

## SENSORY GOALS FOR SPEECH MOVEMENTS: CROSS-SUBJECT RELATIONS AMONG PRODUCTION, PERCEPTION AND THE USE OF AN ARTICULATORY SATURATION EFFECT

Joseph Perkell<sup>1,2,3</sup>, Melanie Matthies<sup>4,1</sup>, Frank Guenther<sup>3,1</sup>, Mark Tiede<sup>1,5</sup>, Majid Zandipour<sup>1,3</sup>, Ellen Stockmann<sup>1</sup> & Nicole Marrone<sup>4,1</sup>

<sup>1</sup> Research Laboratory of Electronics, M.I.T., Cambridge, MA USA

<sup>2</sup> Dept. of Brain and Cognitive Sciences, M.I.T

<sup>3</sup> Dept. of Cognitive and Neural Systems, Boston University, Boston, MA

<sup>4</sup> Dept. of Health Sciences, Program in Communication Disorders, Boston University

<sup>5</sup> Haskins Laboratories, New Haven, CT

**ABSTRACT:** This paper describes two studies of 19 young-adult speakers of American English in which measures were made of: (a) *phoneme contrast* – articulatory distance between vowels in each of two vowel pairs and acoustic distance between the sibilants /s/ and /ʃ/, (b) *auditory discrimination* of tokens from synthetic vowel and sibilant continua, and (c) *contact* of the tongue tip with the lower incisors for /s/. In the vowel study, an articulatory measure of contrast distance was correlated across subjects with auditory discrimination. In the sibilant study, acoustic contrast distance was related across subjects to auditory discrimination and also to the use of contact. These findings are compatible with the DIVA model of speech motor planning, in which the goals for phonemic speech movements are in auditory and somatosensory spaces.

### INTRODUCTION

The two studies described below are based on the hypothesis that the goals of phonemic speech movements are in auditory and somatosensory spaces. Such phonemic goals are an essential characteristic of DIVA, Guenther's model of speech motor planning (cf. Guenther & Ghosh, 2003; Guenther, et al., current proceedings), in which production and perception are closely linked. In order to probe this linkage, we have performed two studies of relations between production and perception.

Recent brain imaging studies have provided evidence supporting the hypothesis of an intimate relation between speech production and speech perception. Investigators have shown that motor areas of the brain are active during speech perception (cf. Rizzolatti & Arbib, 1998) and auditory areas are active during speech production (cf. Hickok and Poeppel, 2000). The studies described below address this hypothesis in another way, by seeking correlations between measures of production and perception across speakers. Specifically, we hypothesize that speakers who discriminate well between phonetic stimuli with subtle acoustic differences will produce sound contrasts that are relatively clear-cut while speakers who discriminate less well between the same stimuli will produce less distinct contrasts.

In DIVA, planning of movements to achieve phonemic goals is influenced by the listener's need for *clarity*, competing with the speaker's motivation to achieve an economy of effort (cf. Guenther, 1995; Lindblom, 1986). Clarity can be defined as the distinctness of auditory goal regions for different sounds, which is in turn determined by the size and location of those regions in auditory space. The model forms goal regions by monitoring the sounds of the language, and learning, for each phoneme, the region that encompasses all the examples of that sound. According to the model's functionality, speakers who can perceive fine acoustic details will learn goal regions that are smaller than speakers with less acute perception, because speakers with acute perception are more likely to reject poorly produced tokens when learning the goal regions.

### STUDY 1: PRODUCTION AND PERCEPTION OF VOWEL CONTRASTS (Perkell et al., submitted-a)

#### *Production experiment*

The subjects for this and the subsequent studies were 19 young adult speakers of American English, 10 males and 9 females. Each subject participated in a production experiment in which we recorded

Proceedings of the 6th International Seminar on Speech Production, Sydney, December 7 to 10, 2003.

articulatory movements with our EMMA system and the acoustic signal. The subjects pronounced multiple repetitions of utterances of the form “Say \_\_\_ hid it.”, where \_\_\_ = *cod*, *cud*, *who’d* or *hood*, in normal and fast conditions. We extracted vowel formants and positions of an EMMA transducer on the tongue blade at the point of minimal articulator speed during the vowel (the vowel target).

