IRis

Northeastern University

Psychology Faculty Publications

Department of Psychology

October 01, 2004

The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts

Joseph S. Perkell
Massachusetts Institute of Technology

Frank H. Guenther
Massachusetts Institute of Technology

Harlan Lane
Northeastern University; Massachusetts Institute of Technology

Melanie L. Matthies
Massachusetts Institute of Technology

Ellen Stockmann
Massachusetts Institute of Technology

See next page for additional authors

Recommended Citation

Perkell, Joseph S.; Guenther, Frank H.; Lane, Harlan; Matthies, Melanie L.; Stockmann, Ellen; Tiede, Mark; and Zandipour, Majid, "The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts" (2004). *Psychology Faculty Publications.* Paper 19. http://hdl.handle.net/2047/d20000866

This work is available open access, hosted by Northeastern University.

Author(s)
Joseph S. Perkell, Frank H. Guenther, Harlan Lane, Melanie L. Matthies, Ellen Stockmann, Mark Tiede, and Majid Zandipour

This article is available at IRis: http://iris.lib.neu.edu/psych_fac_pubs/19

The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts

Joseph S. Perkell,[a] Frank H. Guenther,[b] Harlan Lane,[c] Melanie L. Matthies,[d] Ellen Stockmann,[e] Mark Tiede,[e] and Majid Zandipour[f]
Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology, Room 36-511, 50 Vassar Street, Cambridge, Massachusetts 02139

(Received 14 March 2003; revised 30 April 2004; accepted 8 July 2004)

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, clear, 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 discrimination experiment, sets of seven natural-sounding stimuli ranging from *cod* to *cud* and *who'd* to *hood* were synthesized, based on productions by one male and one female speaker. The continua were then presented to each of the 19 speakers in labeling and discrimination tasks. Consistent with the hypothesis, speakers with discrimination scores above the median produced greater acoustic contrasts than speakers with discrimination scores at or below the median. Such a relation between speech production and perception is compatible with a model of speech production in which articulatory movements for vowels are planned primarily in auditory space. © *2004 Acoustical Society of America.* [DOI: 10.1121/1.1787524]

PACS numbers: 43.70.Aj, 43.70.Bk, 43.71.Es [AL] Pages: 2338–2344

I. 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. Mathiak *et al.*, 2002; Rizzolatti and Arbib, 1998) and auditory areas are active during speech production (cf. Hickok and Poeppel, 2000). A study using transcranial magnetic stimulation showed an increase of motor-evoked potentials in listeners' tongue muscles when they heard words whose production strongly involves tongue movements (Fadiga *et al.*, 2002).

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 relatively more clear-cut vowel contrasts 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 multidimensional auditory-temporal space (Guenther, 1995; Guenther *et al.*, 1998). In this model, DIVA,[1] 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 (Guenther; 1995; Lindblom and Engstrand 1989). 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 (see Guenther *et al.*, 1998 for details). According to the way the model functions, speakers who can perceive fine acoustic-phonetic details will learn goal regions spaced further apart. This may be because they are more likely than people with less acute auditory perception to reject poorly produced tokens of a phoneme when learning the goal regions.

II. BACKGROUND

Studies of the relation between production and perception have employed a variety of experimental paradigms and measures. In the most common approach, experimenters examine a group of speakers who vary in measures of production and perception and look for relationships between the two types of measures. For example, Fox (1982) had 11 subjects judge similarities of synthetic vowel stimuli in paired comparisons and identified three underlying dimensions, two of which were interpreted as representing the first two formants. The dimensional weightings that best predicted a given speaker's judgments of vowel similarities correlated with the formant values of that speaker's produced corner

[a]Corresponding author; Research Laboratory of Electronics, M.I.T. Dept. of Brain and Cognitive Sciences, M.I.T.; Department of Cognitive and Neural Systems, Boston University, Boston, MA; email address: perkell@speech.mit.edu
[b]Also of Department of Cognitive and Neural Systems, Boston University, Boston, MA.
[c]Also of Department of Psychology, Northeastern University, Boston, MA.
[d]Also of Department of Communication Disorders, Boston University, Boston, MA.
[e]Also of Haskins Laboratories, New Haven, CT.
[f]Also of Department of Cognitive and Neural Systems, Boston University, Boston, MA.

