Cross-subject relations between measures of vowel production and perception

Joseph S. Perkell,[a, b] Frank H. Guenther,[b, a] Harlan Lane,[c, a] Melanie L. Matthies,[d, a] Ellen Stockmann,[a] Mark Tiede,[a, e] and Majid Zandipour[a]

a – Research Laboratory of Electronics, Massachusetts Institute of Technology, Boston, MA

b – Department of Cognitive and Neural Systems, Boston University, Boston, MA

c – Department of Psychology, Northeastern University, Boston, MA

d – Department of Communication Disorders, Boston University, Boston, MA

e – Haskins Laboratories, New Haven, CT

ABSTRACT

This study addresses the hypothesis that the more accurately a speaker discriminates a vowel contrast, the more distinctly the speaker produces that contrast. Measures of speech production and perception were collected from 19 young adult speakers of American English. In the production experiment, speakers repeated the words *cod, cud, who'd* and *hood* in a carrier phrase at normal and fast rates. Articulatory movements and the associated acoustic signal were recorded, yielding measures of contrast distance between /ɑ/ and /ʌ/ and between /u/ and /ʊ/. In the perception experiment, sets of seven stimuli ranging from *cod* to *cud* and *who'd* to *hood* were synthesized, based on natural productions by one male and one female speaker. The continua were presented to each of the speakers in labeling and discrimination tasks. Consistent with the hypothesis, measures of produced vowel contrast were correlated across subjects with a measure of vowel discrimination. This finding is compatible with a model in which articulatory movements for vowels are planned primarily in auditory space.

1. INTRODUCTION

Recent brain imaging studies have provided evidence supporting the hypothesis of an intimate relationship between speech production and speech perception. Investigators have shown that motor areas of the brain are active during speech perception cf. [1] and auditory areas are active during speech production cf. [2].

The current study addresses 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 vowel stimuli with subtle acoustic differences will produce vowel contrasts that are relatively clear-cut while speakers who are less able to discriminate between the same vowel stimuli will produce less clear-cut vowel contrasts.

This hypothesis is based on a model of speech production in which goals for vowel movements are regions in multi-dimensional auditory-temporal space [3,4,5]. In this model, DIVA, speech motor planning is influenced by two competing constraints: the listener's need for clarity and the speaker's motivation to achieve an economy of effort [6]. The degree of clarity is related to the amount of separation among the auditory goal regions for different sounds, which is determined by their location and size in auditory space. The model forms goal regions initially by monitoring sounds from the speaker's native language and learning, for each phoneme, the region of auditory space that encompasses examples of that phoneme [4]. According to DIVA, speakers who can perceive fine acoustic details will learn goal regions that are smaller and spaced further apart because they are more likely than people with less acute perception to reject poorly produced tokens of a phoneme when learning the goal regions.

2. BACKGROUND

In prior studies, experimenters have examined speakers who varied in measures of production and perception and found relationships between the two types of measures, cf. [7,8,9]. Most of these studies are consistent with the idea that speakers who have relatively sensitive perceptual capabilities produce more distinct sound contrasts. The current study investigates this idea in more detail, using articulatory, acoustic and perceptual measures, with an explicit, model-based hypothesis: subjects who have relatively higher vowel discrimination scores will produce sharper vowel contrasts than those who do not.

3. METHODS

Subjects: Production and perception experiments were performed on a group of 19 young adult speakers of American English, 9 females and 10 males who had no history of speech or hearing disorders.

Production Experiment: Each subject participated in a speech production experiment, in which his or her articulatory movements and speech signal were recorded.

The speech materials consisted of the words *cod, cud, who'd* and *hood* embedded in the phrase, "Say ___ hid it." There were 27 repetitions of each *cod* and *who'd* utterance

ISBN 1-876346-48-5

and 9 repetitions of each *cud* and *hood* utterance (which were included initially as foils for a slightly different design). The corpus was read in two different conditions, "normal" and "fast." For the fast condition, the subject was asked to speak as rapidly as possible without eliminating any sounds.