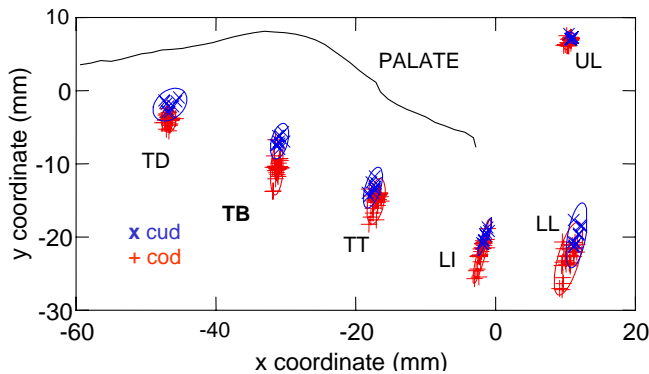


Figure 1: Locations of EMMA transducers – TD: Tongue dorsum, TB: Tongue blade, TT: Tongue tip, LI: Lower Incisors, LL: Lower lip, UL: Upper lip

Figure 1 shows an example of articulatory data from a female subject, for 27 repetitions of /a/ (+) and 9 repetitions of /ʌ/ (x) in the normal condition. Our focus is on data from the TB (tongue blade) transducer, because it evidenced the least amount of coarticulation from the surrounding sounds. From such articulatory data and the values of F1 and F2, we calculated “category separation” for each vowel pair. TB (articulatory) category separation is the distance in mm between the centroids of the TB /a/ and /ʌ/ distributions. Acoustic category separation is the distance in Hz in the formant plane between the distributions of F1 vs. F2 for /a/ and /ʌ/.

*Perception experiment*

Based on natural productions of the end-point utterances, natural-sounding stimuli were synthesized (with the Klatt synthesizer) in 7-step continua for *cod* to *cud* and *who’d* to *hood*. Each subject gave responses to the stimuli in labeling and ABX discrimination tasks. The discrimination tests used stimulus pairs (for stimuli A and B) that were 1, 2 and 3 steps apart on the continua. Both the slopes of labeling functions in their transitions from perceived *cod* to *cud* and *who’d* to *hood* and also peak discrimination scores varied considerably across subjects. The results from the 2-step ABX stimuli were used for further analysis, because discrimination of tokens separated by three steps on the synthetic speech continuum very often reached 100 percent while that of tokens separated by one step was at chance or near-chance levels in some subjects.

*Relations between production and perception*

Figure 2 shows values of TB category separation vs. peak 2-step ABX percent correct for the *who’d-hood* utterances (left panels) and the *cod-cud* utterances (right panels), with results from the normal condition in the bottom panels and the fast condition in the top panels. Each point shows the results from one subject (numbered 1-19).

Even among the 2-step ABX scores there are some ceiling effects (values at 100%), indicating that those subjects may have had better discrimination than this task measured. For this reason, and because the *who’d-hood* distributions are right skewed, the subjects were divided into two groups for further analysis: For each combination of vowel pair and speaking condition, subjects with peak 2-step discrimination scores of 100% were categorized as HI discriminators. The remaining subjects, with percent correct < 100, were categorized as LO discriminators. The *point biserial r* was used to measure the relation between the categorized ABX measures and the continuous values of TB category separation.