Downloaded 22 Feb 2011 to 129.10.104.144. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

vowels, providing evidence of a relation between production and perception. Recently, Newman (2003) found modest but significant cross-subject correlations in measures of 20 subjects' speech perception and properties of their speech productions. Speakers having perceptual prototypes with longer voicing onset times (VOTs) for the /p/ in /pɑ/ also tended to produce /pɑ/ with longer VOTs. Similarly, there was a relation between the frequency of the spectral peak of their prototypes and productions of the /ʃ/ in /ʃɑ/.

Other results have supported the idea that perception and production are linked, but somewhat less directly. Based on electromyographic (EMG) measurements, Bell-Berti *et al.* (1979) inferred that subjects formed two groups in their implementation of tense-lax differences among the vowels /i, ɪ, e, ɛ/: one group apparently used tongue height, the other, tongue tension. In a labeling test of an /i/ to /ɪ/ continuum with an anchoring condition, the magnitude of the induced boundary shift was greater for the tongue-height group. Frieda *et al.* (2000) had subjects produce the vowel /i/ in citation and hyperarticulated conditions and also had them select their ideal exemplar of /i/ from among 330 synthetic stimuli. A synthesized stimulus was classified as a prototype if chosen as an exemplar more than 13% of the time it was presented. According to this method, 24 of the 35 subjects had perceptual prototypes for /i/. The formant values of those perceptual prototypes were approximated more closely by subjects' hyperarticulated productions of /i/ than they were by their citation productions, but only for the subject group with relatively clear-cut prototypes.

There have also been studies of parallel changes in production and perception. For example, Bradlow *et al.* (1997) studied perceptual learning of English /r/ and /l/ by 11 Japanese speakers. The subjects recorded /r/ and /l/ productions before and after a number of sessions in which they were trained to identify /r/ and /l/. For measures of those productions, a group of English-speaking listeners identified and rated the phonemes produced before and after training. Improvements in these production measures covaried across speakers with measures of their perceptual learning. Rvachew (1994) obtained a similar finding when children were given perceptual training to correct phonological errors in their speech production.

In a study of eight postlingually deafened cochlear implant users, Vick *et al.* (2001) examined covariation of production and perception of vowel contrasts. Measures were made prior to the subjects' receiving their implants and 24–52 weeks postimplant. For the most part, subjects who produced vowel pairs with reduced contrast pre-implant and who showed improved perception of the contrast post-implant, also had enhanced production contrasts postimplant.

A link between perception and production has also been shown within individual subjects, using a ''sensorimotor adaptation'' paradigm. Houde and Jordan (1998, 2002) observed compensatory changes in the productions of vowels whispered by subjects whose auditory feedback had been altered. The feedback alteration shifted the first two formants along the /i, ɪ, ɛ, æ, ɑ/ axis, which had the effect of changing the identity of the vowel that was fed back to the subject. For example, when subjects initially pronounced *bed* and then were fed back an increasingly *bead*-like acoustic result, they compensated for the perceived vowel raising by vowel lowering and ended up pronouncing *bad*. The effect generalized to various consonant environments and vowels. For example, training with *get* yielded compensations in test words *geck*, *guess*, *debt*, and *pet* and also in test words *geet*, *git*, *gat*, and *got*. This demonstration of a tight link between production and perception within individual speakers is compatible with the functionality of mappings between articulatory and auditory frames of reference in DIVA (Guenther, 1995; Guenther and Ghosh, 2003).

Although there have been negative findings (cf. Paliwal *et al.*, 1983; Ainsworth and Paliwal, 1984—discussed further below), on balance, these rather diverse studies provide support for some kind of link between production and perception. Many of them 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 more distinct vowel contrasts than those who do not.

III. METHODS

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

A. Production experiment

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

1. Speech materials

The speech materials, which took an hour to read aloud, 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 and 9 repetitions of each *cud* and *hood* utterance (which were initially included as foils for a slightly different design). The two contrasts were chosen because the members of each contrasting pair are acoustically and articulatorily close to one another.[2] These utterances were part of a larger set that included other materials. In order to investigate the effects of speaking condition, the entire corpus was read in three different conditions, ''fast,'' ''normal,'' and ''clear.'' For the fast condition, the subject was asked to speak as rapidly as possible without eliminating any sounds. If necessary, the subject was reminded of this criterion as the experiment progressed. For the clear speech condition, the subjects were asked to pronounce the words carefully without increasing their loudness.