An electromagnetic midsagittal articulometer system (EMMA – [10]) was used to record the position versus time of small transducer coils mounted on the subject's tongue lips and jaws in the midsagittal plane. Transducers were attached with biocompatible adhesive to the vermillion border of upper lip (UL) and lower lip (LL), the gingival papilla between the lower central incisors (LI), and three places on the tongue, 1 cm from the tongue tip (TT), the tongue blade (TB – about 3 cm from the tongue tip) and the tongue dorsum (TD – about 5 cm back from the tongue tip). Transducers were also attached to the bridge of the nose and the gingival papilla between the upper central incisors for a maxillary frame of reference. A custom-written program was used to control the experiment, record the movement and acoustic signals to disk and display the utterance materials one at a time on an LCD screen located about three feet in front of the subject.

The data of interest were the x and y positions of a transducer coil on the tongue at the time the tongue reached its vowel target position, and measures of the corresponding acoustic spectrum. As the first step in data extraction, one of the experimenters labeled the beginning and end of the vowel for each token, using an interactive display of the acoustic waveform and spectrogram.

The remaining data extraction procedures were performed algorithmically. For reasons explained below, the TB transducer was chosen as most representative of the tongue body position for the vowel. TB "target" x and y coordinates were extracted at the time of the minimum in velocity magnitude during the vowel. The first three formant frequencies were extracted from the acoustic signal with a method designed to minimize the occurrence of missing or spurious values.

Two measures of category separation, i.e., vowel contrast, were derived for each vowel pair, /ɑ-ʌ/ and /u-ʊ/, in each of the two conditions, normal and fast. To derive articulatory category separation, the mean values of TBx and TBy were calculated for each vowel and condition. Articulatory separation for the pair was calculated as the Euclidian distance between the points defined by mean TBx and TBy values. Correspondingly, for formant category separation, mean values of F1 and F2 were calculated, and formant separation was calculated as the Euclidian distance between the means in the F1, F2 plane.

Perception experiment. The same 19 subjects also participated in perception experiments, consisting of vowel labeling and discrimination tasks.

Two continua of seven stimuli were synthesized for each of the two word pairs, *cod-cud* and *who'd-hood*, based on natural productions of the words in isolation by a male speaker and a female speaker. Values of the first three formants at 10 ms intervals during the vowel were calculated by interpolation between the naturally produced end-point values. The fourth and fifth formants were kept constant. For each continuum, the vowel portions were synthesized with the same duration and the same naturally produced F0 contour. For each stimulus, the synthesized vowel was inserted between the same naturally produced initial-consonant release and final-consonant voicing signal to avoid discontinuities in the waveform. The resulting stimuli sounded quite natural in informal listening tests.

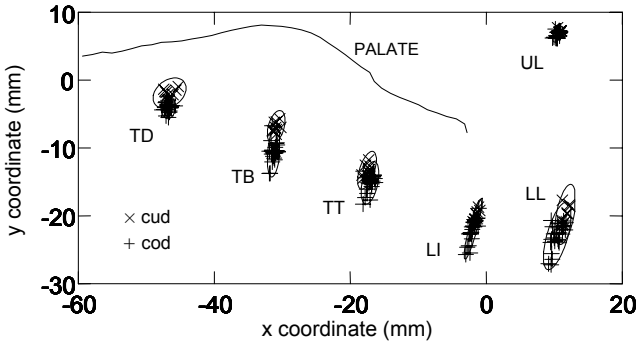

Figure 1: transducer coil locations.

Stimuli from all four of the continua were presented to each of the subjects in labeling and discrimination tasks. The labeling test was administered first, then the discrimination test. Stimuli were blocked by word pair, and all tasks were subject-paced. In the labeling task, each stimulus was presented individually and the subject was asked to identify the word using a computer mouse to select from the two choices on the monitor (*cod* or *cud*, *who'd* or *hood*). Each of the 7 stimuli was presented 18 times. The discrimination task was a classic ABX design. Stimuli were grouped into 60 sets of three stimuli where the first and second were 1, 2, or 3 steps apart on the synthesis continuum and the third was the same as either the first or second. After each set was played, the subject decided whether the third stimulus was the same as the first or second and indicated the decision by selecting a button on the computer screen.

