PAPERS FROM THE FOURTEENTH REGIONAL MEETING CHICAGO LINGUISTIC SOCIETY

APRIL 13-14, 1978

EDITED BY

DONKA FARKAS
WESLEY M. JACOBSEN
KAROL W. TODRYS

CHICAGO LINGUISTIC SOCIETY
GOODSPEED 205
UNIVERSITY OF CHICAGO
1050 E. 59TH STREET
CHICAGO, ILLINOIS 60637

TABLE OF CONTENTS

Individual Perceptual Variation and a Perception/Production Link in Vowels

Robert A. Fox
University of Chicago

The concept that a nontrivial link exists between the perception and the production systems in speech has been explicitly incorporated in a great number of perceptual models. Though space does not permit an enumeration of all the models which have specific components involving this perception/production link, these models include: Motor Theory (Liberman, Cooper, Harris, MacNeilage, and Studdert-Kennedy, 1967), Analysis-by-synthesis (Stevens and Halle, 1967), Fant's passive/active model (Fant, 1967) and the Logogen model (Morton and Broadbent, 1967). Though theoretically incompatible, they each contain components which link together some aspect of speech production with speech perception.

However, most perceptual models incorporate a perception/production link for reasons other than empirical evidence suggesting such a link. These reasons include: (1) arguments of simplicity -- since the two components of perception and production must be linked at some point, if only at the level of cogitation, then the lower the level of contact, the fewer cognitive components required; (2) lack of invariance between the acoustic wave and the speech percept -- the relationship between the acoustic wave and the articulatory shape is more easily specified than that between the speech sound perceived and the acoustic wave; and (3) theoretical framework -- the production/perception link in analysis-by-synthesis models is, in part, motivated by transformational-generative grammar. Theorists have not considered all the implications of an actual link, particularly in terms of the fact that articulatory heterogeneity, or differences between listeners in terms of their acoustic space, would imply concomitant perceptual differences. Studies have shown this to be the case for listeners differing in terms of native language (Terbeek and Harshman 1971, Terbeek 1977) and dialect (Scholes 1967) and it may therefore be the same for idiolectal variations.

This paper concentrates on two basic questions: (1) Does idiolectal perceptual variation between listeners exist? and (2) If such significant variation does exist, to what extent can these variations be predicted on the basis of the listener's own acoustic space.

An experiment was designed to investigate these questions using multidimensional scaling. The technique of multidimensional (MD) scaling has been described very completely elsewhere (c.f. Carroll and Chang, 1970) but briefly, it is an experimental procedure used to discover the perceptual dimensions or criteria used by subjects in perceiving certain stimuli. The analytic process involves obtaining perceptual distance information from listeners among a set of speech stimuli (e.g. by dyadic or triadic comparisons) and subsequently viewing the speech stimuli as points in an n-dimensional perceptual space. The researcher seeks to find the proper number of dimensions and the coordinates for the stimuli on each of these dimensions. These perceptual dimensions are considered to be the features used in perceiving the speech stimuli.

Studies using MD scaling techniques (including Hanson 1967, Singh and Woods 1971) have used the paradigm mainly to discover the important perceptual features of the speech stimuli themselves. But one particular MD analysis technique, INDSCAL, not only provides coordinate values for the speech stimuli along each perceptual dimension extracted, but indicates the importance of each dimension in the perceptual responses of each listener. The paradigm thus provides a measure by which listeners' perceptual responses and, derivatively, listeners' perceptual structures can be compared.

The experiment to be described required a group of listeners to perceptually scale six separate sets of vowel stimuli. Since each listener scaled six sets of vowels, the experimentally obtained perceptual responses could be compared both <u>within</u> listeners (across the stimulus sets scaled) and <u>across</u> listeners, allowing the use of the statistical test of analysis of variance (ANOVA).

