OSU Working Papers in Linguistics 50 (115-125)

Production and perception of individual speaking styles *

Keith Johnson & Mary E. Beckman
kjohnson@ling.ohio-state.edu
mbeckman@ling.ohio-state.edu

Abstract: As explanation of between-speaker differences in speech production moves beyond sex- and age-related differences in physiology, discussion has focused on individual vocal tract morphology. While it is interesting to relate, say, variable recruitment of the jaw to extent of palate doming, there is a substantial residue of arbitrary differences that constitute the speaker's "style". Style differences observed across a well-defined social group indicate group membership. Other style differences are idiosyncratic "habits" of articulation, individual solutions to the many-to-many mapping between motoric and acoustic representations and to the many different attentional trading relationships that can exploit the typical patterns of redundant variation in independent acoustic correlates of any minimal contrast. Perceptual studies of social style differences suggest that perceptibility depends upon the task and upon the hearer's own group membership. The few studies of idiosyncratic differences suggest that speakers perceive each others' productions in terms of their own habits. Thus, perceptual compensation for speaker differences must go beyond mere vocal tract normalization. A promising route for describing how listeners compensate for the arbitrary variation of style is an instance-based (or exemplar) model of speech perception in which the distribution of exemplars is heavily weighted by instances of the speaker's own productions.

Introduction

Until very recently, most discussion of between-speaker differences has been couched in the framework of "speaker normalization". In this framework, gestures are equated with the dimensions of invariant linguistic contrast between phonemes ("distinctive features"), and between-speaker variability is treated as an artifact of the transmission line — a kind of noise which needs to be filtered out of the signal in order to get at the meaningful category variation.

This paper illustrates several ways in which gestures for the same phoneme category can differ meaningfully across speakers, and then discusses the implications for our models of the listener. If listeners can categorize speakers, then the problem is not merely one of normalizing over speakers to perceive phonemes, but a more general problem of how to extract categories in one dimension of classification in the face of meaningful variation in another dimension of classification. We propose a model of how listeners might process utterances for all of the linguistically relevant categories that the signal encodes.

* Talk presented at the 3rd joint meeting of the Acoustical Society of America and the Acoustical Society of Japan, Honolulu, 2-6 December, 1996. Thanks to Mariapaola D'Imperio for helpful comments on an earlier version.

Meaningful Variation

The earliest work on between-speaker differences, of course, categorized speakers entirely in terms of age- and sex-related changes in vocal tract morphology (see Peterson & Barney, 1952 and use of these data in testing nearly all subsequent proposals of vowel normalization algorithms). In particular, we know that adult male talkers tend to have lower formants and lower fundamental frequencies than adult females do, because of hormonal changes at puberty that lead both to a descent of the larynx that elongates the vocal tract and a simultaneous change in the morphology of the thyroid cartilage that elongates the vocal folds. Such observations prompted algorithms for normalizing formant values by fundamental frequencies, and the like, to find the invariant underlying gesture (Nearey, 1978; Miller, 1989).

Articulatory studies, however, suggest that there are real between-speaker differences in gesture. For example, figure 1 shows x-ray traces of jaw and tongue surface at vowel mid-point in front vowels produced by two adult male speakers of American English. The speaker on the left shows a large variation in jaw height that is systematically related to the contrasts between the two high and two mid vowels and between the mid and low vowels. The speaker on the right shows hardly any variation in jaw height across the five vowels. We speculate that these different gestural strategies may be related to between-speaker differences in palate shape. That is, a more steeply domed palate might be associated with an individual articulatory style that does not recruit the jaw much in tongue raising and lowering gestures for vowels.