4. RESULTS

Figure 1 is a plot of the EMMA transducer coil locations during all the productions of tokens containing the /ɑ/ in *cod* (+) and the /ʌ/ in *cud* (×) by a female subject in the normal speaking condition. The transducer coils are located at points on the tongue dorsum (TD), tongue blade (TB), tongue tip (TT), lower incisor, (LI), lower lip (LL) and upper lip (UL). It was determined that the middle of the three tongue transducers, TB , represented the /ɑ/ target location with the least amount of coarticulatory influence of the preceding /k/ and the following /d/. Therefore, the analysis of the articulatory target for the /ɑ/ in *cod* and the /ʌ/ in *cud* focused on the location of the TB transducer. For uniformity across the four test words the same transducer coil was used to represent the tongue body location during the /u/ in *who'd* and the /ʊ/ in hood. In this example, the TB category separation is the distance in mm between the centroids of the *cod* and *cud* TB distributions.

ISBN 1-876346-48-5

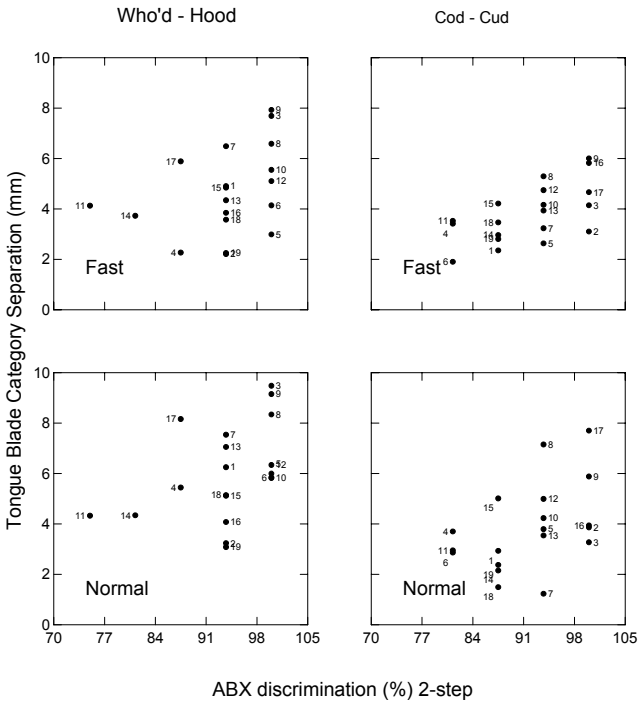

Figure 2: TB category separation (mm) vs.2-step ABX discrimination (%)

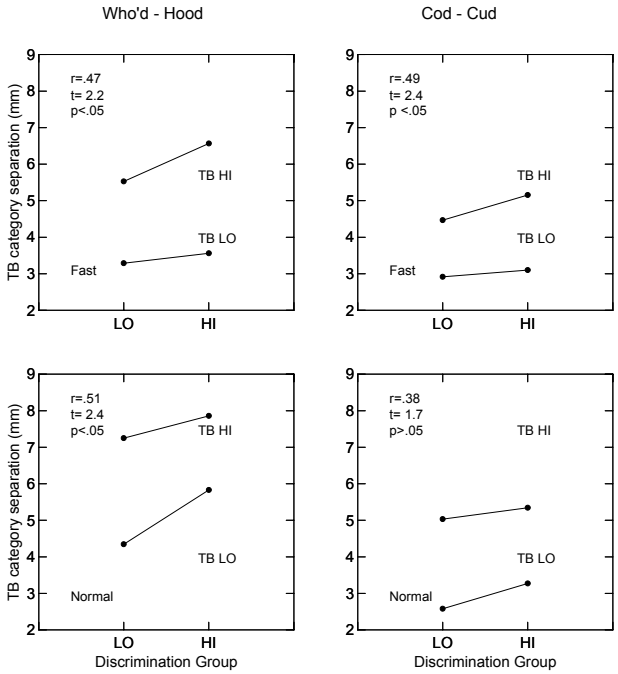

Figure 3: TB category separation (mm) vs. peak 2-step ABX discrimination (%)

In the perception results, there was considerable variation among the subjects in the steepness of the slope of their labeling functions and in the shapes and peak amplitudes of their discrimination functions. Many of the 3-step discrimination functions evidenced pronounced ceiling effects. Many 1-step intervals did not cross the labeling category boundary and their peak discrimination scores did not correlate significantly with TB category separation. For these reasons we report on the peak 2-step scores.

Figure 2 presents plots of subjects' TB category separation (mm – vertical axis) vs. their peak 2-step discrimination scores (percent correct – horizontal axis). The top two panels show results from the fast condition; the bottom two panels, from the normal condition; the left panels for *who'd-hood*; and the right panels for *cod-cud*. Each point represents mean values for one subject; the subjects are labeled 1-19. Across subjects, vowel pairs and conditions, values of TB separation range between about 2 and 10 mm. The peak discrimination scores range from 75 to 100%.

For the following reasons, we classified subjects into two groups based on their peak discrimination scores. There were substantial ceiling effects, with 37% of the subjects (*who'd-hood*) or 26% of the subjects (*cod-cud*) scoring 100 percent. All subjects' scores fell at one of five values (*who'd-hood*) or 1 of 4 values (*cod-cud*) between 75 and 100 % correct. (Scores varied in increments of 6.25% because there were 16 discrimination trials per two-step stimulus comparison.) Finally, the distribution for *who'd-hood* was right-skewed. These considerations and the observation that production contrast appeared to grow nonlinearly with phoneme discrimination led us to classify all subjects who scored 100% as HI discriminators. Some in this group presumably had greater underlying discriminative capacity than their scores indicated, but its expression was limited by the ceiling effect. All others subjects were classified as LO discriminators. (This division coincided with groupings based on a median split of the ABX 2-step scores.) On the other hand, the frequency distributions of production contrast were distributed roughly normally over the set of 19 speakers. For the combination of dichotomous discrimination and continuous separation variables, the point biserial r was then employed as a measure of the relation between perception and production. This measure does not assume normality of the dichotomous variable [11]. The null hypothesis that the correlation is zero is evaluated using Student's t distribution.

Figure 3 plots TB category separation (mm – vertical axis) against the bimodal value (HI or LO) of the subjects' 2-step discrimination scores for *who'd-hood* (left panels) and *cod-cud* (right panels) in the fast condition (upper panels) and normal condition (lower panels). The values of TB category separation are shown averaged separately for a high separation group (TB HI) and a low separation group (TB LO), based on a median split of TB category separation values. Within each TB group, HI discrimination subjects produced greater phoneme category separation than LO discrimination subjects. Each panel contains the results (r, t and p) of the point biserial correlation calculated across the combined results from the two TB groups. The correlation is positive and significant in three of the four cases ($p < .05$), with normal rate *cod-cud* the exception.

Point biserial correlations were calculated for formant contrast distance versus discrimination category (HI, LO) in the same way. Three of these four correlations were not significant; for *cod-cud* in the fast condition, however, $r = .60$ ($t = 3.06$, $df = 17$, $p < .01$).

ISBN 1-876346-48-5

Correlation analyses were calculated for TB category separation for *who'd-hood* (mm) versus TB category separation for *cod-cud* (mm). The correlations were reliable in both the normal ($r = .45$, $p < .05$) and fast ($r = .55$, $p < .05$) condition data, indicating that the more a speaker separated tongue positions for contrasting *cod* and *cud*, the more that speaker did likewise for *who'd* and *hood*.

