Mechanisms of Vowel Production: Auditory Goals and Speaker Acuity

Joseph S. Perkell[a,c,]*, Harlan Lane[b,a], Satrajit Ghosh[a], Melanie L. Matthies[c,a], Mark Tiede[d,a], Frank Guenther[c,a] and Lucie Ménard[e,a]

a: Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
b: Northeastern University, Boston, MA; *c:* Boston University, Boston, MA; *d:* Haskins Laboratories, New Haven, CT; *e:* Université du Québec à Montréal, Montréal, Québec, Canada
*Email: perkell@mit.edu

Abstract

Acoustic recordings were made of vowel productions by 20 young adult speakers of American English to investigate the effects of speaker and speaking condition on produced phoneme contrast. The vowels /i/, /ɪ/, /ɛ/, /æ/, /ɑ/ *and* /u/ *were embedded in "compound words" consisting of two CVC words. The compound words, embedded in a carrier phrase, were spoken in normal, clear, and fast conditions. To investigate relations between contrast and acuity, each subject's vowel discrimination was measured using stimuli from synthetic continua.*

Produced vowel contrast decreased with speaking condition, from clear *to* normal *to* fast, *as did vowel duration, but the two were not correlated. As found previously, speakers with greater spectral acuity produced vowels with more spectral contrast than speakers with lesser spectral acuity. In a new finding, measures of speaker acuity and vowel dispersion were negatively correlated: speakers with higher acuity had smaller vowel target regions. These results are interpreted with respect to the functionality of the DIVA model of speech motor planning.*

1 Introduction

This study concerns one of the central questions in speech motor control and motor control in general: what is the *task space* or highest-level control domain of a particular kind of movement? Since a speaker's main objective is to produce an intelligible percept in the listener, we hypothesize that at least some of the goals of speech movements are *regions in auditory space*. Several lines of evidence support this view, including the results of motor equivalence experiments [cf. 1] , sensorimotor adaptation experiments [cf. 2, 3] and studies of the effects of changes in hearing status in cochlear implant users and normal-hearing speakers [cf. 4]. Auditory goal regions are also a basic component of the DIVA neurocomputational model of speech motor control [cf. 5]; the experiments reported here employ acoustic and perceptual measures to test hypotheses based on that model. Using speaking condition, vowel stress and phonetic environment as probes, they investigate the nature of phonemic goals for vowels and how they are influenced by speakers' auditory capabilities.

2 Methods

The study is comprised of two experiments: 1) a production experiment on the effects of speaking conditions and speaker differences on goals for vowels, and 2) an experiment relating speakers' auditory acuity for vowel contrasts to their production data from Experiment 1.

2.1 Experiment 1: Production measures

Acoustic recordings were made of vowel productions by young adult speakers of American English – 10 females and 10 males. The corpus was designed to investigate the effects of speaker and speaking condition on measures of contrast and dispersion. The vowels /i/, /ɪ/, /ɛ/, /æ/, /ɑ/ and /u/ were embedded in "compound words" consisting of two CVC words, each of which comprised a real word. Variations of phonetic context and stress were used to induce as much dispersion as possible around each vowel centroid (e.g., *peepPICK, tapPET, tickCOCK, keepPOT,* where the capitalized word was stressed and the two central consonants were the same). The compound words were embedded in a carrier phrase and were

spoken in *normal, clear*, and *fast* conditions. Formants and durations were extracted and formants were converted to mels. Overall vowel contrast was estimated for each speaker and across speakers using Average Vowel Spacing (AVS) – the average of all inter-vowel distances, in mel space. Dispersion for each vowel was measured as the average distance in the F1 x F2 plane (in mel coordinates) of vowel tokens from their corresponding vowel-type means.

These data were used to test two hypotheses:

H1: Effects of speaking condition. According to the DIVA model [5], changes from *clear* to *normal* to *fast* speech for vowels are implemented in part by reducing the spectral ranges of their goal regions. As schematized in Fig. 1, the model predicts that, because of economy of effort [cf. 6], auditory trajectories will pass through the parts of auditory goal regions that are closest to those of temporally adjacent sounds. The figure schematizes the hypothesis that relatively peripheral vowels will move toward the center of the vowel space and contrast (AVS) will decrease as the condition changes from *clear* (to *normal* – not shown) to *fast*.

H2: Trading relation. Because vowel intelligibility is a function of both contrast and duration, and the speaker's primary objective is to maintain intelligibility, there will be a trading relation across speakers such that, comparing *clear* to *fast* speech, some speakers change spectral contrast more and duration less while others do the converse.