In figure 2 we see similar between-speaker variation in the coordinated movement of jaw and tongue in a set of magnetometer studies reported by Harrington and Fletcher (1996). They compared high and low vowels of Australian English produced in accented versus unaccented positions in the intonation contour, and showed that some speakers (such as JMF) lower the jaw more in accented syllables, whereas other speakers (such as LML) have very little variation in jaw position across accented versus unaccented position. Here individual speaker style seems to be associated not with different morphologies,

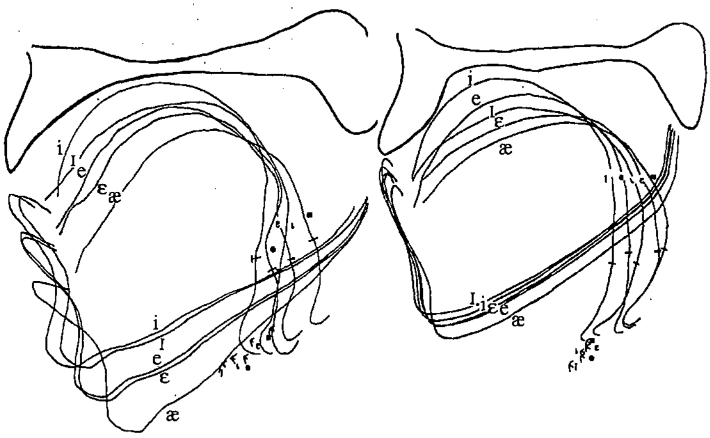

Figure 1. Tongue shapes during vowels for two speakers in Ladefoged, DeClerk, Lindau & Papcun (1972).

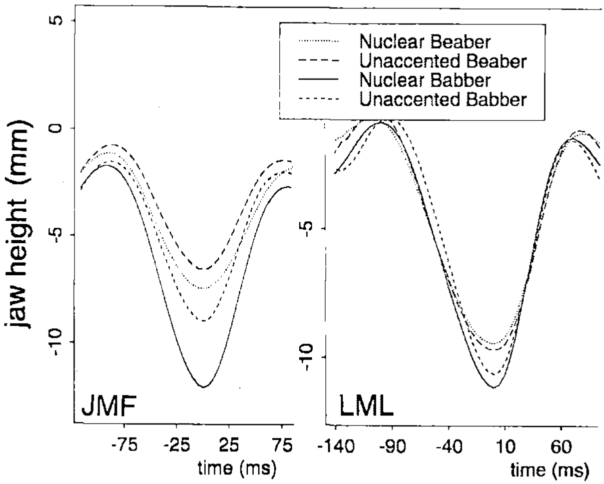

Figure 2. Average jaw trajectories for speakers JMF and LML from Harrington & Fletcher's (1996) study.

but with subtle differences in the prosodic system. JMF consistently makes vowels in accented syllables longer and more peripheral, whereas LML does not use these redundant non-tonal cues to pitch accent placement. See also Edwards, Beckman, and Fletcher (1991) and deJong (1995) for comparable inter-speaker differences in prosodic strategies for American English. Harrington & Fletcher's study also suggested another kind of difference between the two talkers in figure 2. Figure 3 shows traces for the tongue body vertical position and for the first formant in representative tokens of utterances with the high tense vowel [i:] produced by these two talkers. JMF's production has a high tongue body throughout the vowel, and a relatively flat and very low F1, whereas LML's production shows a pronounced diphthongal movement, with a distinct peak in tongue body position late in the vowel and a much higher F1 at the beginning of the vowel. We suspect that this difference is part of a larger pattern of variation in style defined by the continuum of Australian English features. That is, JMF's higher vowel here is typical of so-called "Cultivated Australian", which is closer to British English, whereas LML's lower onset and decidedly diphthongal pattern is closer to the "Broad Australian" end of the continuum.

Figure 4, from Harrington and Cassidy (1994), shows average formant values in a database of Australian English and a comparison plot of typical British English formant values. The diphthongal lowering at the beginning of the tense front vowel in *heed* in Australian English doesn't show up very well, because these averages are from the vowels' midpoint values. But the figure does show another salient feature of Australian English — namely, the raising of the lax front vowels in *hid*, *head*, and *had*. To us, this pattern is very reminiscent of some differences in regional dialects that we've found in our ongoing studies of American English vowel systems.