2. Recordings

An electromagnetic midsagittal articulometer system (EMMA—Perkell *et al.*, 1992) was used to record the posi-

Downloaded 22 Feb 2011 to 129.10.104.144. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

tion versus time of points on the subject's tongue, lips and jaws in the midsagittal plane. The subject was seated in an adjustable chair in a sound-attenuated room. A directional microphone was positioned about 14 in. from the subject's lips, and the subject read three repetitions of the utterance set to make an acoustic recording without the movement transducer system in place.[3] Then the transmitter assembly of the EMMA system was fit snugly to the subject's head. Small EMMA transducer coils (2 mm×5 mm) were attached at the midline 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). Transducer coils 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 a computer screen located about three feet in front of the subject. To allow monitoring of the progress of the experiment, the program also generated a real-time display of the acoustic signal and movement trajectories of the transducers.

The audio signal was low-pass filtered at 7.8 kHz and sampled at 16 kHz. Each signal channel from the EMMA system (three per transducer coil) was hardware low-pass filtered at 100 Hz and sampled at 500 Hz.

3. Data extraction

The data of interest were the *x* and *y* positions of a transducer coil on the tongue at the time the tongue reached its 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 target vowel in the sound pressure waveform for each token, using an interactive MATLAB procedure with displays of the sound pressure waveform and a spectrogram.

Approximately 4100 tokens were processed for this study (27 repetitions each of *cod* and *who'd*, 9 each of *cud* and *hood*, three conditions, 19 subjects). To make the analysis time tractable, algorithms were developed to automate the remaining data extraction steps.

The first step in the extraction of articulatory data for each token was the conversion of the raw EMMA voltage signals from each transducer to midsagittal-plane values of *x* (horizontal) and *y* (vertical) versus time. The voltage signals are first lowpass filtered using a Butterworth filter with a 100 Hz corner frequency and a roll-off rate of 18 dB/oct. Next, the transducer position versus time signals are translated and rotated into the maxillary frame of reference. This is followed by lowpass filtering using a ninth-order Butterworth with a 12.5 Hz corner frequency. For reasons explained below, the TB transducer was chosen as being most representative of the tongue body position for the vowel. Next, the time of the minimum in the velocity magnitude (speed) of the TB transducer during the vowel interval was located as the target point at which *x* and *y* data were extracted. For this purpose, TB velocity magnitude versus time was computed as the square root of the summed squares of central-differenced *x* and *y* values at each time step.

The first three formant frequencies were extracted with a method designed to minimize the occurrence of missing or spurious values. The audio signal was first pre-emphasized by first-order differencing (mu=0.98). Formants were obtained by peak-picking from a spectral envelope that was derived by averaging linear predictive coding (LPC) spectra taken over the interval from 30 to 40% of the delimited vowel interval using a sliding 30 ms window overlapped at 1 ms steps. In order to ameliorate problems with missing or incorrectly identified formants, each analysis was repeated using LPC orders of 16 through 22 inclusive, plus an "optimal" order chosen by a heuristic similar to that suggested by Vallabha and Tuller (2002). The intermediate result of this LPC analysis was eight largely overlapping sets of possible formant frequencies for each token (one for each LPC order). These frequency values from all tokens of a given vowel type were then binned within a histogram of 100 Hz resolution. Starting from expected values (mean results from Peterson and Barney, 1952) the closest peaks of the smoothed histogram were then used to determine subject-specific formant "targets" to use in choosing among the formants computed for each token, with the three closest values to the F1, F2, and F3 targets retained as the values for that token. Outliers exceeding two standard deviations were marked as missing data.

Two measures of vowel contrast were derived for each vowel pair, /ɑ-ʌ/ and /u-ʊ/, in each of the three speaking conditions, fast, normal, and clear. To derive *articulatory contrast distance* in each speaking condition, the mean values of TB*x* and TB*y* were calculated for each vowel and the Euclidean distance between the two sets of coordinates was found.