2.2 Experiment 2: Perception measures

Eighteen of the original 20 subjects could return for perceptual testing and served as subjects. The Klatt synthesizer was used to generate gender-neutral vowel continua in 1001 steps in Bark space for the continua /pɪp/-/pɛp/ and /pɛp/-/pæp/. In labeling tasks, the subjects heard tokens that were selected from 11 equally spaced intervals between the endpoints. Logistic functions were fit to the labeling data and the intersections were used to establish each subject's phoneme boundary for each continuum. For a measure of auditory spectral acuity, a discrimination test was used to determine each subject's JND for each continuum at the subject's category boundary. The discrimination test used a 4-interval, 2-alternative forced choice task, in which the subjects heard the sequence A-B-A-A or A-A-B-A and had to select

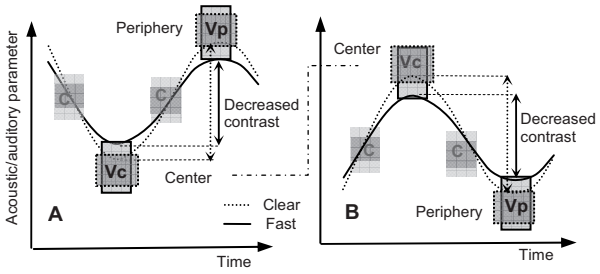

Figure 1: *Schematic diagrams of planned trajectories for an auditory parameter vs. time, passing through CVCV goal regions. Vc: Central Vowel; Vp: Peripheral Vowel. Solid trajectory:* fast*; Dotted trajectory:* clear. *Part A: Vp is "higher" than Vc; Part B: Vp, lower than Vc. (Dotted-dashed line: plots are offset vertically.) Expanding the spectral dimensions of goal regions from* clear *to* fast *speech leads to decreased contrast.*

whether the 2nd or 3rd item was different from the rest. Subjects were provided feedback about the accuracy of their responses. An adaptive staircase procedure was used in which the initial interval between A and B was large. After each trial, the interval decreased following a correct response and increased following an incorrect response. The staircase was terminated after 14 reversals. The JND for that run was estimated as the mean of the separation between A and B of the final four reversals. JNDs were estimated from the last of four such staircase runs for each subject and each synthetic-speech continuum, on the assumption that the last run represented the best estimate of the JND for that subject and continuum. Across subjects, the JND frequency distributions were right-skewed (since there was a lower limit on discriminable difference but no upper limit). Thus, each subject's acuity was defined as the reciprocal of the JND averaged over the two continua.

These data were used in combination with results from Experiment 1 to test two additional hypotheses:

H3: Acuity and contrast are related. Higher-acuity speakers, i.e., those who are better able to discriminate exemplars of vowel sounds with subtle spectral differences, will produce those sounds with more contrast than speakers with lower acuity.

H4: Acuity and dispersion are inversely related. Higher-acuity speakers will have smaller vowel goal regions and hence produce those sounds with less token dispersion than lower-acuity speakers.

3 Results

Average values of AVS (mels) as a function speaking condition and word stress are plotted in Fig. 2.

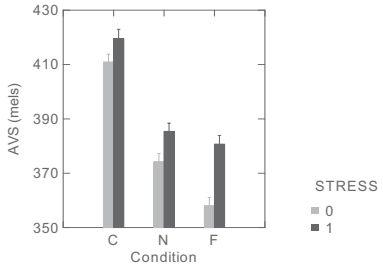

Figure 2. *Vowel contrast (AVS - mels) in clear (C), normal (N) and fast (F) speaking conditions. Data were pooled over all repetitions of each vowel in stressed and unstressed positions from 20 speakers.*

A repeated-measures ANOVA showed statistically significant effects of speaking condition, phonetic context and stress and their interactions on AVS (except for the interactions involving speaking condition x context). A post-hoc test comparing *fast* and *clear* speech showed that AVS was significantly reduced from *clear* to *fast* speech (F=70, df=20,120, p<.01).

Referring back to Fig. 1, the schematic diagrams in parts 1A and 1B show that peripheral vowels are expected to "move" toward the center of the vowel space in changes from *clear* to *fast* speech. In addition, central vowels should move upward toward higher vowels (1A) and downward toward lower vowels (1B) in *fast* vs. *clear* speech. In a variety of contexts that include high, low and central vowels, more central vowels are expected to show relatively little net movement and greater overlap of *clear* and *fast* distributions than peripheral vowels. Thus the AVS changes hypothesized in H1 should be accompanied by differences in the amount of condition-related overlap of central and peripheral vowel distributions.

Figure 3 shows 95% confidence ellipses aligned with the axes of the first two principal components for *clear* and *fast* instances of each vowel type. A measure of overlap (in mels) was computed as the ratio of the area in common between the *clear* and *fast* ellipses (the "INTERSECT") to the full area of each *clear* ellipse. Higher values thus indicate a greater degree of overlap. That ratio was averaged across vowels and speakers separately for peripheral and central vowels. The overlap was 77.3% for the two

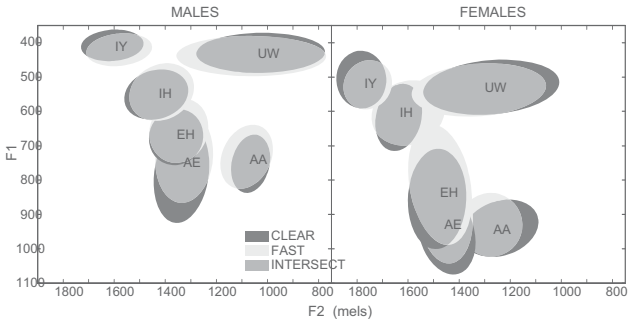

Figure 3. *95% confidence ellipses aligned with the axes of the first two principal components for* clear *and* fast *instances of each vowel type.*

central vowels (/ɛ/=EH and /ɪ/=IH) and 69.8% for the four peripheral vowels – a reliable difference (t for matched pairs, 2.2, df=19, p<.05 two-tailed).

