THE PSYCHOLOGICAL REPRESENTATION
OF VOWEL SOUNDS

Henri Cohen

A Thesis
in
The Department
of
Psychology

Presented in Partial Fulfillment of the Requirements
for the degree of Master of Arts at
Concordia University
Montréal, Québec, Canada

November, 1978

ABSTRACT

THE PSYCHOLOGICAL REPRESENTATION OF VOWEL SOUNDS

Henri Cohen

This thesis examines the relationship between physical properties of vowel sounds and their psychological representation as given in a multidimensional scaling analysis. Subjects in the study were asked to judge the degree of dissimilarity between pairs of synthetic vowel sounds in three phonetic contexts: in isolation, embedded in /p-p/ and in /h-d/. Multidimensional scaling analyses yielded a final configuration in five dimensions in /h-d/ and in four dimensions in the other two contexts. The first three dimensions corresponded best to the features of tongue advancement, vowel height and discrimination between monophthongs and diphthongs and these were found to be invariant across the three contexts examined. Contextual effects were found to be significant for the fourth dimension in that the feature tenseness seemed to best account for sounds presented in /p-p/ and h-d/ only. The fourth dimension in the context of isolated vowels and the fifth in /h-d/ could not be defined. These findings suggest that all the dimensions in the psychological representations that have been defined were derived from information carried in the first two lower formant frequencies. The results are interpreted in light of the current literature on vowel perception. Extensions of the technique used in the study are proposed.

Acknowledgements

I would like to sincerely express my appreciation to Dr. Norman Segalowitz for his support, encouragement and advice throughout the research and composition of this thesis.

I would also like to extend my gratitude to Dr. Melvin Komoda whose suggestions and criticisms were invaluable, and to Dr. Ching Suen for making available the speech synthesizer and for his comments and suggestions.

Thanks are also due to Dr. John Christian and Dr. Jim Ramsay for their assistance in statistics and to Jeff Mulherin for designing the computer programs.

Table of Contents

Table of Contents (cont.)

List of Figures

List of Figures (cont.)

List of Figures (cont.)

List of Tables

The fact that one cannot order speech sounds, vowels and consonants, along a unidimensional physical scale points to the presence of a complex physical representation. Although the exact nature and number of acoustic and articulatory features needed for adequate description are not agreed upon, it seems reasonable to suppose that speech sounds can be considered as occupying positions in a multidimensional space in which each of the features used in describing the speech sounds corresponds to a dimension of the space.

Consequently, there has been a concerted effort to investigate the problem of relating differences in the physical properties of speech sounds to their perceptual representation. In absolute identification experiments, subjects are presented with isolated phonemes and asked to indicate after each random presentation which phoneme they think has been presented. The observations gathered are arranged in an n x n confusion matrix indicating for each of the n phonemes presented how often it was confused with one of the n-1 other phonemes. The confusion data thus obtained can help us discover how a set of speech sounds is processed within the listener or, in other words, to discover the psychological structure underlying a particular set of speech sounds. It should be borne in mind that the construction of a confusion matrix does not require any knowledge of the physical parameters or properties of the stimuli for it is based on the responses of the listeners. In substance, a structure or a representation derived from a confusion matrix is a subjective interpretation of the interrelations among the stimuli as they are perceived psychologically and not as they are measured physically (Shepard, 1972).

In a classical study of confusions among single initial consonants, Miller and Nicely (1955) measured perceptual confusions among single initial consonants under various conditions of noise added to the speech signal. In their experiments, subjects were asked to identify each of sixteen different consonant-vowel (CV) syllables. All syllables had the same ending /a/, as in father. The consonants used were the following: /p, t, k, f, θ, s, ʃ, b, d, g, v, ð, z, ʒ, m, n /. Miller and Nicely (1955) obtained 17 complete 16 x 16 confusion matrices, each under a different condition of signal to noise ratio. Shepard (1972) was able to derive a similarity matrix from their data by pooling their results. From this, he was able to construct a spatial representation of the speech sounds. Combining data in this fashion, Shepard argues, has little effect, if any, on the internal structure of the confusions. Shepard recovered the two dimensional spatial configuration in Fig. 1 by applying the method of "exponential analysis of proximities", fol-