Correspondingly, to derive *acoustic contrast distance*, mean values of F1 and F2 were calculated for each vowel in each condition, and formant separation was calculated as the Euclidean distance between the means in the F1, F2 plane.

B. Perception experiments

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

1. Stimuli

For each of the two word pairs, *cod-cud* and *who'd-hood*, two sets of seven stimuli ranging from *cod* to *cud* and *who'd* to *hood* were synthesized, based on natural productions of the words in isolation by a male speaker and a female speaker (who were not subjects in the experiment). 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. The Klatt synthesizer (Klatt, 1980) as implemented in the SPEECHSTATION 2 SPEECH ANALYSIS WORKSTATION software (Sensimetrics, Inc., Somerville, MA) was used to generate the seven stimuli in each continuum. Within each continuum, the vowel portions of all seven

Downloaded 22 Feb 2011 to 129.10.104.144. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

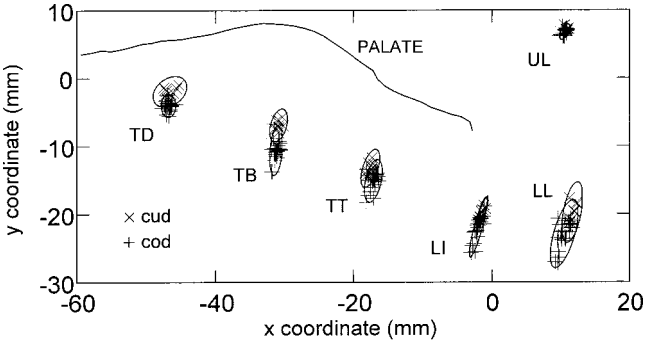

FIG. 1. EMMA transducer coil locations during all the productions of tokens containing the /ɑ/ in *cod* (+) and the /ʌ/ in *cud* (×) by one female subject in the normal condition. The transducer coils are located at points on the tongue dorsum (TD), tongue blade (TB), tongue tip (TT), lower incisor, (LI), lower lip (LL), and upper lip (UL). The midsagittal palatal contour (up to about the dento-alveolar junction on the right) is shown for reference.

stimuli were constructed to have the same naturally produced F0 contour and duration; the chosen duration was a compromise between the durations of the naturally produced endpoint utterances. For each stimulus, the synthesized vowel was inserted carefully between the same naturally produced initial-consonant release and final-consonant voicing signal so as to avoid discontinuities in the waveform. The resulting stimuli sounded quite natural in informal listening tests.

2. Tasks

Stimuli from all four of the continua (2 contrasts×2 genders) were presented to each of the subjects in labeling and discrimination tasks. Stimulus presentations for the tasks were created and administered using the Ecos/Win interface (AVAAZ Innovations, 1997 London, Ontario) and played through Bose noise-reducing headphones™. The labeling test was administered first, then the discrimination test. Stimuli were blocked by word pair, and all tasks were subject-paced.

The labeling task served as a control to verify the expected behavior of the listeners. In this 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 seven stimuli was presented 18 times.

The discrimination task was a classic ABX design (Liberman *et al.*, 1951). Stimuli were grouped into 60 sets of three stimuli where the first and second were one, two, or three 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 button 1 or 2 on the computer screen using the mouse. Each set was played only once in a block and the task consisted of two blocks, for a total of 120 trials.

IV. 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 one female subject in the normal 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. As exemplified in this figure, the TB category separation is the distance in millimeters between the centroids of the *cod* and *cud* TB distributions.

To clarify how the perceptual results were parametrized, Fig. 2 displays results for six of the subjects for the cod-cud continuum. These six illustrate the ranges of variation in labeling slopes and discrimination accuracy that were observed. Each panel contains a labeling function (filled circles connected by solid lines) whose values are the percentage of instances in which the presented stimulus (numbered 1–7 on the horizontal axis) was labeled *cod*. The remaining three functions show the percent correct for the one-step (number 1 connected by dotted lines), the two-step, and the three-step ABX stimuli. Subjects 3, 9, 14, and 7 are males and subjects 19 and 6 are females. The top three plots show responses to the male synthesized stimuli and the bottom three plots, responses to the female synthesized stimuli. There is 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. Such cross-speaker variation is consistent with previous findings (cf. Pisoni, 1971).