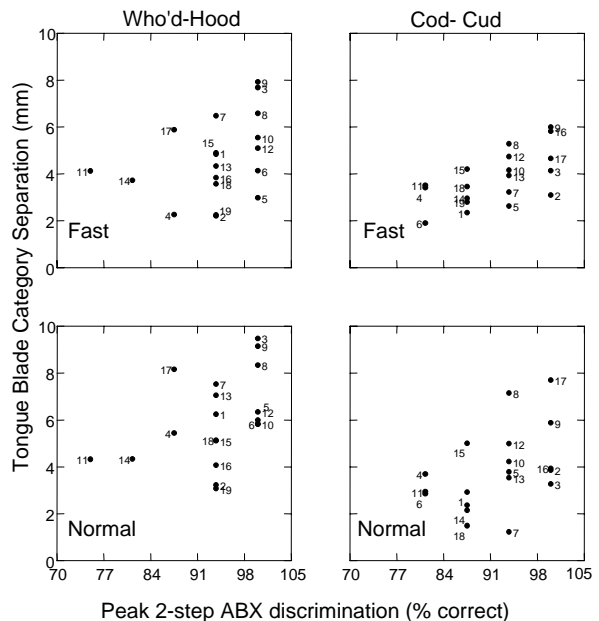


Figure 2: TB category separation vs. peak 2-step ABX score. Each point shows the mean of 40 ABX determinations and 27 measures of TB category separation for one subject (1-19).

Figure 3 shows values of TB category separation vs. ABX discrimination group (LO vs. HI). The panels are arranged as in Fig. 2, and the values of  $r$ ,  $t$  and  $p$  ( $df = 17$ , one-tailed) are from point biserial correlations. The figure shows that HI-discrimination subjects read the test words with greater TB category separation than LO-discrimination subjects. Correlations were statistically significant in three of the four cases (one tailed tests). Correlations of acoustic category separation (in the formant plane) with discrimination group were significant in only one of the four cases. This may be because formants do not provide sensitive enough measures of the auditory cues being used by the speakers (see Discussion and Conclusions).

Pearson product moment correlations of TB category separation for *who'd-hood* with category separation for *cod-cud* were significant ( $p < .05$ ) for both normal ( $r = 0.45$ ,  $df=17$ ,  $p < .05$  one-tailed) and fast utterances ( $r = 0.55$ ,  $df=17$ ,  $p < .01$  one-tailed). Since the speakers behaved similarly in producing the two contrasts, the data were combined, after first converting them to z-scores. Point biserial correlations for the combined data were significant for both the normal ( $r = 0.54$ ,  $t = 2.63$ ,  $p < .01$  one tailed) and the fast conditions ( $r = 0.58$ ,  $t = 2.90$ ,  $p < .01$ , one-tailed). Thus, the more accurately a speaker discriminates a vowel contrast, the more distinctly the speaker produces that contrast.

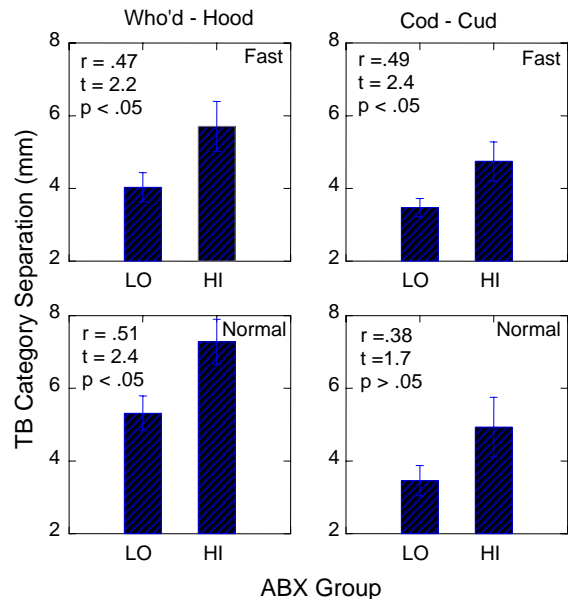


Figure 3: TB category separation (vertical axes) vs. ABX group (horizontal axes – LO, HI discriminators).  $r$ ,  $t$  and  $p$  values are from point biserial  $r$  correlations. The error bars show  $\pm$  one standard error about the mean.