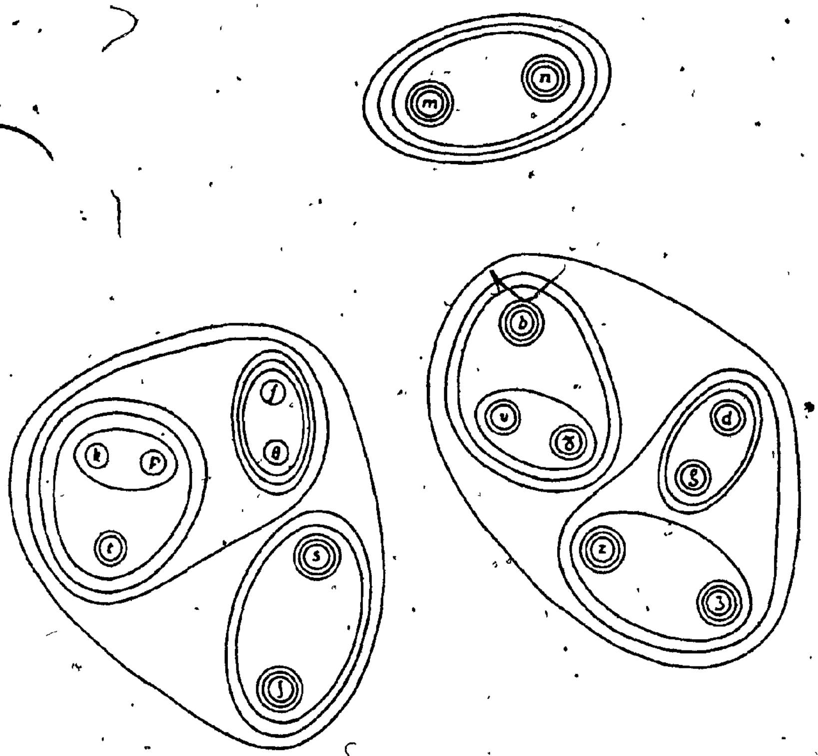

Figure 1. Spatial and hierarchical representation of the perceptual similarity between consonants derived from Miller and Nicely's (1955) data. From Shepard (1972).

lowed by a hierarchical clustering technique which sequentially orders sound pairs according to their degree of similarity. His results indicate that two orthogonal dimensions corresponding to the perceptual features of voicing and a combination of affrication and nasality partitioned the 16 consonants into clusters representing the nasals / m, n/, the unvoiced stops and fricatives /p, t, k, f, θ, s, ʃ/ and the corresponding voiced stops and fricatives /b, d, g, v, ð, z, ʒ/. In addition to replicating the results of Miller and Nicely, who had already shown that the two dimensions of nasality and voicing best met the data, Shepard provided some validity for this type of spatial representation.

The two perceptual features in the spatial configuration of Fig. 1 account for 99.4% of the variance for confusion among the 16 consonants. Miller and Nicely, on the other hand, considered the speech stimuli to vary with respect to as many as five distinctive features (i.e., voicing and nasality as well as affrication, duration and place of articulation). The fact that over 99% of the variance is accounted for by two dimensions, does not imply that variations occur along two distinctive features only. In the lower part of Fig. 1, for example, we can discern the parallel and vertical ordering between the unvoiced fricatives and their voiced counterparts with respect to place of articulation. Moving down from /f/