The one-, two-, and three-step stimulus comparisons were not equally successful in differentiating among the subjects' perceptual performances. The peak three-step discriminations produced a ceiling effect for most subjects, and the peak one-step stimuli led to highly variable scores, many at chance level. Moreover, many one-step intervals did not cross the labeling category boundary and the peak one-step scores did not correlate with any production measures. We retained, therefore, the peak discrimination scores obtained with stimuli separated by 2 steps on the synthetic continua.[4]

Figure 3 presents frequency distributions of the 19 subjects with respect to those two-step peak discrimination scores for each of the two vowel contrasts. For the following reasons, we classified subjects into two groups based on those scores. Even the two-step discrimination triads yielded substantial ceiling effects, with 37% of the subjects (*who'd-hood*) or 26% of the subjects (*cod-cud*) scoring 100 percent. Those subjects presumably had higher discriminative acuity than this discrimination task was able to measure. All subjects' scores fell at one of five values (*who'd-hood*) or one of four 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 "high"

Downloaded 22 Feb 2011 to 129.10.104.144. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

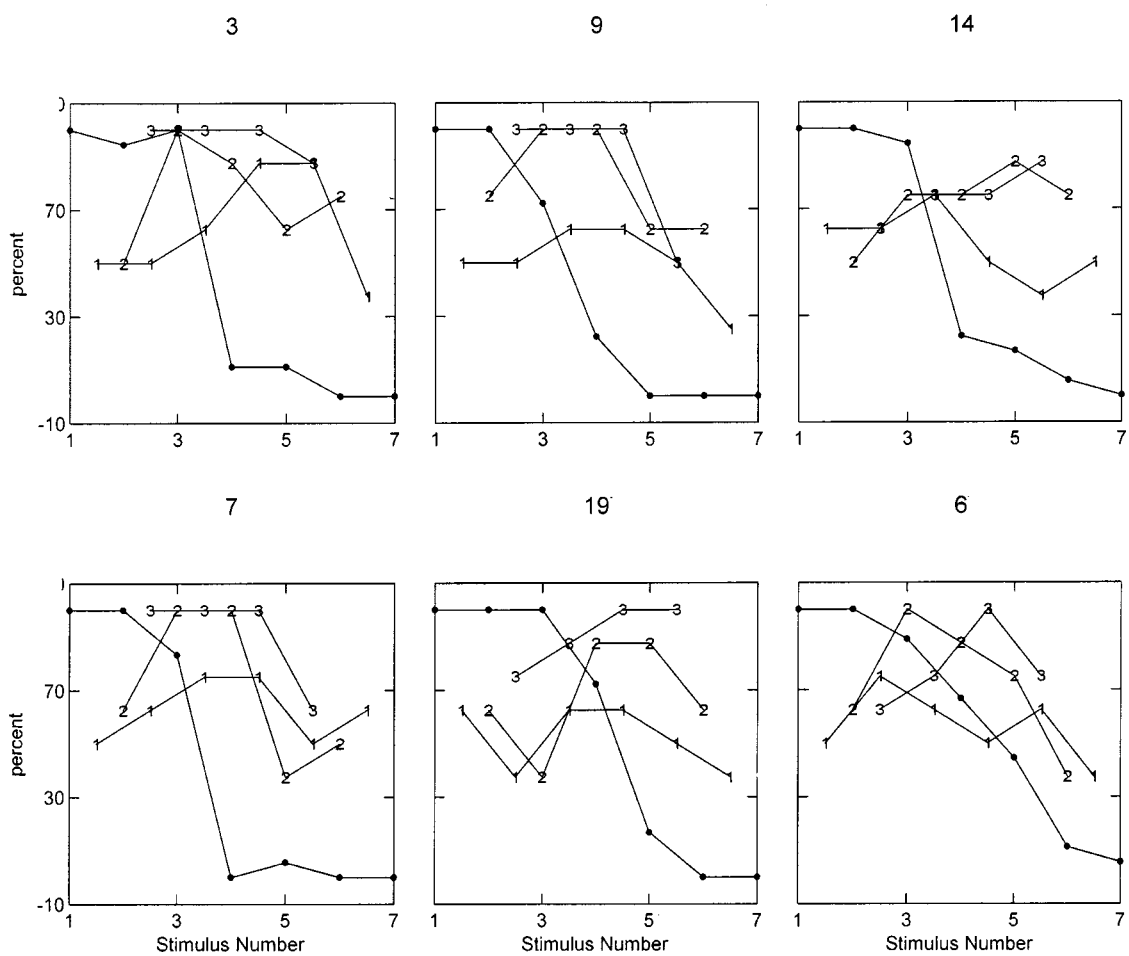

FIG. 2. Examples of perception test results for the cod-cud continua from six of the subjects. Each panel contains a labeling function (filled circles connected by solid lines) whose values are the percentage of instances in which the presented stimulus (numbered 1–7 on the horizontal axis) was labeled cod. The remaining three functions in each panel show the percent correct for the one-step (number 1 connected by dotted lines), the two-step, and the three-step ABX discrimination stimuli.