These results support the elaborated H1: when vowels were produced in different contexts, they moved toward the center of the vowel space as speaking condition changed from *clear* to *normal* to *fast*, with greater movement for point vowels, as evidenced by decreases in AVS and differences in overlap between central and peripheral vowels in *clear* vs. *fast* conditions.

H2 predicted a trading relation across speakers such that, comparing clear to fast speech, some speakers change contrast distance more and duration less while others do the converse. AVS decreased from clear to fast speech (a 10% drop), as did syllable duration (a 13% drop). When all utterances were pooled across speakers, AVS (mels) and duration were not correlated (r=0.09). Individual speakers ranged widely in that correlation; the average value was (r=0.15). Thus, H2 is not supported; no trading relation was found between AVS and duration changes under changes in speaking condition from *clear* to *fast*.

Differences among speakers in acuity were correlated with AVS in each of the three speaking conditions, as hypothesized (H3). The product-moment correlations in clear, fast and normal speaking conditions, respectively, had values of .44, .51 and .44 (all p <.05, one tailed). As predicted by H4, differences among speakers in acuity were correlated inversely with the size of their average vowel dispersion, in each of the three speaking conditions. The product-moment correlations in *clear*, *fast* and *normal* speak-

ing conditions, respectively, had values of -.51, -.57, and -.51 (all $p<.05$, one-tailed).

4 Summary and Discussion

The results of Experiment 1 show that vowel contrasts decreased and vowel productions moved toward the center of the acoustic vowel space as speaking condition changed from *clear* to *normal* to *fast*. These findings confirm H1 and are consistent with a trade-off between clarity and economy of effort [6]. H2 was disconfirmed: rather than the predicted trading relation between condition-related changes in vowel contrast and durations, measures of vowel contrast and durations were not correlated.

In support of H3 and consistent with previous findings [cf. 7], results from Experiment 2 showed positive correlations between produced contrast and speaker acuity for vowels: speakers with better acuity produced greater vowel contrasts. H4 was confirmed in the current study by a new finding. There were negative correlations between speaker acuity and measures of vowel dispersion: speakers with higher acuity had smaller vowel target regions. An interpretation of these findings is shown in Fig. 4.

The figure shows schematic goal regions of a high-acuity speaker and a low-acuity speaker for the vowels /ɪ/ and /ɛ/ in F1 x F2 space. The solid black circles represent the goal regions for the high-acuity speaker; the dashed gray circles, for the low-acuity speaker. The smaller size and greater distance between the high-acuity target regions is consistent with the current findings of smaller dispersions and greater AVS for high- than for low-acuity speakers. We assume that when acquiring speech, children learn that it is advantageous to speak clearly; therefore, they will acquire goal regions for sounds that are relatively small and spaced far apart. According to the DIVA model, high-acuity speakers would acquire goal regions that are smaller and spaced further apart compared to low-acuity speakers. That is because when learning the goal regions, the high-acuity speaker would reject outlying exemplars (such as the sound indicated by an X in the figure) as being produced badly, whereas the low-acuity speaker would consider such sounds to be acceptable.

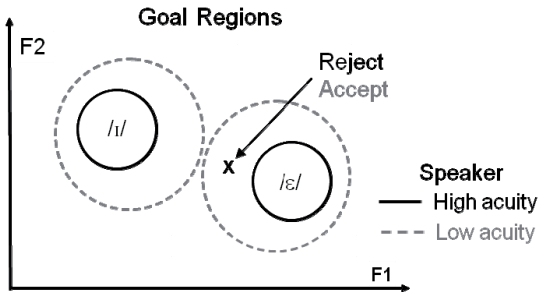

Figure 4. *Schematic illustration of goal regions of a high-acuity speaker and a low-acuity speaker for the vowels /ɪ/ and /ɛ/ in F1 x F2 space.*

5 Acknowledgement

This research was supported by Grant no. R01-DC01925 from the National Institute on Deafness and Other Communication Disorders, N.I.H.

6 References

[1] J.S. Perkell, M.L. Matthies, M.A. Svirsky, & M.I. Jordan. 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, 1993.

[2] J.F. Houde & M.I. Jordan. Sensorimotor adaptation of speech I: Compensation and adaptation. *J. Speech, Language, Hearing Res.* 45:295-310.

[3] V. Villacorta, J.S. Perkell & F.H. Guenther. Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception, *J. Acoust. Soc. Am.* 122:2306-2319, 2007.

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

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

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

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