PAPERS FROM THE PARASESSION ON LANGUAGE AND BEHAVIOR CHICAGO LINGUISTIC SOCIETY

MAY 1-2, 1981

EDITED BY
CARRIE S. MASEK
ROBERTA A. HENDRICK
MARY FRANCES MILLER

CHICAGO LINGUISTIC SOCIETY
THE UNIVERSITY OF CHICAGO
CLASSICS 314A
1050 E. 59TH ST.
CHICAGO, ILLINOIS 60637

Perceptual Structure of English Monophthongs and Diphthongs

Robert Allen Fox
The Ohio State University

Much effort has been directed at discovering the perceptual features or dimensions utilized in the perception of American English vowels. Evidence on the nature of these features have been obtained primarily through the use of multidimensional scaling--the features extracted in such studies corresponding to traditional linguistic distinctions such as front/back, high/low, and round/nonround (Hanson, 1967; Singh & Woods, 1970; Shepard, 1972; Terbeek, 1977; Fox, 1978). However, a problem arises in terms of understanding the perception of vowel quality which changes over a short course of time. The studies noted above ostensibly examined the perceptual structure of monophthongs, though obviously in real speech these vowel qualities do change dynamically, especially in so-called diphthongized vowels such as [eɪ oo]. In addition, little effort has been made to include true diphthongs (such as [aɪ aʊ ɔɪ]) in such studies.

One experiment directed specifically at this general question was carried out by Terbeek (1974; Terbeek and Fox, 1975). This was an unpublished MD-scaling experiment examining the perceptual structure of a set of 9 diphthongal vowels which included three vowels with [i] offglides ([eɪ aɪ ɔɪ]), three vowels with [u] offglides ([u oo aʊ]) and three vowels with level or schwa-like offglides ([ɪə æə ɔə]). A six-dimensional solution was obtained, but was relatively uninterpretable. No clear-cut front/back dimension emerged and though a dimension reflecting a high/low distinction (F1 onset) was obtained, it represented the least salient of the six dimensions. The most salient property seemed to be presence of rounding anywhere in the vowel vs. complete absence of rounding, while other features defied a straightforward acoustic interpretation. Such results would suggest that a set of features completely different from those used to explain monophthong perception is required for nonsteady-state vowels, but this conclusion is both unappealing and unlikely. More realistically, these results reflect some fortuitous skewing factor affecting subject responses and say little about the nature of diphthong perception. A separate problem is that of understanding the interaction of vowel duration and perceptual features. It is remarkable that duration has not appeared as a perceptual dimension in any of the scaling experiments given that many studies (Tiffany, 1953; Ainsworth, 1972; Mermelstein, 1977; Verbrugge & Isenberg, 1978) have shown that durational differences can affect identification of vowel quality--particularly for tense/lax vowel pairs. However, it is not clear whether the lack of a duration dimension is a function of the perceptual structure being modeled or the nature of the stimulus sets. For example, some studies (e.g., Pols et al., 1969) equalized durations over all vowels, while other studies (e.g., Singh & Woods, 1969) simply have not reported durations and it is unclear whether such durational differences were actually present.

This paper describes a multidimensional scaling study designed to discover the perceptual structure of both monophthongs and diphthongs in American English using a much larger and more complete vowel stimulus set than any of the previously cited experiments. Perceptual distance data were obtained through dyadic comparisons and were then analyzed using INDSCAL.

The stimuli consisted of 15 American English vowel [i ɪ eɪ ɛ æ ɑ ʌ ɔ oʊ o ʊ ɔɪ aɪ aʊ ju] representative of the nonrhotic vowels in the speech of a typical Midwestern speaker (Ladefoged, 1975). The vowels were produced by a male speaker in the phonetic environment [h__d]. The stimulus tokens appeared in pairs, 300 msec apart while different stimulus pairs were separated by 5 sec. There were 105 different vowel pairs, disregarding vowel sequence within a pair. Four separate similarity judments were obtained from each subject for each vowel pair in order to establish stable perceptual distance estimates. Presentation order of vowel pairs were pseudo-randomized (i.e., random order except for the requirement that no vowel appear twice in consecutive pairs) across the 4 different trials.

Since the goal of this study was to determine the features used in perceiving vowels without presupposition as to the nature of these features, an effort was made to allow the vowel tokens to resemble those which normally appear in speech, therefore vowel quality for most vowels was not steady-state throughout and no effort was made to equalize durations across different vowel qualities.

The steady-state vowel tokens used in scaling experiments are usually described in terms of mean F1 & F2--however, since we are using vowels changing in phonetic quality in time, it is necessary to capture the acoustic characteristics of such vowel tokens in a more robust manner. To this end each vowel was divided into 4 equal portions as shown in Fig. 1., yielding 5 separate reference points within each vowel. Point 1 represents the onset of F1 (considered the beginning of the vowel) and point 5 the onset of the stop closure. Points 2, 3, 4 are equidistant between points 1 and 5. Acoustic measurements made for each vowel included F1, F2, F2-F1, and F3 at points 1-4, extent of change in these formant measures from points 1-4 (both signed and unsigned values), F0, and duration. Point 5 was not included since formant values here reflect the transition to the following dental stop. Fig. 2 is a F1 x F2 plot of the stimulus vowels.

Twenty-four subjects, all undergraduate students at OSU, participated in the experiment. All subjects were from central Ohio and spoke a Midwestern dialect of American English. The scaling task required these subjects to listen to the two vowels of a stimulus pair and to judge the similarity/dissimilarity of the tokens on a 9-point scale. Each subject's similarity judgments were checked for consistency before they were included in the INDSCAL analyses. This was done to ensure that excessive 'noise' was not included in the perceptual distance information submitted to the analysis program.

The primary analytic tool utilized in this study was multidimensional scaling. This procedure involves viewing similarity judgments between each of a set of objects and accounting for these estimated distances in terms of perceptual dimensions or features underlying the subjects' perceptual responses. Though discussed at length elsewhere (e.g., Kruskal & Wish,

(1978), MD-scaling models the perceptual space--this space composed of the set of perceptual dimensions underlying the subject's similarity judgments and, by extension, the vowel features normally used in perception. INDSCAL, the procedure used here, is a particularly powerful program in that it considers not only intervowel distance estimates, but differences across individual subjects which, in turn, allow INDSCAL to avoid problems such as orientation of the resulting spatial model. This permits the researcher to make a stronger claim concerning the psychological reality of the solution. For the purpose of this short paper suffice it to say that INDSCAL analysis extracts perceptual features from the data which have a strong claim to psychological reality and reflect features utilized by subjects during the perceptual process.

INDSCAL analyses were done using the entire 15 vowel stimulus set as well as 4, 9, and 11 vowel subsets (in order that these results could be compared to appropriate results in the literature and to see the degree of difference between the perceptual space for monophthongs alone and the combined space). Analyses were done in 1-9 dimensions and the 4-dimensional solution was found to provide the best overall representation of the perceptual structure of the stimuli. Since one of the most crucial decisions in any INDSCAL study involves a judgment as to what the 'correct' dimensionality of the solution is, and because several other studies (e.g., Terbeek, 1974, 1977; Terbeek & Fox, 1975) have extracted a larger number of dimensions for an equal or smaller number of vowels, some defense of this dimensionality is in order. The fit curve shown in Fig. 3 (indicating the cumulative proportion of variance accounted for --a traditional guide in such a decision) gives relatively little relevant information. Normally one looks for an 'elbow' indicating a steep decrease in variance accounted for from one dimension to the next. However, several other criteria were more useful. Interpretability required that no more than 4 dimensions be extracted since at dimensionality 5 and above the clear perceptual structure indicated at lower dimensionalities degenerated. Four dimensions also represent the upper limit on the stimuli to number of dimensions ratio (i.e., number of input data values/number of dimensional coordinates) recommended by Kruskal & Wish (1978) for 15 stimuli. In addition, a version of Gandour & Harshman's (1978) 'split-half' procedure was employed which indicated that these four dimensions were reliably replicated in INDSCAL analyses of separate halves of the data. In this procedure the data are divided into two halves, INDSCAL analyses done on each separate half at a number of dimensionalities, and the dimensions from these separate analyses compared. Any dimension which does not replicate probably accounts only for noise in the data. This procedure, along with the discovery of what seemed to be multiple INDSCAL solutions at 5 dimensions supported selection of the 4-dimensional solution.

This solution is presented in Fig. 4 and accounts for about 65% of the variance with the dimensions numbered in terms of order of appearance. The most salient perceptual dimensions are D1 and D2, accounting for 32% and 18% of the variance, respectively. The two remaining dimensions, D3 and D4 account for somewhat less of the variance (8% and 6%, respectively). In traditional terminology one could label D1 as front/back, D2 as high/low, and D3 as low-back onset. D4 does not readily yield to

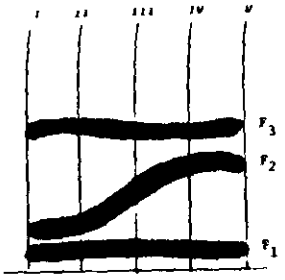

Figure 1. Formant measurements made on each vowel.

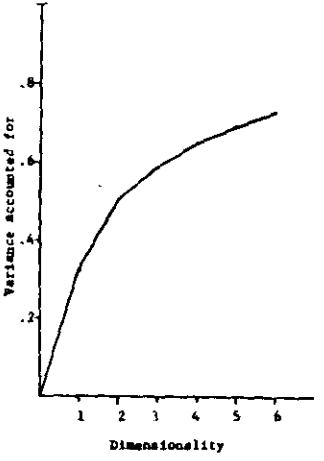

Figure 3. Fit curve for INDSCAL solutions in 1-6 dimensions.

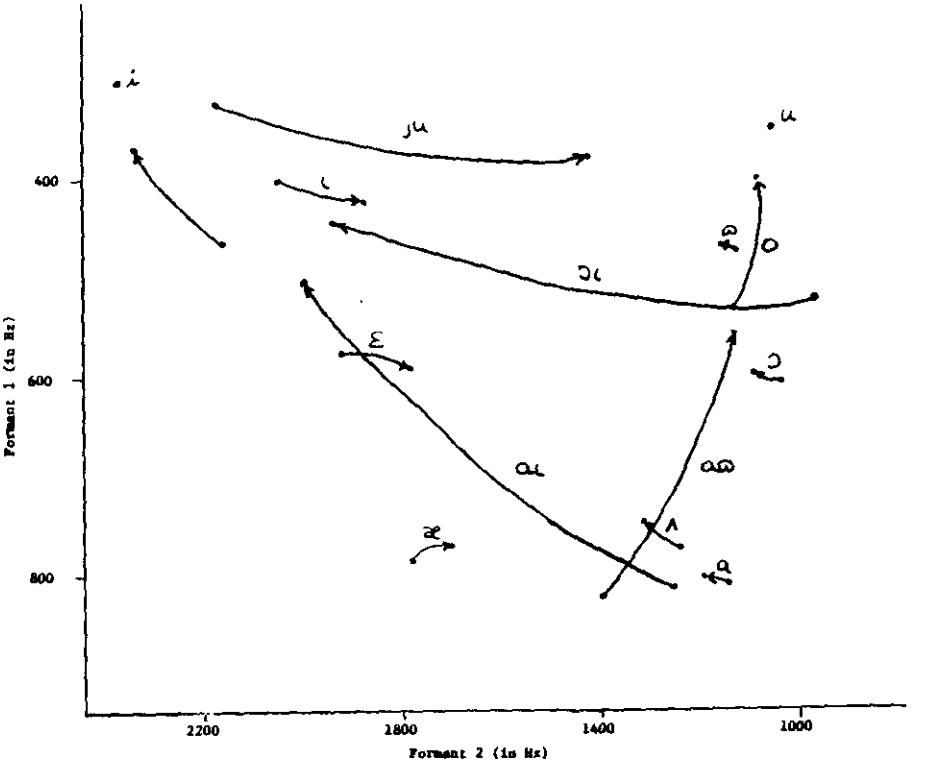

Figure 2. F1 x F2 plot of the stimulus vowels.

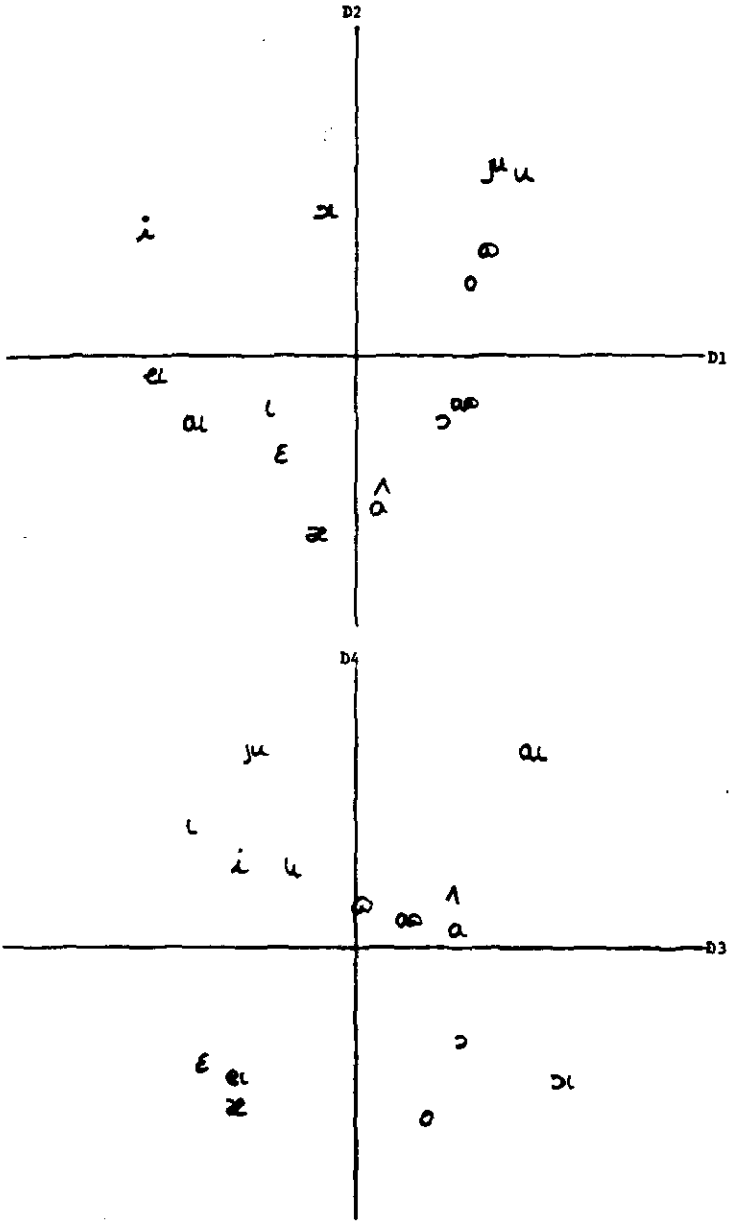

Figure 4. 4-dimensional INDSCAL solutions. Plots of D1 x D2 & D3 x D4.

any such interpretation. How do these dimensions compare to various distinctive features? Table 1 summarizes the results of regression analyses of the vowel coordinates on each dimension against a set of categorical distinctive features including Chomsky & Halle's (1968) binary High, Low, Tense, and Round; and Ladefoged's 4-level Height feature (marked LHigh).

Table 1. Summary of regression analyses between perceptual dimensions and selected distinctive features.

Dim.		Feature	Simple r	Multiple r	Variance acct.for	F-test for individual regression variables
D1		Back	.866	.866	.750	$F=26.98, df=1/9, p<.001$
D2	(a)	High	.755			$F=21.47, df=1/7, p<.005$
		Low	.515[a]	.880	.775	$F=4.62, df=1/7$, n.s.
		Tense	-.482[a]			$F=2.12, df=1/7$, n.s.
	(b)	LHigh	.923	.923	.851	$F=51.50, df=1/9, p<.001$
D3		Back	.890	.947	.896	$F=57.90, df=1/8, p<.001$
		High	-.707[a]			$F=7.99, df=1/8, p<.05$
D4		High	.775	.775	.612	$F=13.57, df=1/9, p<.01$

[a]These are partial correlations, i.e., the variance accounted for by the other variables in the regression equation have been partialled out.

Shown are the most significant predictors. The regressions were done on the basis of the 11 vowels which are not true diphthongs since there may be some disagreement as to the proper featural markings for the diphthongs. D1 is most significantly related to Back, while D2 is best related to Ladefoged's Height feature. Note that the combination of the binary features High, Low and Tense cannot predict the D2 coordinates as well as does Ladefoged's multivalued feature. In terms of Tense, it did not appear as a significant predictor of any perceptual dimension, nor did it appear as a separate dimension in the scaling experiments cited earlier, calling into question its reality as a psychologically real perceptual feature. D3, which we have labelled low-back onset, is significantly related to both Back and the inverse of High. D4 is correlated significantly with High, but clearly at a relatively marginal level and it does not account for much of the variance on that dimension. In terms of the diphthongs, [aɪ] patterns with the mid-front vowels, [aʊ] with the mid-back vowels, and [ju] with [u]. The vowel [ɔɪ] is judged as high, but neither front nor back. Though not shown, a separate INDSCAL analysis (reliably replicated) was done using perceptual distance estimates between diphthong pairs only. This resulted in a single front/back dimension with the same ordinal pattern as in D1 and accounted for almost half of the interdiphthong variance.

How do these perceptual dimensions relate to the acoustic structure of the vowels, and in particular, is there any evidence that listeners are making phonetic similarity judgments on the basis of direction or extent

of formant movement or duration? A gross indication of these relationships can be determined by correlating the vowel coordinates on each dimension against the acoustic measures described earlier. Many of these correlations are shown in Table 2. As can be seen, D1, the

Table 2. Correlations between perceptual dimensions and selected acoustic measures.

Measure	D1	D2	D3	D4
F1 I	-.01	-.796*	.519	-.161
F1 II	-.011	-.785*	.529	-.112
F1 III	-.041	-.811*	.497	-.128
F1 IV	-.110	-.833*	.262	-.228
F2 I	.575	.017	-.812*	.166
F2 II	.672	-.050	-.798*	.100
F2 III	.858*	-.048	-.632	.005
F2 IV	.935*	.003	-.331	.002
F2-F1 I	.477	.262	-.831*	.187
F2-F1 II	.561	.213	-.827*	.124
F2-F1 III	.728	.225	-.689	.045
F2-F1 IV	.843*	.245	-.363	.068
Delta F1 u	.103	-.104	.472	.206
Delta F1 s	-.142	.196	-.543	-.039
Delta F2 u	.092	.323	.405	.264
Delta F2 s	.362	-.019	.634	-.205
Duration	.078	.155	.461	.038
F0	.018	-.059	.194	.521

*$p < .001$

front/back dimension was most significantly related to F_2IV, accounting for almost 88% of the variance on that dimension, thus suggesting that subjects were attending to F2 offglide in their perceptual judgments. D2, the height dimension, was inversely related to F1, with the best acoustic predictor being F_1IV. Regression analyses indicated that prediction of D2 could be increased by including F_2I and duration into the regression equation, but the increase in variance accounted for was small. D3, the low-back onset dimension, was best correlated with F2-F1 (and F2) measures; in addition, suggestive correlations were obtained with the F_1I & F_1II measures. Regression of D3 against various acoustic measures revealed that signed change in F_2 from point 1-4, in conjunction with F2-F1 II, made a significant increase in the prediction of D3 thus suggestive that it is serving as as salient acoustic cue--however, the significance of Delta F_2 signed was at the .05 level only (F-test for individual variable, $F=3.67$, df=1/12, $p<.05$). D4 was not significantly correlated with any acoustic measure at the .01 level or better. The correlation between D4 and F_0 is marginally significant (at the .05 level) but given the number of correlations done on the data does not indicate a significant relationship.

In summary, there was no evidence that subjects use radically different perceptual features when perceiving diphthongal vowels as opposed to monophthongal vowels--the resulting perceptual dimensions were very similar to those from monophthongal studies. Even including

true diphthongs in the vowel stimuli did not produce a separate dimension which reflected dynamic information such as change in F1 or F2, though there was some evidence that subjects were attending to change in F2 in making their similarity judgments. Duration still did not appear as a separate dimension, though it did contribute (though at only a marginally significant level) to the prediction of at least one dimension (D2) emphasizing its role as a supplementary rather than primary cue in vowel identification.

The perceptual results here again indicate the important role which formant peaks play in the perceptual process, even though some theorists (e.g., Klatt, 1979) have questioned the assertion that formant frequencies represent psychological real dimensions in perception. However, given that INDSCAL results have such strong claims to representing psychologically real phenomenon, and the fact that studies utilizing different stimulus sets as well as perceptual tasks have obtained very similar results, it is very difficult to deny a conclusion concerning the importance of features defined in terms of formant frequencies at some level of perceptual processing.

Acknowledgments

I would like to thank the College of Humanities, The Ohio State University, for providing computer funds for this project.

References

Ainsworth, W. (1972) Duration as a cue in the recognition of synthetic vowels. J. Acoust. Soc. Am., 51:648-651.

Chomsky, N. and Halle, M. (1968) The Sound Pattern of English. Harper and Row.

Fox, R. (1978) Individual perceptual variation and a perception/production link in vowels. Papers from the 14th Meeting, CLS.

Gandour, J. and Harshman, R. (1978) Crosslanguage differences in tone perception: A multidimensional scaling investigation. Lang. Sp. 21:1-33.

Hanson, G. (1967) Dimensions in speech sound perception: An experimental study of vowel perception. Ericsson Technics 23:175 pp.

Klatt, D. (1979) Speech perception: A model of acoustic-phonetic analysis and lexical access. J. Phon. 7:279-312.

Kruskal, J. and Wish, M. (1978) Multidimensional Scaling. Sage University Papers.

Ladefoged, P. (1975) A Course in Phonetics. Harcourt Brace Jovanovich.

Mermelstein, P. (1977) On the relationship between vowel and consonant identification when cued by the same acoustic information. Haskins SR-51/52, 201-212.

Pols, L., van der Kamp, L., and Plomp, R. (1969) Perceptual and physical space of vowel sounds. J. Acoust. Soc. Am. 46:458-467.

Shepard, R. (1972) Psychological representation of speech sounds. In David and Denes (eds), Human Communication: A Unified View, NY: McGraw-Hill.

Singh, S. and Woods, G. (1970) Perceptual structure of 12 American vowels. J. Acoust. Soc. Am., 49:1861-1866.

Terbeek, D. (1974) Multidimensional scaling of the perception of diphthongs. Presented to The Classification Society, April 29-30.

Terbeek, D. (1977) A cross-language multidimensional scaling study of vowel perception. Working Papers in Phonetics, UCLA, 37.

Terbeek, D. and Fox. R. (1975) INDSCAL study of the perceptual space of American diphthongs. Presented at the 90th Meeting, Acoustical Society of America, J. Acoust. Soc. Am., 58:S91 (A).

Tiffany, W. (1953) Vowel recognition as a function of duration, frequency modulation, and phonetic context. J. Sp. Hear. Res. 18:289-301.

Verbrugge, R. and Isenberg, D. (1978) Syllable timing and vowel perception. Haskins SR-55/56, 113-122.