and /v/ to /ʃ/ and /ʒ/ the respective places of articulation for the unvoiced fricatives and their voiced counterparts are labiodental, dental, alveolar and palato-alveolar. Thus, in addition to the previously mentioned distinctive feature of affrication, place of articulation appears also to be correlated with the ordering of the speech sounds along the major dimensions in this spatial representation. Surely, a higher dimensional configuration would account for a larger proportion of the variance and some investigators, using multidimensional scaling methods, have endeavoured to extract as many as four spatial dimensions from the Miller and Nicely data (Wilson, 1963). But with less than 1% of the variance unaccounted for in the example described here, it is doubtful whether the added dimension(s) would have any psychological significance. Nevertheless, it should already be clear that a spatial representation is useful in that it transforms the implicit information contained in the original matrix of numbers into an explicit and immediately accessible structure (Shepard, 1972).

The confusion between phonemes has been considered here as a measure of psychological similarity, but direct subjective measurements of the similarities of pairs of stimuli have also been obtained. In an experiment using the same 16 consonant stimuli employed by Miller and Nicely (1955), Peters (1963) had subjects rate each of

120 pairs of consonants on a nine-point scale (ranging from "extreme similarity" (1) to "extreme dissimilarity" (9)) after having pronounced the pair aloud. It is important to note that subjects in this study responded to the stimuli not only as listeners but also as speakers.

Shepard applied the clustering algorithm used previously to the resulting matrix of similarity ratings obtained by Peters. The most prominent clusters are represented in Fig. 2. A single heavy line links pairs with high similarity ratings. For purposes of comparison, the consonant phonemes are placed in the same relative positions as in Fig. 1. The results of the clustering analyses are indicated by the dashed lines encircling the clusters. At the level of four clusters, the consonants are divided into four groups comprising the nasals /m, n/, the stops /b, d, g, p, t, k/, the fricatives /f, θ, v, ð/, and the sibilants /s, ʃ, z, ʒ/. At the level of eight clusters, every voiced consonant is linked with its unvoiced counterpart, and the two nasals (as in Fig. 1) are still paired together.

The greatly outstretched representations of these clusters indicate that the pattern obtained here bears little resemblance to those derived from the confusion data. The discrepancy between the structures recovered from the similarity and the confusion data may be due, as was previously remarked, to the difference between

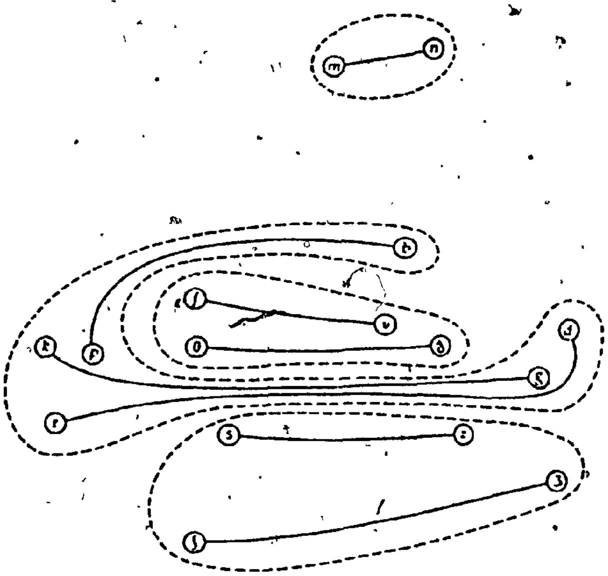

Figure 2. Spatial and hierarchical representation of the perceptual similarity between consonants derived from Peters' (1963) data. From Shepard (1972).

responding solely as a listener, as in Miller and Nicely's experiment, or as both a listener and a speaker, as in Peters'. A speaker, for example, might be paying more attention to the articulatory characteristics of the sounds he is pronouncing than would a listener.