Since production measures for the two vowel contrasts were correlated, the correlation coefficients relating discrimination to production were computed for the two contrasts pooled. To pool those results in each speaking condition for each speaker, that speaker's two discrimination scores (one for each contrast) were converted to standardized (z) scores using the set of data for the 19 subjects. Next each speaker's two discrimination z scores were summed and high and low discriminators were identified with a median split (this time on z scores). Then, each speaker's two scores for TB category separation were converted to z scores and summed. Finally, the point biserial r was computed as before, relating ABX discrimination class (HI or LO) to TB category separation — this time using standard scores. The point biserial r at fast rate relating discrimination scores to the two production contrasts pooled was $r = .54$ ($t = 2.63$, $p < .05$); at normal rate, it was $r = .58$ ($t = 2.90$, $p < .01$).

5. DISCUSSION

We have found, for two vowel contrasts, that the more accurately a speaker discriminates a contrast, the more distinctly the speaker produces that contrast.

The finding of cross-speaker correlations of an articulatory measure of vowel contrast with vowel discrimination scores, largely in the absence of correlations involving acoustic measures, is consistent with the results of some other studies [12,13], which showed that listener ratings of productions can reveal relationships between production and perception when acoustic measures of production do not. The relatively low-dimensional acoustic measures used in the current study (equivalent to a single time slice during the time-varying vowel) may not adequately capture the kinds of differences that the speakers are attending to auditorily when producing sound contrasts.

Our general hypothesis, that perception influences production, is compatible with the DIVA model [3,4,5], in which the basic phonemic units for vowels are multidimensional regions in auditory-temporal space. These regions are utilized in speech perception and they are also goals for the planning of articulatory movements. As explained in the introduction, according to the DIVA model, speakers who have more acute perception of fine acoustic differences between vowels will learn auditory goal regions for vowels that are smaller and spaced further apart than speakers with less acute vowel perception. Such differences in goal regions among speakers would account for the current experimental results.

6. ACKNOWLEDGEMENT

Supported by Grant no. DC01925, National Institute on Deafness and Other Communication Disorders, N.I.H.

REFERENCES

[1] Rizzolatti, G. & Arbib, M.A. (1998). Language within our grasp, Trends Neurosci., 21, 188-194.

[2] Hickok, G. & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. Trends Cogn. Sci., 4, 131-138.

[3] Guenther, F.H. (1995). Speech sound acquisition, coarticulation and rate effects in a neural network model of speech production, Psych. Review 102, 594-621.

[4] Guenther, F.H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements, Psych. Review, 105, 611-633.

[5] Guenther, F.H. (in press). A model of cortical and cerebellar function in speech, Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Aug. 3-9.

[6] Lindblom, B. & Engstrand, O. (1989). In what sense is speech quantal?, J. Phonetics, 17, 107-121.

[7] Fox, R.A. (1982). Individual variation in the perception of vowels: Implications for a perception-production link, Phonetica 39, 1-22.

[8] Bradlow, A.R., Pisoni, D.B., Akahane-Yamada, R. & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production, J. Acoust. Soc. Am. 101, 2299-2310.

[9] Vick, J., Lane, H., Perkell, J., Matthies, M.L., Gould, J. & Zandipour, M. (2001). Covariation of cochlear implant users' perception and production of vowel contrasts and their identification by listeners with normal hearing. J. Speech, Language and Hearing Res. 44, 1257-67.

[10] Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., and Jackson, M. (1992). Electro-magnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements, J. Acoust. Soc. Am. 92, 3078-3096.

[11] Perry, N.C. & Michaedl, W. B. (1954). The reliability of a point-biserial coefficient of correlation. Psychometrika, 16, 313-325.

[12] Savariaux, C., Perrier, P., Orliaguet, J-P. & Schwartz, J-L. (1999). Compensation strategies for the perturbation of French [u] using a lip tube. II. Perceptual analysis. J. Acoust. Soc. Am., 106, 381-93.

[13] Jones, J.A. & Munhall, K.G. (2003). Learning to produce speech with an altered vocal tract: the role of auditory feedback. J. Acoust. Soc. Am., 113, 532-43.