Figure 5 shows vowel spaces for 13 female talkers from Birmingham, Alabama, and 7 from Los Angeles, California. The Alabama data are from Johnson's unpublished work, and the California data are from Johnson, Flemming, & Wright (1993). As in figure

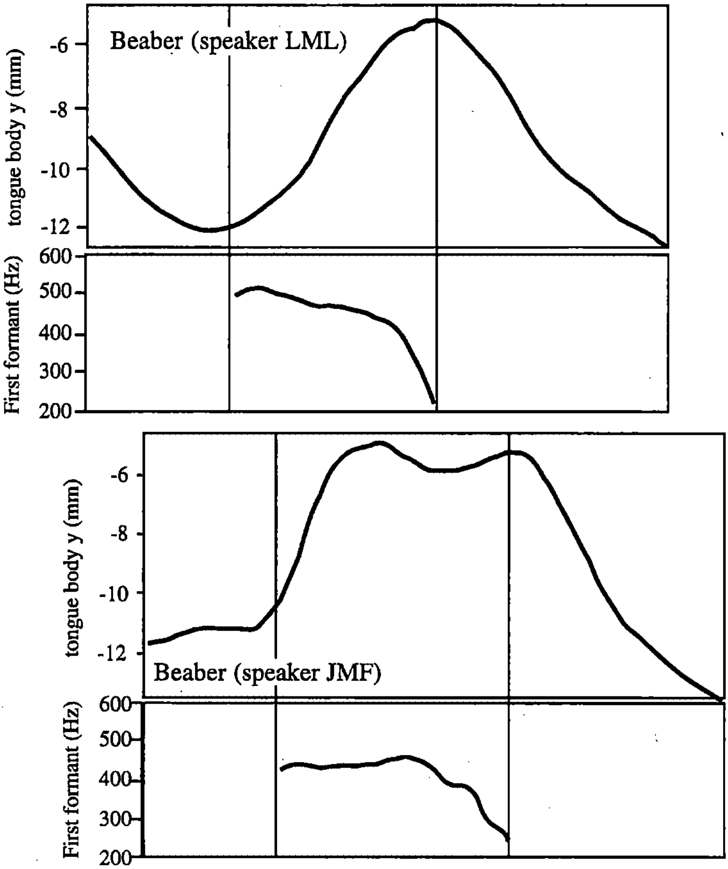

Figure 3. Tongue body height and first formant trajectories of sample tokens of "Beaber" from speakers JMF and LML. Vertical lines mark [i] onset and offset.

4, the formant values that are observed here are taken from vowel midpoints, so the plot does not show that many of the Alabama speakers had a lower onset value for the tense front vowel. However, the figure does show that for the Alabama speakers the lax front vowels in *head* and *had* are raised relative to productions by the Los Angeles speakers. These kinds of speaker-style differences observed across a well-defined social or regional group indicate group membership, and sociolinguistic studies have shown that listeners can be very acutely aware of them, particularly when the differences are associated with differences in social prestige or stigma (see, e.g., Labov, 1966; Trudgill, 1974). A normalization algorithm which treats this kind of between-speaker variability as noise to be factored out of the signal could not be an accurate model of how real listeners extract relevant categories from the signal.

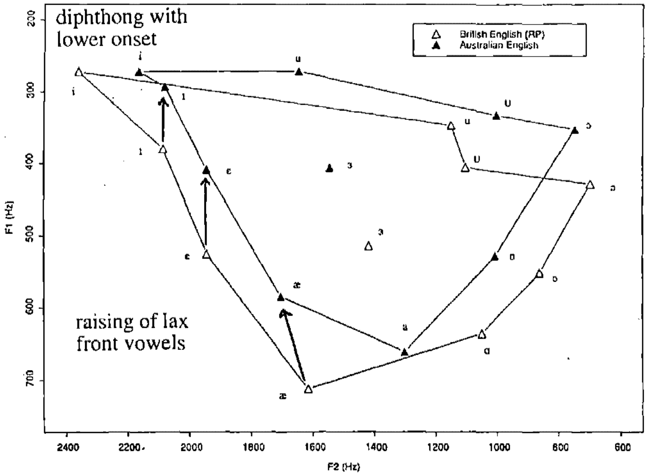

Figure 4. Harrington & Cassidy (1994) acoustic vowel formant measurements comparing Australian versus RP vowel spaces.