Shepard (1972), however, argues for a more cognitive interpretation of these results, and disagrees with Peters' conclusion that the consonants were sorted according to manner of articulation first, and that voicing and place of articulation were respectively next in importance. The feature of voicing, in fact, appears to have been ignored by the subjects as is evident from the groupings in Fig. 2. This contrasts sharply with the confusion data solution where this distinctive feature was clearly the most prominent. Shepard, therefore, suggests that Peters' subjects treated this judgmental task as an analogy task rather than as a similarity task. In other words, subjects might have been most impressed by the parallelism between the voiced and unvoiced consonants, thus leading them to discount the manifest difference between the two parallel sets. Furthermore, Peters' subjects had had some practice in phonetics and the fact that they pronounced the consonant themselves might tend to focus their attention on the feature of manner of articulation. The disparity between the two representations might be due to the fact that they are based on the results of two differently derived sets of data, confusion and similarity data. Although it is difficult at this point to ascertain which method has greater validity, this disparity argues for caution in generalizing from the confusions among phonemes to subjective estimates of similarity.

Studies on the perceptual representation of speech sounds have not been confined to consonants only. In a well known study by Peterson and Barney (1952) listeners attempted to identify, after an aural presentation, which of ten monosyllabic words each beginning with /h/ and ending with /d/ and differing only in the vowel had been pronounced. The words used were heed, hid, head, had, hod, hawed, hood, who'd, hud and heard and the vowels were respectively /i, I, ɛ, æ, ɑ, ɔ, U, u, ʌ, ɝ/. The task was rendered difficult by the fact that different speakers varying both in sex and age pronounced the same words on different presentations.

Shepard (1972) computed a 10 by 10 symmetric matrix of proximity measures from the resulting confusion matrix obtained by Peterson and Barney and subjected it to an exponential analysis of proximities. The recovered perceptual representations were fitted in the dimensional space shown on Fig. 3. The resulting configuration has some face validity. Words that sound similar (e.g., hod and hawed) are represented close to each other while words

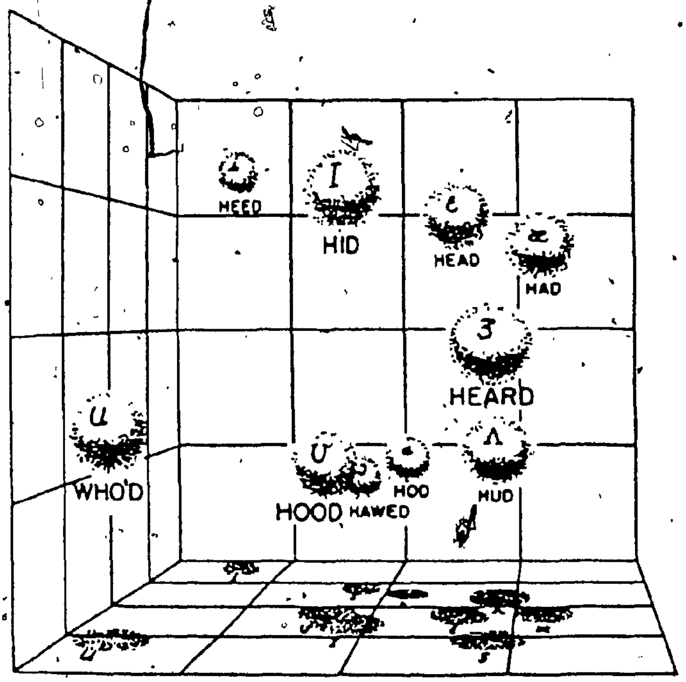

Figure 3. Three-dimensional spatial representation for 10 vowels derived from Peterson and Barney's (1952) data. From Shepard (1972).

that sound relatively dissimilar (e.g., who'd and heed) are further apart.

Additional demonstration for the validity of such a spatial configuration can be obtained from its relationship with other, external variables. In this connection, Shepard warns, there is no ground to believe that the projections of the points on any of the reference dimensions are related to an external variable since the structure of the recovered configuration is uniquely determined by the data. With respect to vowels, however, it is only the frequencies of the two or three lower formants (Fi) - groups of overtones, or energy peaks in the frequency spectrum, that correspond to resonating frequencies of the air in the vocal tract - that have been considered to be of any psychological significance (Potter and Peterson, 1948; Darwin, 1976). Though the three dimensions in Fig. 3 account for most of the variance (99%), the first two dimensions "explain" 97% of it and correspond roughly to the first and second formant frequencies of the vowels.

Shepard fitted three new rotated axes through the representation of Fig. 3 such that each axis would best fit the average center frequency of the vowel formants as measured by Peterson and Barney. He found that the angles between the three new axes corresponding to the first three formants were 99° between F1 and F2, 104°

between F1 and F3, and 49° between F2 and F3. That the psychological effect or contribution of F1 is relatively independent of the other formants is suggested by the near orthogonality of the first axis to the other two. Psychological effects of F2 and F3, on the other hand, appear to be somewhat interdependent. This may be attributed to the restricted range in which F2 and F3 can vary: the lower F2, the more determined is F3. Fig. 4 shows the F1-F2 and F2-F3 planes of the acoustic space.

Because, in this case, a reduced spatial configuration of two dimensions still gives a good approximation of the data, accounting for 97% of the variance, and because the 10 points in the space are roughly aligned along the traditional vowel loop - the closed curve which delimits the area in the F1-F2 plane (Fig. 4) of the acoustical space where all vowels are contained -, we can discern a clear relationship between the physical properties of the stimuli and their recovered psychological representations derived independently of those physical measurements. The notion that the frequencies of the first two formants play a critical role in the recognition of spoken vowels thus receives additional support.

It is important to note that Peterson and Barney (1952) tried to relate vowel qualities with formant patterns. They plotted in the F1-F2 plane 10 vowels spoken by different speakers. The graph showed some overlap

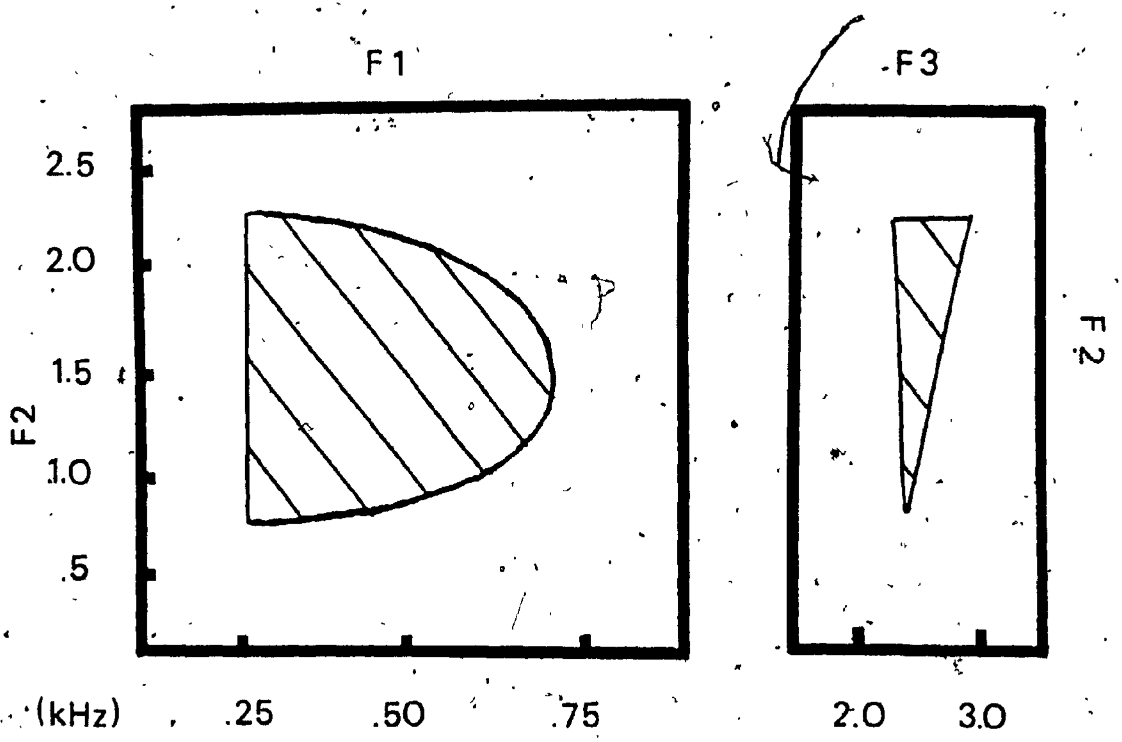

Figure 4. Stylized version of acoustic space in three dimensions.

for certain vowel areas. The addition of the F3 frequency did not greatly alter this overlapping. Applying statistical techniques of multivariate analysis to the Peterson and Barney data, Welch and Wimpress (1961) showed that adding the F3 information to F1-F2 did little to enhance recognition. It simply reduced the error rate from 13% to 9%. With the fundamental frequency (F0) also taken into account, the error rate was further reduced to approximately 6%.

Individual differences are partly responsible for the overlap in the formant space. As is shown by the Peterson and Barney (1952) spectrographic measurements, the center frequencies of the vowel formants vary considerably across men, women, and children as well as between speakers of the same sex and age group. No spread, however, is observed when the same person repeats the vowels a number of times (Potter and Steinberg, 1950). In this connection, a speaker-dependent correction would be of great importance. Some experiments suggest that a "reference space" taking into account the incoming "speaker-dependent" information is built up in the course of speech perception (Ladefoged and Broadbent, 1957; Lieberman, Crelin and Klatt, 1972). Suggestive evidence for this reference space comes from Gerstman (1968) who, using also Peterson and Barney's data, introduced a speaker normalization based on the vowels /i, a, u/. He found

97.5% of the vowels were recognizable in the resulting two dimensional configuration and as had been suggested before (Potter and Steinberg, 1950), it turned out that a third dimension was necessary to account for /ɝ/. Results from more recent experiments suggest, however, that listeners can identify a high proportion of vowels spoken differently and in different contexts and that speaker-dependent acoustic variation poses no great perceptual problems within a common dialect group (Verbrugge, Strange, Shankweiler and Edman, 1976).

Plomp, Pols and Van de Geer (1967) presented another alternative to study the nature of vowel spectra. They carried out a dimensional analysis on the frequency spectra of 15 Dutch vowels spoken by 10 persons and determined this with 18 bandpass filters. The analysis showed that four dimensions accounted for 84.1% of the total variance. They identified the first two dimensions only as the frequencies of the first and second formants. Correct identification of vowels reached 90% when the shortest distance between the mean vowel positions (averaged over the 10 speakers) and the individual vowel points in the four dimensional configuration was used as a criterion. The percentage of correct identifications dropped to 85% when the spatial configuration was reduced to three dimensions.

The perceptual representation of vowel sounds may, then, to a certain extent be related to the physical space since some of the physical dimensions must affect the way in which listeners discriminate between speech sounds (Wilson and Saporta, 1965; but see Shepard, 1972). Cohen, Slis and Hart (1967) found that the three dimensions corresponding to center frequencies of F1, F2, and duration, apart from specific formant bandwidths, had to be combined in an optimal way for maximal recognition of synthetic vowels. Pickett (1957) investigated perceptual confusions among 12 vowels masked with noise and concluded also that the same three physical characteristics were the most important ones. Mohr and Wang (1968) computed similarity matrices for vowels following a paired comparison procedure and paired physiological features (high, mid, labial, palatal, nasal) with the rank order of the similarity indices. Some of these physiological features significantly affected similarity scores.