discriminators. All others subjects were classified as "low" discriminators. (As it turned out, high discriminators scored above the median peak two-step score and low discriminators at or below the median.)

Figure 4 plots articulatory contrast distance (for tongue body position) and acoustic contrast distance (for separation in the formant plane) in the upper and lower panels, respectively, as a function of the three speaking conditions. The left-hand panel gives results for *who'd-hood*, the right for *cod-cud*. For both articulatory and acoustic parameters and for both vowel contrasts, high discriminators (values labeled "*H*") produced greater contrast distances than low discriminators ("*L*") on the average. This effect was reliable for the articulatory parameter for *who'd-hood* ($F=28.6$, df=1,131, $p<.01$) and *cod-cud* ($F=17.9$, df=1,116, $p<0.01$). The effect was also reliable for the acoustic parameter for *cod-cud* ($F=26.4$, df=1,70, $p<0.01$) but not for *who'd-hood* where unexplained between-speaker variability was great ($F=1.6$, df=1,73, $p>0.05$; error bars are one standard error of the mean). As inspection of Fig. 4 reveals, speaking condition had an effect on both articulatory and acoustic contrast distance for *who'd-hood* (respectively, $F=49.1$, df=2,262, $p<.01$; $F=5.5$, df=2,146, $p<0.01$) but not for *cod-cud* ($F=0.5$, df=2,232, $p>.05$; $F=1.8$, df=2,140, $p>0.05$).

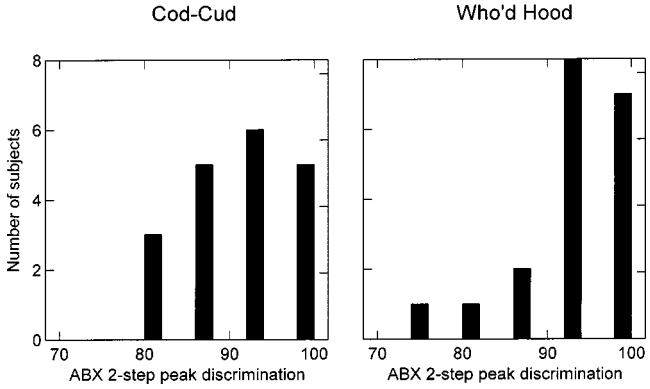

FIG. 3. Frequency distributions of the 19 subjects with respect to two-step peak discrimination scores on an ABX task for each of the two vowel contrasts.

V. 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. This finding is consistent with and extends the results of other investigations

Downloaded 22 Feb 2011 to 129.10.104.144. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