The stimuli consisted of nine different American English vowels: /i,I,ɛ, æ, a,ʌ, o, U, u/. These vowels include most of the relatively monophthongal vowels in American English and are representative of a variety of articulatory distinctions (e.g. front/nonfront, high/nonhigh, round/nonround, closed/open, etc.). The stimulus vowels appeared in the environment /h__d/.

Dyadic comparisons were used to assess the perceptual distance between each vowel pair. Four separate similarity judgements were obtained from each listener for each vowel pair in order to establish stable perceptual distance estimates. The presentation order of the vowel pairs were randomized to reduce expectation effects. Four distinct presentation orders, each comprising all 36 vowel pairs, were constructed and each order had dummy initial and final vowel pairs to avoid scaling 'noise' related to first or last pair responses.

Six different speakers produced separate and complete stimulus vowel sets. Each of these sets were composed of four distinct trials, all differing in presentation order. The speaker groups were composed of 3 men and 3 women whose vowels were of the same phonetic value.

The scaling task used in the experiment required subjects to listen to two sequentially presented vowel stimuli and to judge the similarity or dissimilarity of the two tokens on a 9-point scale as follows:

S . . . . D

The point nearest the S represented a judgement of very similar and the point nearest the D represented a judgement of very different.

Sixteen subjects took part in the scaling experiment and this group was evenly divided into men and women. All the subjects were non-linguists. Both speakers and listeners were chosen so that they were from the same general dialect group, i.e. that the vowels which each produced were of the same phonetic quality, as determined by the ear of the experimenter. While it is strictly true that neither the listeners nor the speakers are from the identical dialect group (which would require that they all come from the same geographical area, level of social status, educational level, etc.) they cannot be differentiated in terms of the vowels which they produce. Any slight vowel differences among the speakers and listeners are thus considered to be idiolectal differences. At the end of the final scaling session, the listeners were required to read a list of words which contained the same vowels that appeared in the stimuli. Spectrograms were made of these words and measurements, including F_0, F_1, F_2 F_3 and duration, were made, thus giving an indication of each listeners acoustic space.

As mentioned earlier, these scaling data were anlysed using INDSCAL, as developed by Carroll and Chang at Bell labs. Three perceptual dimensions were extracted in the MD scaling analysis, accounting for 76% of the total variance in the data. These three dimensions are shown in figure 1 and figure 2 which are the dimension 1 by dimension 2 and dimension 1 by dimension 3 planes, respectively.

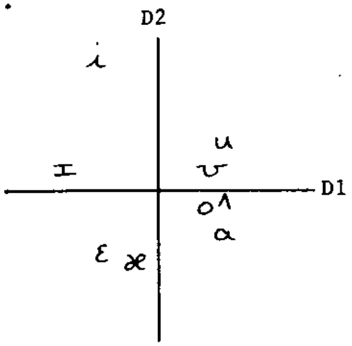

Figure 1. D1 by D2 plot

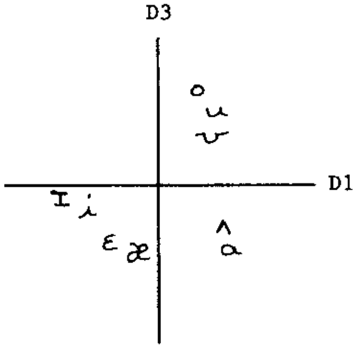

Figure 2. D1 by D3 plot

These dimensions correspond to F_2 height (front/nonfront), F_1 height (high/nonhigh), and rounding.

INDSCAL generates a subject weight, or salience, for each input perceptual distance matrix for each of the extracted dimensions. In this analysis, a subject weight was generated for each of the 6 vowel sets which each of the listeners scaled. One way analysis of variance was used to examine these subject weights for significant variance in terms of both listener and stimulus vowel set for each of the three perceptual dimensions. A summary of the results from the analyses of variance appears in tables 1 and 2.