## STUDY 2: PRODUCTION AND PERCEPTION OF THE /s-/j/ CONTRAST (Perkell et al., submitted-b)

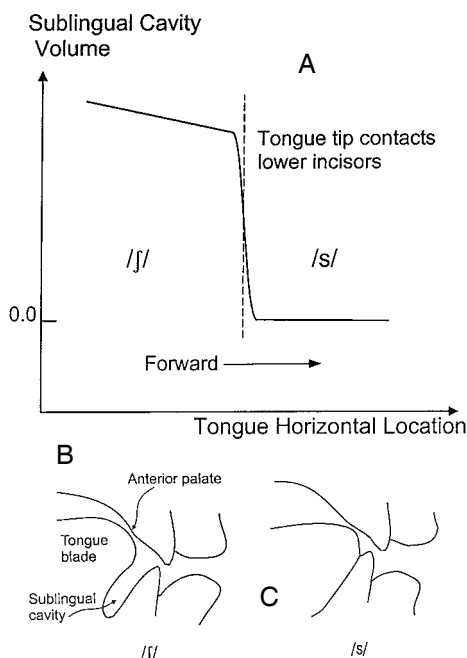


Figure 4: Schematic illustration of the use of a saturation effect to facilitate production of the /s-/j/ contrast.

/j/ configuration, the sublingual cavity volume gradually decreases; then when contact is made, the volume suddenly drops to zero.

We have stated above that the DIVA model posits goals for phonemic speech movements in auditory and somatosensory spaces. To explore the role of the latter and its relation to discriminative capacity, we examined the sibilant contrast, /s-/j/. Many consonants may have important somatosensory goals. For example, somatosensory goals for stop consonants could be patterns of articulator contact. The sibilant sounds /s/ and /j/ may have both kinds of goals. We hypothesize that the auditory goals for sibilants consist of particular distributions of energy in the noise spectrum; the somatosensory goals could consist of patterns of contact of the tongue blade with the palate and teeth.

Figure 4 schematizes how one somatosensory goal for /s/ may be determined by a "saturation effect" that helps to define the /s-/j/ contrast. As schematized in panel B, /j/ is produced by positioning and shaping the tongue blade so there is a relatively long, narrow groove between the tongue blade and palate, and a resulting space between the lower incisors and under side of the tongue blade, called a "sublingual cavity". In contrast, /s/ is produced (panel C) by forming a shorter groove with the tongue blade in a more anterior position, so there is contact between the tongue tip and lower incisors and no sublingual cavity. Panel A shows that as the tongue blade is gradually moved forward from a

Since the frequency distribution of the radiated noise for the sibilants depends partly on the size of the resonant cavity anterior to the constriction, the elimination of the sublingual cavity causes a sudden increase in the spectral center of gravity from one that is characteristic of /ʃ/ to one that is characteristic of /s/. Obviously, once contact is made, a further increase in the activity of the muscles pushing the tongue forward cannot cause any significant additional change in the sibilant noise spectrum. We call this nonlinear (quantal) relation between (motor commands underlying) articulatory movement and the resulting area function and acoustic output a “saturation effect” (cf. Stevens, 1989; Fujimura and Kakita, 1979).

Based on this background and the results of Study 1, we hypothesize the following. 1) Speakers will use the saturation effect just described. This will be evidenced by consistent contact of the tongue tip with the lower incisors during /s/ – to help differentiate /s/ from /ʃ/, and speakers will vary in this respect. 2) Speakers will vary in their ability to discriminate /s/ from /ʃ/. 3) Across speakers, both factors, use of tongue contact (in effect, a somatosensory goal) and ability to discriminate auditorily between the two sounds will predict the strength of the produced contrast – measured acoustically.

### Production experiment

The same 19 subjects (plus an additional female) participated in this production experiment. Each subject pronounced multiple repetitions of “Say \_\_\_ hid it” (where \_\_\_ = *sod*, *shod*, *said* or *shed*) while recordings were made of the acoustic signal and of contact between the underside of the tongue tip and the lower incisors (using a custom-made sensor and electronics). The contact signal was used to calculate the proportion of time contact was made during the sibilant interval. Acoustic spectra were derived from the speech signal during the sibilant, and the spectral mean (center of gravity or COG) was calculated. The acoustic contrast distance between the phonemes of each subject was quantified with  $d'$  calculated from the COG values for the two phonemes.

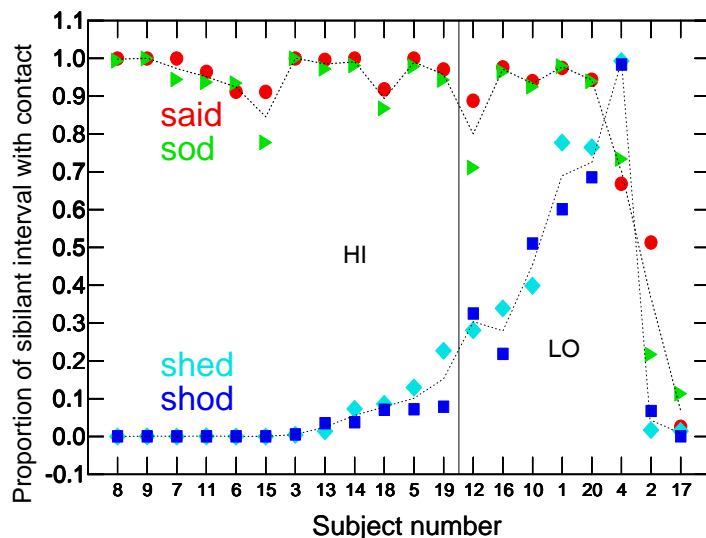


Figure 5: Proportion of sibilant interval with contact in each of 19 subjects. (See text for details.)

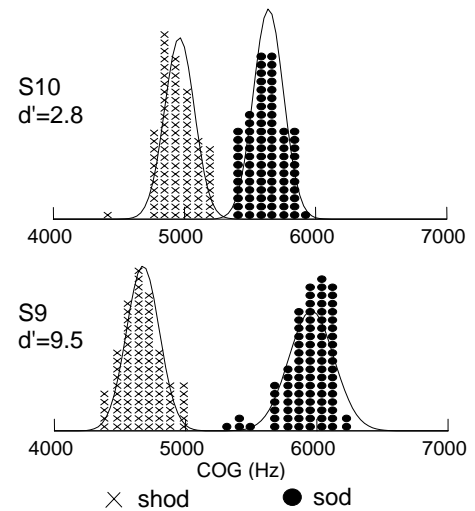


Figure 6: Frequency distributions of spectral center of gravity for tokens of two sibilants and their associated  $d'$

Figure 5 shows, for all subjects, average values of the proportion of sibilant duration in which there was tongue contact. As anticipated, for most subjects, the proportion was close to 1.0 for /s/ and close to 0.0 for /ʃ/. For each subject, “contact difference” was calculated as the difference in proportion between /s/ and /ʃ/, averaged across the two vowel environments. The 12 subjects with a contact difference greater than 0.81 were classified as HI for contact difference (left side of Fig. 5); the remaining 8 subjects were classified as LO for contact difference. For the LO subjects, the mean contact difference was  $\leq 0.7$ , and the variability increased sharply (compared to the HI subjects) so that the distributions of the contact measures for /s/ and /ʃ/ for the LO subjects began to overlap.

Figure 6 shows examples of distributions of values of spectral mean (“Center of Gravity,” COG) for the sibilants in *sod* and *shod* for two subjects, along with the corresponding values of  $d'$ . These two

subjects have among the most extreme values of  $d'$ . Even though all the subjects produced normal-sounding sibilants, the strength of the contrast varied considerably.

### Perception experiment

The original 19 subjects performed labelling and discrimination (ABX) tasks with natural-sounding synthetic stimuli ranging in seven steps from *said* to *shed* (as with the vowel stimuli). As in Study 1 with vowels, the 3-step ABX stimulus pairs of synthetic sibilants yielded an excessive number of ceiling effects and the 1-step pairs produced several cases of performance near chance, so the 2-step data were used for further analysis. Eleven of the subjects had peak values of 2-step percent correct = 100; they were categorized as HI discriminators. The remaining subjects, with percent correct < 100, were categorized as LO discriminators, as in Study 1.

### Contrast separation vs. the use of contact difference and auditory discrimination

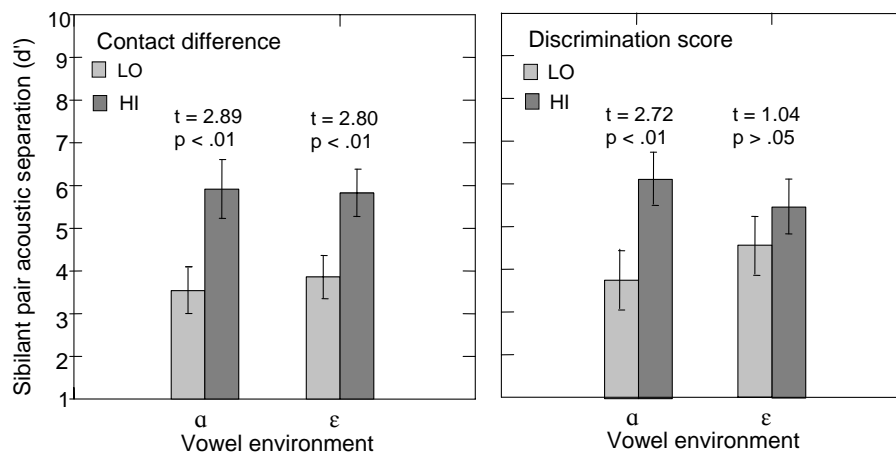


Figure 7: Sibilant pair acoustic separation ( $d'$ ) vs. contact difference group (left panel) and discrimination score group (right panel) for each vowel environment. The error bars show +/- one standard error about the mean.

Figure 7 shows values of  $d'$  for acoustic category separation (vertical axes) vs. contact difference group (LO, HI – left panel) and discrimination score group (LO, HI – right panel) for the sibilants spoken in the two vowel environments. The results of two-sample t-tests (one-tailed) are given above each pair of bars. It is evident that the strength of the produced contrast is positively related to both contact difference and ability to discriminate the contrast. (Note that subjects with HI contact difference values did not necessarily have HI discrimination scores and vice versa.) We used multiple linear regression across speakers to predict the strength of the produced acoustic contrast ( $d'$  averaged across the two vowel environments) from individual subjects' values of contact difference and 2-step ABX score. The result was significant ( $R = 0.65$ ;  $p < .05$ ), with approximately equal weight for each predictor (contact difference:  $\beta = 0.56$ ,  $t = 2.82$ ,  $p < .01$ ; peak ABX:  $\beta = 0.50$ ,  $t = 2.54$ ,  $p < .05$ ). These results show that both factors, use of the saturation effect and discrimination ability, are related to how well the speakers produce the contrast.

We calculated cross-subject correlations between measures of vowel and consonant contrast. Among the several possible correlations, only one was significant, indicating that speakers who produced the best vowel contrasts did not necessarily produce the strongest sibilant contrasts. Several factors could underlie this dissociation, including subject differences in the perception of vowel and sibilant sounds and in the relative weighting of auditory vs. somatosensory goals in vowel vs. sibilant production.