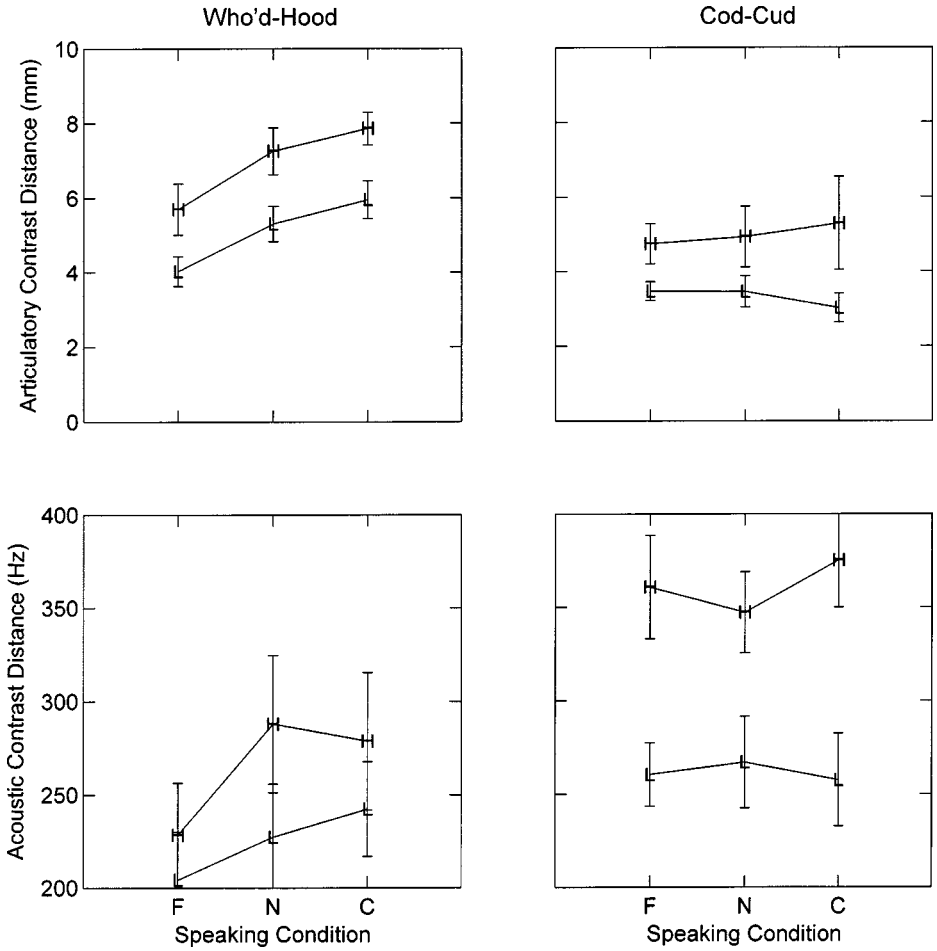

FIG. 4. Articulatory contrast distance (for tongue body position—upper panel) and acoustic contrast distance (for separation in the formant plane—lower panel) as a function of the three speaking conditions. The left-hand panel gives results for *who'd-hood*, the right for *cod-cud*. Findings for high discriminators (labeled "*H*") and low discriminators ("*L*") are plotted separately. Error bars are one standard error about the mean.

(see Background; also Ladefoged *et al.*, 1972). It is also compatible with the results of Perkell *et al.* (in press), who studied individual differences in producing the sibilant contrast in American English and their relation to two speaker characteristics: speakers' use of a quantal biomechanical effect in producing the sibilants and their performance on a test of sibilant discrimination. Parallel to the current results, Perkell *et al.* (in press) found that speakers with more acute sibilant discrimination produced greater sibilant acoustic contrast distances.

In Sec. II, we cited two studies with objectives and methods similar to the present experiment, which failed to find support for the hypothesis that speakers who discriminate a contrast more acutely produce that contrast more distinctly. Paliwal *et al.* (1983) performed a vowel production and perception experiment on ten subjects to evaluate the hypothesis that a listener refers to his own articulation for perceiving speech. Although this hypothesis, that production regulates perception, is the converse of ours, their general objective was the same: to find relations between production and perception. The subjects produced 11 vowels of English in /hVd/ context and they identified synthesized vowels (in the same context) in which values of F1 and F2 covered the entire F1, F2 plane. The results of correlations of the formant frequencies of produced and perceived vowels, within and between subjects, led to rejection of the hypothesis. Ainsworth and Paliwal (1984) used a similar approach in studying English glides /w, r, l, j/, with the same results and conclusion. There are several differences between these two studies and the current one that could account for the different outcomes, most notably the fact that we used discrimination scores not identification results as the perceptual measure.

As mentioned in the Introduction, our general hypothesis, that perception influences production, is compatible with the DIVA model (Guenther, 1995; Guenther *et al.*, 1998) 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.

We hypothesize that language learners find it advantageous in their communicative interactions to be as intelligible as possible across a range of acoustic transmission conditions; therefore, according to the functionality of DIVA, 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.