As can be seen in Table 1, no significant variance was found in the subject weights as a function of stimulus vowel set on two of the three perceptual dimensions. Though significant variance was found in the subject weights on the third dimension, it was only at the .05 level of probability and accounted for very little of the total variance (11%). This set of ANOVAs represented statistical tests checking for significant variance <u>within</u> listeners and indicated that the listeners were consistent in their perceptual responses across the different vowel sets scaled.

Very significant variance was found as a function of listener identity on all three perceptual dimensions, as is shown in Table 2. Not only were the probability levels very low ($p < .001$ for D1 and D2, and $p < .01$ for D3) but the variance accounted for by the listener factor was extremely high; 68%, 77% and 47% on dimensions 1, 2, and 3, respectively.

These results indicate that the listeners varied significantly in their use of the perceptual dimensions extracted, thus giving a positive answer to the question concerning idiolectal perceptual variation. The next question is whether the perceptual variations can be correlated with acoustic space differences among the listeners.

In order to evaluate whether such a correlation exists, a second MD scaling analysis was done. The first type of MD analysis done was called 3-mode scaling -- the perceptual distance data were organized in a tripartite fashion: <u>vowel</u> by <u>vowel</u> by <u>perceptual</u> <u>distance</u> <u>matrix</u>. The INDSCAL solution included a salience for each of the separate matrices (i.e. one for each vowel stimulus set scaled by each listener). The second type of MD analysis used was 4-mode INDSCAL. The

Table 1. Analysis of variance results for stimulus vowel set factor

Dimension 1

Source	SS	DF	MS	F
Vowel sets (between groups)	.0335	5	.0067	$F < 1.0$
Error (within groups)	.4740	60	.0079	Not signficant
Total	.5075	65		

Dimension 2

Source	SS	DF	MS	F
Vowel sets (between groups)	.0135	5	.031	$F < 1.0$
Error (within groups)	.9790	60	.0163	Not significant
Total	.9945	65		

Dimension 3

Source	SS	DF	MS	F
Vowel sets (between groups)	.0587	5	.0117	$F = 2.37$
Error (within groups)	.2962	60	.0049	$p < .05$
Total	.3549	65	Variance accounted for: 11%	

Table 2. Analysis of variance results for listener factor

Dimension 1

Source	SS	DF	MS	F
Listeners (between groups)	.3545	10	.0354	$F = 13.7$
Error (within groups)	.1427	55	.0026	$p < .001$
Total	.4972	65	Variance accounted for: 68%	

Dimension 2

Source	SS	DF	MS	F
Listeners (between groups)	.7897	10	.0796	$F = 21.0$
Error (within groups)	.2058	55	.0037	$p < .001$
Total	.9955	65	Variance accounted for: 77%	

Dimension 3

Source	SS	DF	MS	F
Listeners (between groups)	.1874	10	.0187	$F = 6.56$
Error (within groups)	.1572	55	.0029	$p < .01$
Total	.3446	65	Variance accounted for: 47%	

input distance data was the same in this type of analysis but are organized vowel by vowel by stimulus vowel set by listener. The program generates weights for each vowel set and each individual listener for each of the perceptual dimensions extracted. The listener weights produced in the solution are a measure of the reliance by the listener upon the perceptual dimensions in their scaling responses.

The perceptual dimensions extracted in the 4-mode INDSCAL solution were practically identical to those of the 3-mode solution so both types of analysis locked into the same basic solution.

In order to examine the relationship between the utilization of the perceptual dimensions (indicated by the listener weights) and the acoustic vowel space of each listener (indicated by the acoustic measures made of the listeners' vowels) the technique of stepwise multiple linear regression was used. This technique indicates how well sets of independent variables predict a set of dependent measures. Specifically, how well various measures of the listener's acoustic space predict the listener weights. In the multiple regression analyses to be discussed, the listener weights on each of the perceptual dimensions were the dependent variables, while various sets of measures of the listener's acoustic vowel space were used as the independent variables.

Throughout the regression analyses on the listener weights from the three perceptual dimensions we will be primarily interested in the degree of prediction afforded by the following 7 groups of acoustic measures: (1) and (2) The formant 1 and formant 2 values of the corner vowels /i,u,a/, both including and excluding F_0. (3) and (4) The formant 1 and formant 2 values of a set of noncorner vowels /æ, ʌ,o/ both including and excluding F_0. (5) The values of formant 1 for all 9 vowels. (6) The values of formant 2 for all vowels. (7) The values of formants 1 and 2 for all 9 vowels, plus F_0 (this gives an indication of the best predictors of the listener weights among all the actual acoustic vowel measurements for each listener.

One should keep in mind what these analyses mean apart from the mathematical manipulations and the statistical abstractions involved. The issue is the degree of relationship between the acoustic space of a listener and his or her perceptual space. The hypothesis being evaluated is whether a listener's perceptual structure utilizes information about the listener's own vowels in making phonetic categorizations. This evaluation will be done by examining the listener saliences, since these saliences represent the degree to which each listener is utilizing the extracted perceptual dimensions. If use of the dimensions by a subject in perceiving a vowel token varies as a function of the subjects's own acoustic space, then the listener weights should reflect differentiation among the listeners in terms of their acoustic vowel space.

The results of the regression analyses done on the dimension 1 data (the dimensions corresponding to F_2 height, or front/nonfront) are shown in figure 3. Each column represents separate analyses using different sets of independent variables. Each successively higher data point (vertically) within each column indicates the value of the multiple correlation coefficient between the actual listener weights and the predicted values using only the independent variables up to and including that point in the regression equation.

More attention should be paid to the prediction afforded by 1-3 variables than that obtained using a larger number of variables. Since, though an increase in the number of variables improves the prediction, the additional variables account for progressively smaller degrees of improvement, particularly when the independent variables are intercorrelated as the acoustic measures are.

Column 1 in figure 3 shows the regression results using the formant 1 and formant 2 values of the vowels /i,a,u/ for each listener. The predictive value of the corner vowels are of interest because these vowels normally exhibit the extreme values of the first and second formants for any particular speaker, and are, therefore, good indications of the range and position of each speaker's acoustic space. Using the three variables /F_2u, F_1u, and F_2i/ a multiple correlation of .678 was obtained; use of all 6 of the independent variables gives a multiple correlation of .845. Column 2 show the results using F_1 and F_2 of /i,a,u/ plus F_0. Including F_0 improves the prediction both at the three variable level, where the multiple correlation equals .823 and at the final level, when the multiple correlation equals .907.

Columns 3 and 4 show the results of the regression analyses using the formant 1 and 2 values of a set of non-corner vowels /æ, ʌ, o/ both with and without F_0 respectively. Though the formant values of these vowels for any particular speaker will indicate something about their acoustic space, it does not give as clear an estimate of the entire vowel space as do the corner vowel measures. It is clear that the use of the corner vowel measures affords much better prediction than does the use of non-corner vowel measures. Without F_0 the best prediction that can be made with the non-corner vowels yields a multiple correlation of .584 (which accounts for only .34% of the variance).

Since the vowel loadings on dimension 1 are related to the 2nd formant height of the vowels, it is important to examine to what extent the listener weights are related to the formant 2 heights of the listener's vowels. These results are shown in column 6. Three variables give a multiple correlation of .744 and represents a very noticeably better prediction than that encountered in columns 3 and 4. At the 6 variable level a multiple r of .988 was obtained.

This multiple correlation is the best obtained yet at the 6 variable level. However, at the three variable level the prediction using /F_2u, F_2i and F_2I/ is not as good as that using the corner vowel measures /F_2i, F_2u/ and F_0.

The results of the regression analysis using only the formant 1 values is shown in column 5. Using formant 1 values, the multiple correlation value at the three variable level is .487 and at the 7 variable level, .773. The values of F_2 predict the listener weights better than do the values of F_1. This is very suggestive of an acoustic space/perceptual space link since dimension 1 (front/nonfront) reflects second formant height.

Column 7 shows the regression analysis using all the directly measured acoustic variables including the values of formant 1 and formant 2 for all vowels and F_0 for each listener. This analysis essentially duplicates the results shown in column 2. The three best predictors were F_0 and the F_2 values of the corner vowels /i,u/. The F_2i and F_2u serve to give the maximum and minimum values for formant 2 for any speaker.

Figure 4 is the plot of the results of the regression analyses using the dimension 2 listener weights as dependent variables. Columns 1 and 2 again show the results using the corner vowel measures. The goodness of the prediction is very similar to that seen for the dimension 1 listener saliences. The best single predictor is F_1a, giving amultiple correlation of .386, with the best three prdictors being F_1a, F_2a and F_1i yielding a multiple correlation equal to .747.

A comparison of columns 1 and 2 should be made with columns 3 and 4 (which show results of regression analysis using non-corner vowels). As in the case

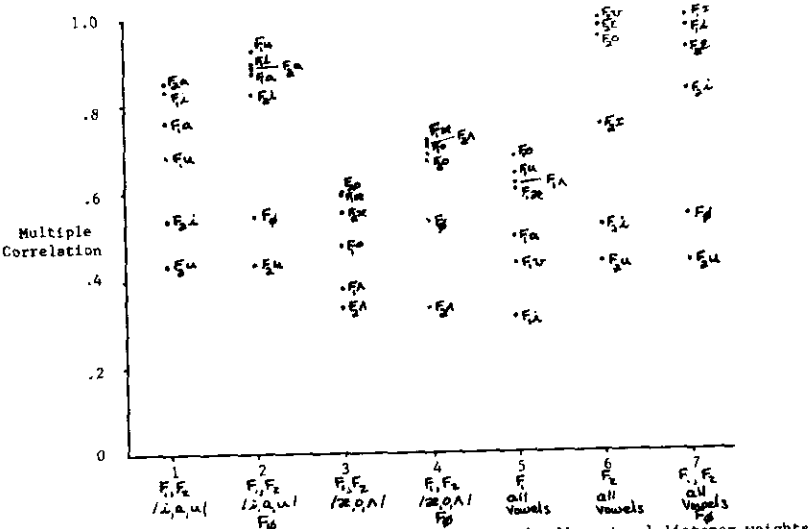

Figure 3. Results from the regression analyses done on the dimension 1 listener weights

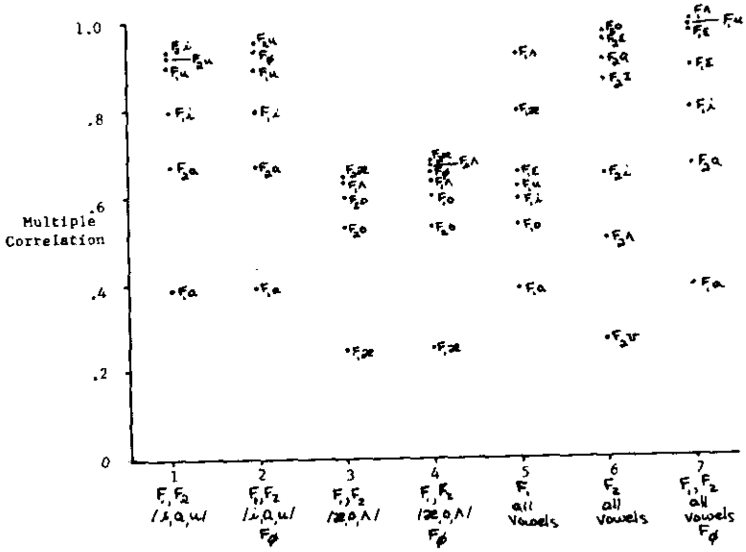

Figure 4. Results from the regression analyses done on the dimension 2 weights

of the saliences on dimension 1, the corner vowels provide significantly better prediction of the listener weights than do the non-corner vowel measures. The best single predictor using the non-corner vowel measures is F_1æ, which give a multiple correlation of .245. At the three variable level the best predictors are F_1æ, F_2o and F_1o (which yields a multiple correlation of .599).

Columns 5 and 6 show the results using the F_1 and F_2 measures of all the listeners' vowels, respectively. Since the 2nd dimension (high/nonhigh) corresponds to the formant 1 values of the stimulus vowels, the F_1 values may predict the listener weights better than do the F_2 values; a situation analogous to that found in the regression analyses found for the dimension 1 weights. The formant 1 measures predict the listener saliences better than the formant 2 values using only 2 variables, though the first and second formant measures give a similar multiple correlation in predicting the listener weights at the 3 variable level.

Column 7 shows the results of the regression analysis using all F_1 and F_2 values plus F_0. The three variables which best predict the listener saliences are measures from the corner vowels, specifically, the best predictors were F_1a, F_2a and F_1i. Note that F_1a usually exhibits the maximum value of the first formant for any particular vowel in a speaker's acoustic space (analogous to the value of F_2 of /i/ for any speaker's formant 2 values. The value of F_1i usually represents a value which is the lowest value for formant 1 in any speaker's vowel space. This is again suggestive that a relationship exists between the formant values of the corner vowels and the listener's use of the perceptual dimensions.

Figure 5 shows the results of the multiple regression analyses done on the dimension 3 (round/nonround) saliences. Columns 1 and 2 show the results

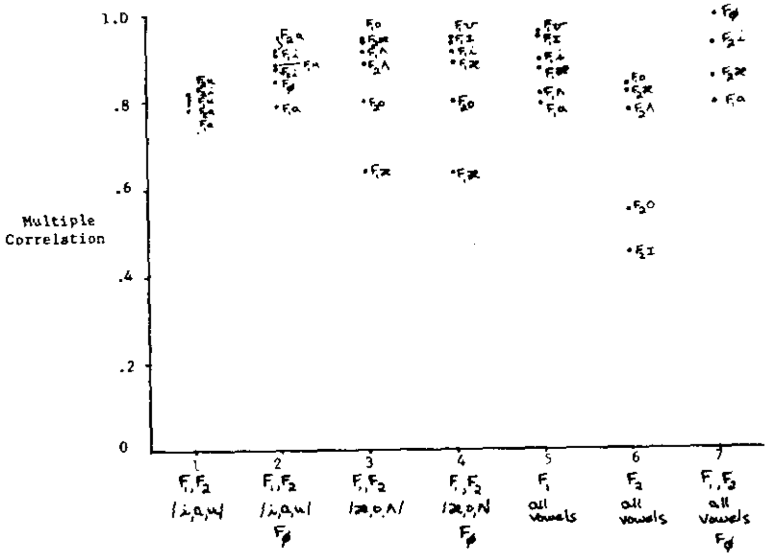

Figure 5. Results from the regression analyses done on the dimension 3 weights

using the corner vowel measures. The single best predictor is F_1a which yields a multiple correlation of .784. The prediction at the three variable level in column 1 (using the variables F_1a, F_2a, and F_1u) gives a multiple *r* of .802. The three best predictors in column 2 are F_1a, F_0, and F_2i, giving a multiple correlation of .875.

A comparison of columns 1 and 2 with columns 3 and 4 does not show a significantly better prediction afforded by the corner vowel measures as opposed to the non-corner vowel measures which has been encountered previously. In a reversal of previous results on dimensions 1 and 2, it can be seen that the non-corner vowel measures actually predict the dimension 3 saliences *better* than do the corner vowels.

Columns 5 and 6 give the regression results using only the formant 1 and 2 values, respectively. The values in column 5 are noticeably higher than column 6, thus the formant 1 values are better predictors of the saliences on this dimension than are the formant 2 values. Turning to column 7 it can be seen that use of 6 independent variables gives a multiple correlation of greater than .999, while at the three variable level, a multiple correlation of .925 is obtained (using F_1a, F_2æ, F_2i). Unlike the previous two dimensions, there does not seem to be a special relationship between the corner vowels and the listener saliences, though a relationship between the listener saliences and the listener acoustic space may exist.

This result is not too surprising. The corner vowels give an indication of the maximum and minimum values of formant 1 and formant 2 -- while F_0 (which figured prominently in the regression analyses as an independent variable) may indicate the scaling factor of the formant normalization needed (since males and females tend to have differentially sized vocal tracts and F_0 can give an indication of the speaker's sex). However, dimension 3 is a rounding dimension, which does not relate to either formant 1 or formant 2 height directly. Rather, this dimension relates to a generally overall lowering of the formants, thus the corner vowels may not be expected to be the best predictor of this dimension.

The results presented show that very suggestive correlations were found between the formant values of the corner vowels (/i,a,u/) and the listener saliences on dimension 1 (related to formant 2 height) and dimension 2 (related to formant 1 height). In particular, two of the best predictors for listener saliences on dimension 1 were F_2i and F_2u. These two vowels exhibit the maximum and minimum values of F_2 for any particular speaker's acoustic space. Two of the best predictors for the listener saliences on dimension 2 were F_1a and F_1i. These vowels exhibit the maximum values of formant one for any particular listener's acoustic space. Thus the saliences of these perceptual criteria for each listener were significantly related to an important measure of the listener's acoustic vowel space.

Since the scaling paradigm extracts perceptual information at the level of phonetic processing, the results therefore suggest that the perception/production link is at the level of phonetic classification. The question of exactly 'how' this link is realized at the phonetic level is a more difficult one -- if only because it is unclear as to exactly what constitutes this level of perception. A very large number of experimental studies (including Hanson 1967, Terbeek 1977, Terbeek and Harshman 1971) have suggested that the phonetic quality of vowels is determined, or assigned, on the basis of perceptual features. In other words, the mental acoustic image of a vowel token to be identified is analyzed in terms of these distinctive features (not necessarily those of any current theory) and is classified into a phonetic category on the basis of the values of these n-ary features. The correlation between production and perception could take the form of a correspondence between the feature values for the listener's own vowels and the feature values required to produce the same phonetic quality response in the perceptual system.

REFERENCES

Carroll, J. D. and Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckert-Young decomposition. Psychometrika 35:283-319.

Fant, G. (1967). Auditory patterns of speech. in Models for the Perception of Speech and Visual Form. Wathen-Dunn (ed.). M. I. T. Press. Cambridge.

Hanson, G. (1967). Dimensions in speech sound perception. Ericcson Technics 23.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P. and Studdert-Kennedy, M. (1967). Some observations on a model for speech perception. Models for the Perception of Speech and Visual Form. Wathen-Dunn (ed.). M. I. T. Press. Cambrige.

Morton, J. and Broadbent, D. (1967). Passive versus active recognition models, or Is your homunculus really necessary? Models for the Perception of Speech and Visual Form. Wathen-Dunn (ed.). MIT Press. Cambridge.

Scholes, R. (1967). Perceptual categorization of synthetic vowels as a tool in dialectology and typology. Language and Speech 10.

Singh, S. and Woods, G. (1971). Perceptual structure of American English vowels. J. Acous. Soc. Am. 49:1861-1866.

Stevens, K. N. and Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. Models for Perception of Speech and Visual Form. Wathen-Dunn (ed.). M. I. T. Press. Cambridge.

Terbeek, D. (1977). A Cross-Language multidimensional scaling study of vowel perception. UCLA Working Papers in Phonetics.

Terbeek, D. and Harshman, R. A. (1971). Cross-language differences in the perception of natural vowel sounds. UCLA Working Papers in Phonetics 19.