## DISCUSSION AND CONCLUSIONS

The studies of vowels and sibilants described above have demonstrated cross-speaker relations between production and perception. In the sibilant study, speakers with more acute sibilant discrimination tended to produce the sounds with greater contrast. This finding is compatible with other recent work (see Newman, 2003 and references in Perkell, *et al.*, submitted a, b) and with the way the Proceedings of the 6th International Seminar on Speech Production, Sydney, December 7 to 10, 2003.

DIVA model learns and uses auditory goals. Speakers who can perceive fine acoustic details will learn goal regions that are smaller (with edges that are spaced further apart) than speakers with less acute perception. Those with acute perception are more likely to reject poorly produced tokens of a phoneme when learning the goal regions. Presumably, this occurs because speakers find that greater clarity is advantageous. The observed cross-speaker relation between produced sibilant contrast and the measure of contact difference is also compatible with DIVA, in its use of somatosensory goals.

In the vowel study, the articulatory measure of contrast was more strongly related to auditory discrimination than the acoustic measure of contrast. The DIVA model predicts that the productions of the better discriminators should be spaced further apart in both articulatory and acoustic spaces. The stronger correlation between discrimination scores and articulations (rather than acoustics) might be interpreted as support for a theory in which basic phonemic units consist of articulatory, rather than auditory, gestures (cf. Browman and Goldstein, 1989). However we believe this result occurs because formant frequencies *per se* are not a sufficient acoustic measure of perceptual distinctiveness; many other aspects of the acoustic signal, such as fundamental frequency, formant ratios, and overall spectral shape, have been shown to affect listeners' perceptions in addition to the absolute magnitudes of formant frequencies.

#### ACKNOWLEDGEMENT

This work was supported by Grant no. R01-DC01925 from the National Institute on Deafness and other Communication Disorders, National Institutes of Health.

#### REFERENCES

- Browman, C.P. and Goldstein, L. (1989). "Articulatory gestures as phonological units" *Phonology* **6** 201-251.
- Fujimura, O. & Kakita, Y. (1979). "Remarks on quantitative description of lingual articulation" In B. Lindblom and S. Öhman (eds.) *Frontiers of Speech Communication Research*, Academic Press, London.
- Guenther, F.H. (1995). "Speech sound acquisition, coarticulation and rate effects in a neural network model of speech production" *Psychological Review* **102** 594-621.
- Guenther, F.H. & Ghosh, S. (2003). "A model of cortical and cerebellar function in speech" *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 169-173.
- Hickok, G. & Poeppel, D. (2000). "Towards a functional neuroanatomy of speech perception" *Trends in Cognitive Science* **4** 131-138.
- Lindblom, B. (1986). "Phonetic universals in vowel systems" In: Ohala JJ, Jeager JJ (eds.) *Experimental Phonology*. Academic Press, pp. 13-44.
- Newman, R.S. (2003) "Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report" *Journal of the Acoustical Society of America* **113** 2850-2860.
- Perkell, J.S., Guenther, F.H., Lane, H., Matthies, M.L., Stockmann, E., Tiede, M. & Zandipour, M. "The distinctness of speakers' productions of contrasting vowels is related to the acuity of their discrimination of the contrasts" (submitted to the *Journal of the Acoustical Society of America*).
- Perkell, J.S., Matthies, M.L., Tiede, M. Zandipour, M., Stockmann, E., Marrone, N., Lane, H., & Guenther, F.H. "The distinctness of speakers' productions of /s/ and /ʃ/ is related to their discrimination of the contrast and use of an articulatory saturation effect" (submitted to the *Journal of Speech, Language and Hearing Research*).
- Rizzolatti, G. & Arbib, M.A. (1998). "Language within our grasp" *Trends in Neuroscience* **21** 188-194.
- Stevens, K.N. (1989). "On the quantal nature of speech" *Journal of Phonetics* **17** 3-45.

Proceedings of the 6th International Seminar on Speech Production, Sydney, December 7 to 10, 2003.