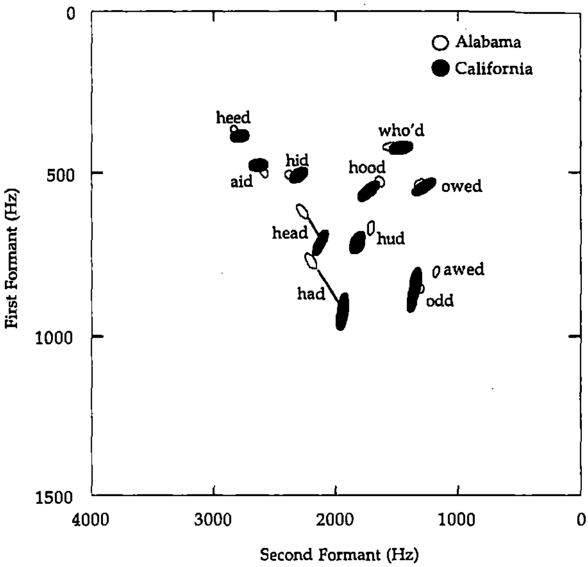

Figure 5. Spectral differences across dialects: Alabama versus Los Angeles vowel spaces. Ellipses show 95% bivariate confidence intervals: filled - Los Angelenos, open - Alabamians.

Using Variability in Speech Perception

We have presented evidence of at least three types of meaningful between-speaker variability in articulation in addition to the average age- and sex-related differences in vocal apparatus size. Clearly, then, perceptual compensation for speaker differences has to go well beyond mere vocal tract normalization. A promising route for describing how listeners compensate for such variation in speaker style is an instance-based (or exemplar) model of speech perception (Nosofsky, 1986; Kruschke, 1992).

In this kind of model (see figure 6), categories are represented cognitively as exemplars in the psychoacoustic space, a map in which the space covered by any one category is the result of actual perceptual experience. A realistic covering map will have many dimensions, corresponding to the many dimensions of the signal to which the listener attends. However, for convenience, we show only a two-dimensional covering map here, which we exemplify with the first and second formants. That is, each of these squares is a point in the listener's auditory F1-F2 space. Categories, then, are represented by distinct sets of weights which code the strength of association between a location in the psychoacoustic space and a category node. This kind of model can account for robust perception of phonemes as produced by a variety of speakers, and it can also account for robust perception of meaningful speaker categories as speakers produce a variety of phonemes.

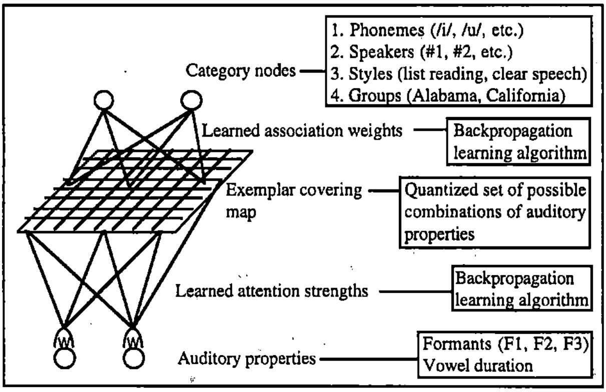
Category nodes
1. Phonemes (/i/, /u/, etc.)
2. Speakers (#1, #2, etc.)
3. Styles (list reading, clear speech)
4. Groups (Alabama, California)
Learned association weights
Backpropagation learning algorithm
Exemplar covering map
Quantized set of possible combinations of auditory properties
Learned attention strengths
Backpropagation learning algorithm
W
W
Auditory properties
Formants (F1, F2, F3)
Vowel duration

Figure 6. A model of speech perception using an exemplar covering map to simultaneously categorize several types of information conveyed by speech.

We used an implementation of the ALCOVE model described by Kruschke (1992). Each token to be categorized was defined by the frequencies of the first three vowel formants at vowel midpoint and by the vowel duration. The covering map was drawn from a set of vowel formant and duration measurements taken from a group of 39 Ohioans (thus the covering map was not representative of either Alabamians or Californians). In calculating the similarity of an input token to the locations in the covering map we used a Euclidian distance measure and a Gausian similarity function. The back-propagation method (Kruschke, 1992) was used to learn both the associations between covering map locations and categories and also the attention strengths given to the stimulus dimensions. Variable parameters in the model, a similarity scale parameter and the attention and association learning rate parameters, were selected by trial and error. With a parameter optimization algorithm we would expect to achieve better vowel classification performance than reported below but no substantial change in the patterns of category structure.