ACKNOWLEDGMENTS

This research was supported by Grant No. DC01925 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

Downloaded 22 Feb 2011 to 129.10.104.144. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

[1]"DIVA" stands for "directions into velocities of articulators."

[2]Other contrasts could have been considered; however, a more systematic exploration of different kinds of distinctions (i.e., tense/lax, low/high, back/front) is beyond the scope of the current study.

[3]These utterances were used to verify informally that vowel productions with the EMMA system in place were not distorted by the presence of the EMMA transducer coils after a short period of accommodation.

[4]For both sets of continua (*who'd-hood* and *cod-cud*), correlations across subjects between peak one-step percent correct and peak two-step percent correct were not significant. A cross-subject correlation between values of peak two-step discrimination percent correct for *who'd-hood* and those for *cod-cud* was close to but missed significance ($r = 0.43$, $p = 0.068$).

Ainsworth, W. A., and Paliwal, K. K. (**1984**). "Correlation between the production and perception of the English glides /w, r, l, j/," J. Phonetics **12**, 237–243.

Bell-Berti, F., Raphael, L. J., Pisoni, D. B., and Sawusch, J. R. (**1979**). "Some relationships between speech production and perception," Phonetica **36**, 373–383.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., and 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.

Fadiga, L., Craighero, L., Buccino, G., and Rizzolati, G. (**2002**). "Speech listening specifically modulates the excitability of tongue muscles: a TMS study," Eur. J. Neurosci. **15**, 399–402.

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

Frieda, E. M., Walley, A. C., Flege, J. E., and Sloane, M. E. (**2000**). "Adults' perception and production of the English vowel /i/," J. Speech Lang. Hear. Res. **43**, 129–143.

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

Guenther, F. H., and Ghosh, S. S. (**2003**). "A model of cortical and cerebellar function in speech," Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Aug. 3–9 (Universitat Autònoma de Barcelona, Barcelona, Spain), 169–174.

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

Hickok, G., and Poeppel, D. (**2000**). "Towards a functional neuroanatomy of speech perception," Trends Cogn. Sci. **4**, 131–138.

Houde, J. F., and Jordan, M. I. (**1998**). "Sensorimotor adaptation in speech production," Science **279**, 1213–1216.

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

Klatt, D. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Ladefoged, P., DeClerk, J., Lindau, M., and Papcun, G. (**1972**). "An auditory-motor theory of speech production," UCLA Working Papers in Phonetics **22**, 48–94.

Liberman, A. M., Harris, K. S., Kinney, J. A., and Lane, H. (**1951**). "The discrimination of relative onset time of the components of certain speech and nonspeech patterns," J. Exp. Psychol. **61**, 379–388.

Lindblom, B., and Engstrand, O. (**1989**). "In what sense is speech quantal?" J. Phonetics **17**, 107–121.

Mathiak, K., Hertrich, I., Grodd, W., and Ackermann, H. (**2002**). "Cerebellum and speech perception: a functional magnetic resonance imaging study," J. Cogn Neurosci. **14**, 902–912.

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.

Paliwal, K. K., Lindsay, D., and Ainsworth, W. A. (**1983**). "Correlation between production and perception of English vowels," J. Phonetics **11**, 77–83.

Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., and Jackson, M., (**1992**). "Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements," J. Acoust. Soc. Am. **92**, 3078–3096.

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., and Stockmann, E. (in press). "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.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Pisoni, D. B. (**1971**). "On the Nature of Categorical Perception of Speech Sounds," Supplement to Status Report on Speech Research, Haskins Laboratories, New Haven, CT.

Rizzolatti, G., and Arbib, M. A. (**1998**). "Language within our grasp," Trends Neurosci. **21**, 188–194.

Rvachew, S. (**1994**). "Speech perception training can facilitate sound production learning," J. Speech Lang. Hear. Res. **37**, 347–357.

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

Vick, J., Lane, H., Perkell, J. S., Matthies, M. L., Gould, J., and Zandipour, M. (**2001**). "Covariation of cochlear implant users' perception and production of vowel contrasts and their identification by listeners with normal hearing," J. Speech Lang. Hear. Res. **44**, 1257–67.

Downloaded 22 Feb 2011 to 129.10.104.144. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp