. The vowel /u/ in American English is produced by forming a narrow constriction with tongue raising in the velo-palatal region and by rounding the lips. Because of the many-to-one relation between vocal-tract shapes and acoustics, approximately the same acoustic output can be produced with more tongue raising and less lip rounding and vice-versa. Figure 1B shows an example of tongue height versus lip protrusion for many repetitions of the vowel /u/ by a single speaker (as shown in Fig. 1A). The negative correlation reflects a motor-equivalent trading relation between the two articulations. Such reciprocal variation of two independently controllable articulations provides evidence that the goal for the vowel /u/ is in an acoustic/auditory frame of reference, rather than a spatial or gestural one [8]. Evidence of an auditory/acoustic goal for /r/ in American English was obtained in a similar motor-equivalence study by Guenther et al. [2].

Figure 1: A: Midsagittal view of points on the tongue body (TB), Upper lip (UL), lower lip (LL) and lower incisors (LI) for many repetitions of the vowel /u/ by a single speaker in a context phrase. B: Tongue height versus lip protrusion for many repetitions of the vowel /u/ by a single speaker.

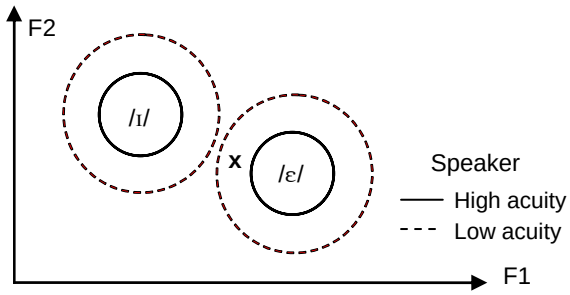


2.2. Auditory goals: Relations between speech production and perception

Further insight about auditory goals can be gained by examining relations between speech production and perception. It is well known that if an individual is born without hearing, that person has a very difficult time learning how to speak intelligibly. On the other hand, if someone acquires speech normally and then becomes completely deaf postlingually, the person's speech can remain intelligible for decades without any useful hearing. However, the speech of such individuals does gradually develop some anomalies following hearing loss. A number of studies have been conducted on speakers who became deaf in adulthood, went without hearing for a number of years and then received a cochlear implant. Results show that phonemic goals are stable, but contrasts can diminish gradually without hearing. Restoration of some hearing with an implant usually results in parallel improvements in perception, measures of contrast in production and speech intelligibility (cf. [5, 16]).

In another kind of approach, we have conducted studies of vowel and sibilant production and perception with 19 normal-hearing young adult speakers of American English. For two vowel contrasts and the sibilant (/s/-/ʃ/) contrast, we measured

Figure 2: Schematic diagram of auditory/acoustic goal regions for the vowels /ɪ/ and /ɛ/ in F1xF2 space. The dashed circles represent goal regions for a speaker with low auditory acuity; the solid circles, goal regions for a high-acuity speaker. The X indicates an example of /ɛ/ that is acceptable to the low-acuity speaker and unacceptable to the high-acuity speaker.



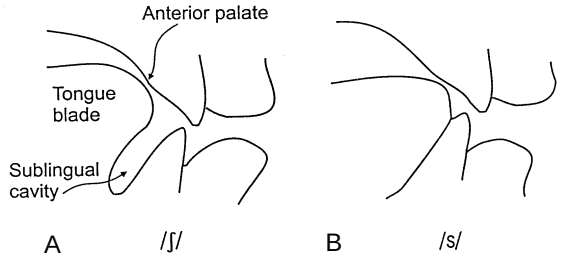
each speaker's degree of produced contrast and the speaker's auditory acuity. Produced vowel contrast distances were measured in articulatory and formant (F1, F2) spaces and the produced sibilant contrast was measured as the difference in spectral means between /s/ and /ʃ/. Auditory acuity was measured as the subjects' ability to discriminate between pairs of natural-sounding synthetic stimuli along continua between each of the contrasting sounds. Both studies found that *speakers with greater acuity produced the sounds with greater contrast*.

To interpret these results, we assume that spoken-language learners find it advantageous to be as intelligible as possible and therefore acquire auditory goal regions that are as distinct as possible. We reason that speakers who can perceive fine acoustic details will learn auditory goal regions that are smaller and spaced further apart than speakers with less acute perception, because, as schematized in Fig. 2, the speakers with more acute perception are more likely to reject poorly produced tokens when acquiring the goals [6, 9].

2.3. A somatosensory goal and saturation effect: The sibilant contrast

We have hypothesized that the sibilant sound /s/ has a somatosensory goal as well as an auditory one. The somatosensory goal is characterized by a saturation effect, which enhances the contrast of /s/ with its homologue, /ʃ/. As schematized in Fig. 3, /ʃ/ is produced by positioning the tongue blade so there is a sublingual cavity. This cavity adds volume and complexity to the resonant cavity anterior to the constriction and thereby contributes to the lower spectral center of gravity of the frication noise. On the other hand, /s/ is produced by pressing the under-side of the tongue blade against the lower alveolar ridge and incisors, which eliminates the sublingual cavity and results in a smaller anterior resonator that contributes to a higher spectral center of gravity. When the tongue blade is moved forward to produce an /s/, once the sublingual cavity is eliminated, further contraction of the muscles that produce the forward movement will increase the contact pressure but will have a negligible effect on the size of the resonant cavity. Thus, making this contact, which can be considered a somatosensory goal for the sound /s/, is characterized as a saturation effect.

Figure 3: Schematic midsagittal-plane representations of tongue blade configurations for producing an /ʃ/ (A) and an /s/ (B). /ʃ/ is produced with a sublingual cavity, which contributes to the lower mean frequency of its acoustic spectrum; /s/ is produced with contact between the under side of the tongue blade and the lower incisors.

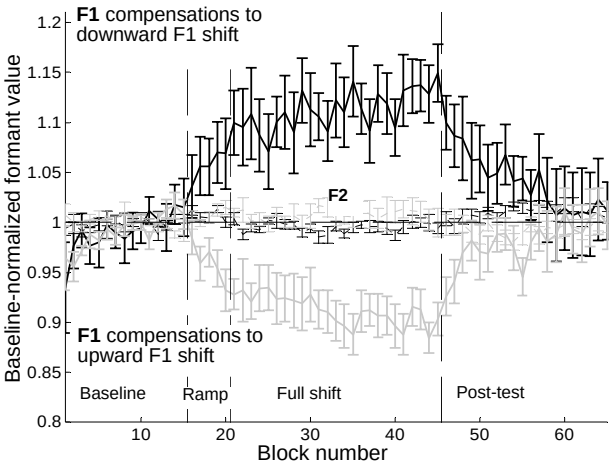


We also made measurements of the consistency of sublingual contact during /s/ production in the above-described perception/production study. *The most distinct sibilant productions were made by subjects who used contact in producing /s/ but not /ʃ/ and had higher acuity for the contrast.* Subjects who did not use contact differentially and had lower acuity produced the least distinct contrasts. Intermediate degrees of contrast were found with subjects who used contact differentially or had higher acuity [9].

3. FEEDBACK AND FEEDFORWARD CONTROL

To investigate feedforward and feedback control mechanisms in speech, investigators have conducted studies in which subjects' feedback has been perturbed and their compensatory responses measured. In the auditory domain, some studies used steady-state perturbations, such as blocking hearing with masking noise; others have used intermittent auditory perturbations that the subjects

Figure 4: Compensatory responses to F1 shifts in normal-hearing subjects. Average values of subjects' baseline-normalized F1 and F2 vs. block number. Each block contains one repetition of each of 18 different words in the corpus. The curves above baseline show the average of 10 subjects' productions in response to a downward shift of F1; the curves below baseline, the average of 10 subjects' responses to an upward F1 shift.



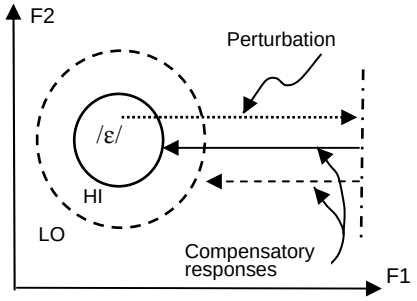
cannot anticipate. Unanticipated modifications of auditory feedback have revealed that mechanisms are available that can detect and correct production errors within about 100 to 150 ms from the onset of the perturbation [15]. Therefore, if a movement lasts long enough, auditory errors can be corrected during the movement itself with the use of closed-loop feedback. Similar results have been found in a number of articulatory perturbation experiments.

However, many articulatory movements in mature, fluent speech do not last long enough for closed-loop feedback-based error correction. It follows that fluent adult speech production is controlled almost entirely by feedforward mechanisms, as in the DIVA model [3].

3.1. Sensorimotor adaptation, goal region size and auditory acuity

Figure 4 shows the results of an experiment that investigated feedforward control in 20 normal-hearing speakers. The subjects pronounced /CɛC/ words while the first formant frequency (F1) in their auditory feedback being was shifted in nearly real time (18 ms delay), without their being aware of the shift [17]. Ten of the subjects received upward shifts and the other 10, downward shifts. The plots show that the subjects partially compensated for the shifts over many trials by modifying their productions so that F1 moved in the direction opposite to the shift (also see [4, 11]). The temporary modification of feedforward commands is evidenced by the persistence of compensation (called "adaptation") for some time even when the perturbation is removed in the "post-test" phase (Fig. 4).

Figure 5: Schematic diagram of goal regions and compensatory responses for /ɛ/ for a high-acuity speaker (solid circle) and a low-acuity speaker (dashed circle). F1 perturbation is indicated by the dotted arrow, and compensatory responses, by the solid and dashed arrows.



The subjects' auditory acuity was also measured. There was a significant correlation between subjects' acuity and amount of compensation to the F1 shift: *speakers with better acuity tended to compensate more* [18].

What underlies this correlation between acuity and compensation? Figure 5 schematizes how two speakers differing in acuity, and therefore in the sizes of their goal regions for the vowel /ɛ/, might respond to a perturbation of F1. The high-acuity speaker has a smaller goal region. The perturbation of F1 is indicated by a dotted arrow pointing to the right, and the shifted value of F1, by a vertical broken line. This high-acuity speaker, in response to the shift in F1, will produce a greater compensatory response (middle arrow) than the one with lesser acuity. This is because the speaker continues to compensate until the F1 of his or her auditory feedback (which includes the shift) moves into the goal region. The distance between the shifted value of F1 (vertical line) and the edge of the goal region is greater for the high acuity speaker. In the DIVA model, auditory feedback provides closed-loop corrections of current motor commands and then modifications of feedforward commands for subsequent movements.

3.2. Time course of speech changes in response to short-term changes in hearing state

In this study, the timing of changes in segmental and suprasegmental speech parameters was

Table 1: Summary of the direction and statistical significance of changes in vowel duration when auditory feedback was blocked and restored (switched) in a group of six cochlear implant users. The changes are between pre-switch utterances and the first, second and third utterances following the switch, which was made within 20 ms of the onset of V_1 in post-switch utterance 1 (shaded column). V_1 = the vowel in the first word; S = the sibilant at the beginning of the second word; V_2 = the vowel in the second word. + = significant increase; – = significant decrease; 0 = no significant change.

Parameter	Hearing Switch	Post-switch Utterance								
		1			2			3		
		V_1	S	V_2	V_1	S	V_2	V_1	S	V_2
Vowel Duration	Block	+		+	+		+	+		+
	Restore	0		–	–		–	–		–

investigated in six cochlear implant users by switching their implant microphones off and on a number of times in a single experimental session. In effect, blocking and restoring hearing in this way imposes a sudden change in acoustic transmission conditions, analogous to the temporary occurrence of loud environmental noise.

The subjects produced multiple repetitions of /dV_1n#SV_2d/ utterances, *Don shad*, *Don sad*, *Dun shed*, and *Dun said,* in quasi-random order. Thus, there were two vowel contrasts, /ɑ/-/ʌ/ (*Don* vs. *Dun*) in the first word position (V_1) and /æ/-/ɛ/ (*shad* vs. *shed*) in the second word position (V_2), and the sibilant contrast /s/-/ʃ/ (see Table 1).

The changes between hearing and non-hearing states were introduced under computer control by a voice-activated switch at V_1 onset (shaded column in Table 1); the number of utterances between switches was varied to minimize subject anticipation of the switches. Measures of the suprasegmental parameters of SPL, duration (reflecting speaking rate) and F0 were made from the vowels, and segmental contrast distances were measured for the vowels and sibilants. Changes in parameter values were computed by averaging data from multiple tokens, lined up with respect to the switch. The changes were calculated between averaged pre-switch values and values from the first, second and third utterances following the switches (post-switch utterances 1, 2 and 3) [7].

Contrast measures for the vowels and sibilants did *not* exhibit significant changes that were maintained consistently during the three post-switch utterances. On the other hand, as shown in Table 1, vowel durations increased during the vowel in which hearing was blocked (V_1, utterance 1) and they decreased for the second vowel (V_2) in utterance 1 when hearing was restored. (Vowel durations were all greater than 150 ms.) Similar results were found for SPL and F0. The changed suprasegmental values were maintained consistently until the time of the next switch in hearing state. We speculate that the duration decrease with hearing restored did not take place until the following syllable (V_2, utterance 1) because neural processing and muscle activation delays made it impossible to truncate motor commands already issued for producing V_1.

Why were there no consistent changes in sound contrasts when hearing state was switched? According to the DIVA model, *short-latency contrast changes* should occur when auditory feedback is *modified,* as in [15], but *not* when it is simply *blocked or restored*. The current results indicate that the mechanism regulating speaking rate is at least partly under closed-loop control since changing the availability of auditory feedback resulted in changes in vowel durations. These differences in changes between segmental contrasts and suprasegmental parameters (e.g., rate, as measured by durations) are consistent with previous findings, which also indicate that the two types of parameters are controlled differently [14, 5, 7].

4. SUMMARY

According to our theoretical overview and experimental results, the control variables for phonemic movements consist of auditory-temporal and somatosensory-temporal goal regions, which correspond to expected sensory consequences of producing speech sounds. Findings of motor-equivalent trading relations for the vowel /u/ and the semivowel /r/ support the idea that their goals are at least partly auditory. Findings that speakers with better acuity produce more distinct sound contrasts indicate that more acute speakers may learn smaller, more distinct goal regions. This idea is also supported by results showing that speakers compensate for shifts in the first formant in their auditory feedback of vowels they are producing, and the amount of compensation is related to their acuity for small differences in the vowel spectra.

As hypothesized in the Introduction, feedback control of segmental parameters involves the detection and correction of mismatches between expected and actual sensory consequences of speech articulation [3]. Experimentally induced, unexpected modifications of auditory feedback can elicit observable rapid responses that seem to be

closed-loop [15]. However, under real-world circumstances auditory disparities between intended and produced speech sounds tend to occur or be maintained over long time spans, e.g., from vocal-tract growth or the insertion of dentures. Therefore, the primary role of auditory feedback control of segmental contrasts is to provide corrections that are incorporated into feedforward commands (as demonstrated in the laboratory in sensorimotor adaptation experiments [4, 11, 17, 18]). On the other hand, changes in acoustic transmission conditions, such as the occurrence of sustained loud noises, are experienced often and call for rapid responses for the maintenance of intelligibility. It follows that unexpectedly blocking and restoring auditory feedback engages a different feedback control mechanism from the one that helps to acquire and maintain segmental contrasts [7].

The results described above are compatible with the function of the DIVA model of speech motor planning – in the way it employs sensory goal regions and feedforward and feedback control mechanisms. Since the DIVA model is formulated in terms of patterns of cortical connectivity and activity, it can also be tested with brain imaging experiments [3]. When imaging studies and behavioral studies are used in combination to test the same DIVA-based hypotheses, they provide a valuable means of quantifying relations among phonemic specifications, brain activity, articulatory movements and the speech sound output

5. ACKNOWLEDGEMENTS

The work from our laboratory that is described in this paper was done in collaboration with a number of colleagues, including Frank Guenther, Harlan Lane, Melanie Matthies, Mark Tiede, Majid Zandipour, Margaret Denny, Jennell Vick and Virgilio Villacorta. Support was from grants R01-DC001925 and R01-DC003007 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

6. REFERENCES

[1] Fujimura, O. & Kakita, Y. 1979. Remarks on quantitative description of lingual articulation. In: Lindblom, B., Öhman, S. (eds.) *Frontiers of Speech Communication Research*. San Diego: Academic Press, 17-24.

[2] Guenther, F.H., Espy-Wilson, C., Boyce, S.E., Matthies, M.L., Zandipour, M., Perkell, J.S. 1999. Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *J. Acoust. Soc. Am.* 105, 2854-2865.

[3] Guenther, F.H., Ghosh, S.S., Tourville, J.A. 2006. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language* 96, 280-301.

[4] Houde, J.F., Jordan, M.I. 1998. Sensorimotor adaptation in speech production. *Science* 279, 1213-1216.

[5] Perkell, J.S., Guenther, F.H., Lane, H., Matthies, M.L., Perrier, P., Vick, J., Wilhelms-Tricarico, R., Zandipour, M. 2000. A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *J. Phonetics* 28, 233-272.

[6] Perkell J.S., Guenther F.H., Lane, H., Matthies, M.L., Stockmann, E., Tiede, M., Zandipour, M. 2004. The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts, *J. Acoust. Soc. Am.* 116, 2338-2344.

[7] Perkell, J.S., Lane, H., Denny, M., Matthies, M.L., Tiede, M., Zandipour, M., Vick, J., Burton, E. 2007. Time course of speech changes in response to short-term changes in hearing state, *J. Acoust. Soc. Am.* 121, 2296-2311.

[8] Perkell, J.S., Matthies, M.L., Svirsky, M.A., Jordan, M.I. 1993. Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot motor equivalence study. *J. Acoust. Soc. Am.* 93, 2948-2961.

[9] Perkell J.S., Matthies, M.L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., Stockmann, E., Guenther, F.H. 2004. 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.

[10] Perkell, J.S., Nelson, W.L. 1985. Variability in production of the vowels /i/ and /a/. *J. Acoust. Soc. Am.* 77, 1889-1895.

[11] Purcell, D.W., Munhall, K.G. 2006. Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *J. Acoust. Soc. Am.* 120, 966-977.

[12] Stevens, K. N. 1989. On the quantal nature of speech. *J. Phonetics* 17, 3-46.

[13] Stevens, K.N. 1998. *Acoustic Phonetics*. Cambridge, MA: MIT Press.

[14] Svirsky, M.A., Lane, H., Perkell, J.S., Wozniak, J. 1992. Effects of short-term auditory deprivation on speech production in adult cochlear implant users. *J. Acoust. Soc. Am.* 92, 1284-1300.

[15] Tourville, J.A., Guenther, F.H., Ghosh, S.S., Reilly, K.J., Bohland, J.W., Nieto-Castanon, A. 2005. Effects of acoustic and articulatory perturbation on cortical activity during speech production. Proc. 11th Annual Meeting of the Organization for Human Brain Mapping, S49.

[16] Vick, J., Lane, H., Perkell, J.S., Matthies, M.L., Gould, J., Zandipour, M. 2001. Speech perception, production and intelligibility improvements in vowel-pair contrasts in adults who receive cochlear implants. *J. Speech, Lang. Hear. Res.* 44, 1257-1268.

[17] Villacorta, V., Perkell, J.S., Guenther, F.H. 2004. Sensorimotor adaptation to acoustic perturbations in vowel formants. *J. Acoust. Soc. Am.* 115, 2430 (A).

[18] Villacorta, V., Perkell, J.S., Guenther, F.H. 2005. Relations between speech sensorimotor adaptation and perceptual acuity. *J. Acoust. Soc. Am.* 117, 2618-2619 (A).