We first trained the model on the utterances produced by the Alabama speakers, a dataset which included between-speaker variation across the 13 female speakers, and also within-speaker variation between normal lab speech and an elicited clear speech style. The model achieved 74% correct vowel classification overall.

Figure 7 shows the association weights between each point in the F1/F2 covering map and the category node for the tense rounded vowel in *who'd*. The open circles plot positive weights, with size scaled to the weight magnitude, and the closed triangles plot negative weights that are substantially less than zero. The weights have a bimodal distribution reflecting the category-internal contrast between the normal lab speech list-reading style and the clear speech style, which had more peripheral F2 values. We've highlighted the two modes here by drawing ellipses around the exemplars with the highest weights. Note also the band of filled triangles just below the ellipses separating the *who'd* category from the *hood* and *owed* categories. These exemplars are the potentially most confusable members of a neighboring category and so are singled out by the training procedure for negative association weights. That is, there is a tuning of the categorization function to sharpen the category boundaries. Note also that there are no such negative weights between the two modes of the *who'd* category weights. In other words, the model does not "normalize" the hyperarticulated clear speech style to convert it to the "normal" style, but represents in the category-internal structure the natural variation actually encountered in the input data. Figure 8 shows the association weights for the vowel in *had* in this model. Again, we have drawn an ellipse to highlight the exemplars with the strongest associations to the category.

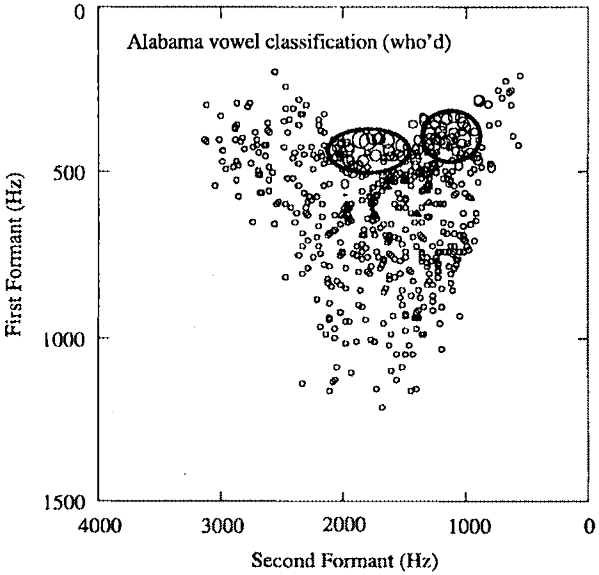

Figure 7. Association weights for who'd in the Alabama data. Each point represents a location in the exemplar covering map. Exemplars which are more strongly associated with *who'd* (large association weights) are given larger points. Points with negative association are plotted with filled triangles.

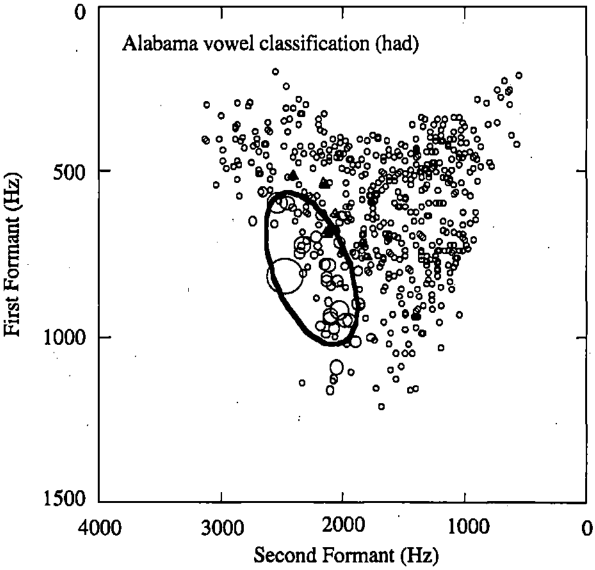

Figure 8. Association for *had* after training on the Alabama data.

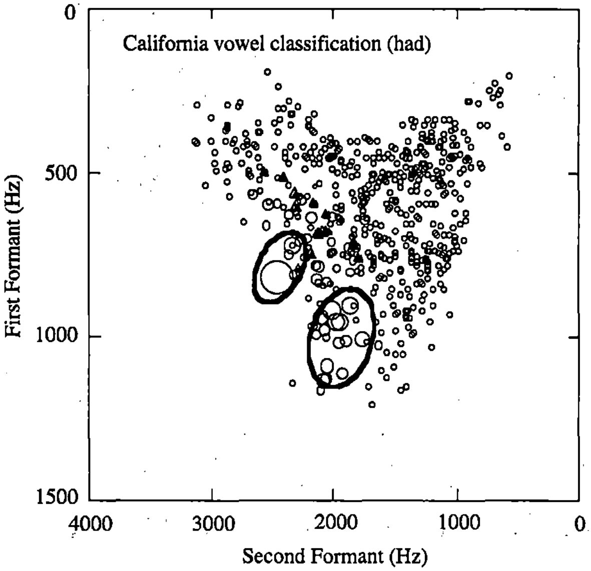

Figure 9. Association weights for *had* after further training on the California data.

We then exposed this Alabama model to utterances produced by the 7 California talkers. At first exposure, classification dropped to only 60% accuracy, but after training, accuracy rose to 77%. Figure 9 shows the association weights for *had* in the elaborated model, which begins to show more category-internal structure, with lower positive values and even some negative weights separating the two modes.

We also trained a model on an orthogonal dimension of classification - to categorize the speaker as either from Alabama or California. The F1/F2 covering map showing the association weights for the category "California speaker" is shown in figure 10. As can be seen in the figure, only tokens with very high F1 values were associated with the category "Californian".

Figure 11 shows the proportion of correctly classified tokens, across the different word types in the corpus. The open bars are for tokens produced by Alabama speakers, and the cross-hatched bars for California speakers.. Since there were more Alabama speakers in the corpus, the model adopted the general strategy of assuming that the speaker was from Alabama in the absence of evidence to the contrary. This, of course, yields 0% correct identification for the California tokens of most words. Words that exemplify vowel categories which have markedly different distributions in the F1-F2 space, however, yield much better dialect classification. In particular, tokens of *had* and *awed* have better than 50% correct classification of the speaker's dialect. The model is sensitive to just those lexical categories which differentiate the two dialects. In other words, just as real listeners do, it hones in on the sociolinguistically meaningful variability in the signal.

We have yet to analyze these corpora for idiosyncratic differences in the relative weighting of F1 and duration for differentiating head from had. Nor do we have data on whether speakers also differentially weight these dimensions in perception. However, studies such as Di Paolo & Faber (1990) and Newman (1996) suggest that there is a relationship. For example, Newman (1996) found a correlation between the average

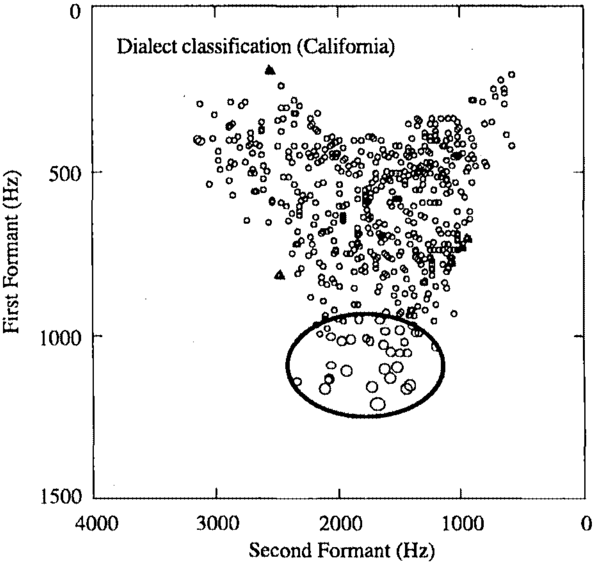

Figure 10. Association weights between locations in the vowel covering map and the speaker category "Californian" in the combined data set.

VOT in subjects' productions of /pa/ and the VOT of synthetic tokens that they rated as the best examples. Di Paolo & Faber (1990) similarly found that younger speakers of Utah English who differentiate the tense vs. lax vowels in *pool* vs. *pull* primarily on the dimension of voice quality rather than F1/F2, also can attend to that "redundant" dimension in categorizing words. We anticipate that an exemplar model can account for such patterns because speaker's own productions and the speech produced in the immediate speech community are likely to be a large component of the exemplar space.

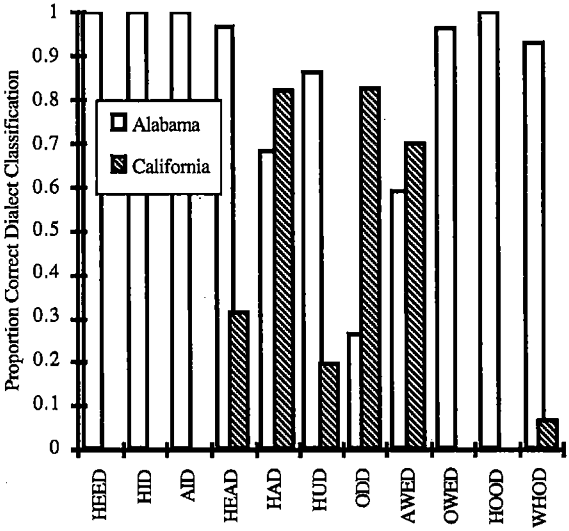

Figure 12. Proportion correct classification of dialect by word.

Conclusions

In this paper we have reviewed evidence for between-speaker differences in speech production that go beyond sex- and age-related differences in physiology. Some of these differences may be related to more subtle morphological differences such as steepness of palate doming. However, there is a substantial residue of arbitrary differences that constitute the speaker's "style". An important component of style differences is the set of differences that can be observed across a well-defined social group, and which indicate group membership. These can be perceptually salient. Thus, perceptual compensation for speaker differences must go beyond mere vocal tract normalization. A promising route for describing how listeners compensate for the arbitrary variation of style is an instance-based (or exemplar) model of speech perception in which the distribution of weights in a covering map are determined by the relative sum of exemplars that the listener encounters for each category. This works for covering maps that let the listener classify the speaker's dialect as well as for covering maps that classify the vowel category.

References

de Jong, K.J. (1995) The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, **97**, 491-504.

Di Paolo, M. & Faber, A. (1990) Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Language Variation and Change*, **2**, 155-204.

Edwards, J., Beckman, M.E., & Flechter, J. (1991) Articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America*, **89**, 369-82.

Harrington, J. & Cassidy, S. (1994) Dynamic and target theories of vowel classification: Evidence from monophthongs and diphthongs in Australian English. *Language and Speech*, **37**, 357-73.

Harrington, J. & Fletcher, J. (1996) Acoustic (non)consequences of gestural variability in the production of accentual prominence. *Journal of the Acoustical Society of America*, **100**, 2826-7.

Johnson, K., Flemming, E. & Wright, R. (1993) The hyperspace effect: Phonetic targets are hyperarticulated. *Language*, **69**, 505-28.

Kruschke, J. (1992) ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22-44.

Ladefoged, P., DeClerk, J., Lindau, M., & Papcun, G. (1972) An auditory-motor theory of speech production. *UCLA Working Papers in Phonetics*, **22**, 48-75.

Labov, W. (1966) *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

Miller, J.D. (1989) Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, **85**, 2114-34.

Nearey, T.M. (1978) *Phonetic Feature Systems for Vowels* (IU Linguistics Club, Bloomington, IN).

Newman, R.S. (1996) Individual differences and the perception-production link *Journal of the Acoustical Society of America*, **99**, 2592.

Nosofsky, R.M. (1986) Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.

Peterson, G. & Barney, H. (1952) Control methods used in a study of the identification of vowels. *Journal of the Acoustical Society of America*, **24**, 175-84.

Trudgill, P. (1974) *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press.