Finally, in a study by Pols, van der Kamp and Plomp (1968), 15 subjects judged 165 triads of Dutch vowels by indicating which pair was most similar and which pair was most dissimilar. They derived a three dimensional spatial representation which correlated quite well (.992, .971 and .742) with the physical configuration. They concluded "that subjects used for their perceptual judgments information comparable with that present in the physical representation of these signals" (p. 466). Their results also support the idea that F1 and F2 are the most important factors in vowel perception. They found the third dimension hard to describe.

It appears that the determination of the perceptual representation of a particular set of vowel sounds is, in part at least, related to the associated physical (phonetic) space. We can therefore expect the configuration recovered in the perceptual space to be dependent on the configuration expressed in the physical space. The psychological space must be ~~related~~ directly to at least some subspace of the physical space, since some dimensions of the physical representation correspond to the stimuli to which the subjects respond. The difference between the psychological and physical representations would represent a transformation in the ordering of speech sounds produced by the perceptual frames of a set of speakers/listeners of a given language.

METHODOLOGICAL APPROACHES TO THE REPRESENTATION OF SPEECH SOUNDS

There have been a number of methodological approaches to the study of underlying perceptual representations of speech sounds and of vowels in particular with the purpose of defining a psychological stimulus space derived from observations concerning the similarity between stimuli. Some of the methods employed in psycho-acoustic experiments are reviewed and evaluated.

One of the methods used is that based on perceptual confusion (Miller and Nicely, 1955; Peterson and Barney, 1952). Perceptual confusions will seldom occur with undistorted speech and noise must therefore be introduced in the acoustic signal to prevent a large number of empty cells in the resulting confusion matrix. Difficulties inherent to this particular method are related to response bias and asymmetry in such matrices. In a regular matrix, for example, the distance between a stimulus and itself is always zero. This appears somewhat unreasonable since, as Wilson (1963) pointed out, the random white noise presented with the consonant sounds has a masking effect which could be regarded as creating a random distribution of effects such that the mean distance between repeated consonants with differing S/N would be non-zero. Moreover, the confusion data themselves rarely point to the features or properties of speech sounds that are important.

Short-term recall (Wickelgren, 1965) is a procedure where subjects are asked to repeat a number of presented speech stimuli, followed by a recall of the whole series. An intrusion error matrix is computed on the basis of the errors made and gives some information about encoding mechanisms that comprise, among others, memory functions. A drawback of this method, apart from the fact that subjects pronounce aloud the stimulus sounds, is that it is not suitable for stimuli that cannot be easily named and the use of subjects trained in phonetics would therefore be required.

In the triadic comparison technique, listeners have to select which two of three stimuli they think are most similar, and which pair is least similar, for all possible subsets of three stimuli in a series (Mohr and Wang, 1968; Pols et al., 1969). The scores cumulated for all triads result in a similarity matrix with a number in every cell representing the number of times a particular pair has been judged more similar than others. This procedure is an improvement over the others previously mentioned and considers the rank order of the similarity judgments as most important.

In another scaling approach, subjects are required to judge the extent to which the two speech stimuli in a pair sound similar or dissimilar relative to each other, by marking a line representing, for example, a nine-point scale (Peters, 1963) or a continuous scale. Ramsay (1978) has argued for the use of a similarity-dissimilarity continuum. Firstly, multidimensional scaling is based on a relationship that is assumed to hold between the mathematical concept of distance and the psychological concept of dissimilarity. Secondly, dissimilarity has an inverse relationship to the concept of similarity. Thirdly, dissimilarity is a property possessed by a pair of objects and not by either one taken by itself. And lastly, a similarity-dissimilarity continuous scale offers the listener the possibility to express finer subjective gradations. Another convenient aspect of such a scale is that subjects consider it to be rather simple; furthermore they are freed from the constraint of making judgments in relation to specific classes or categories (for a more comprehensive discussion of this problem, cf. Ramsay, 1978, pp. 1 - 11).

Introduction to the Experiment

As can be seen, there has apparently been no attempt to systematically compare the perceptual representation of speech sounds in different contexts. Some studies have indeed looked at the perceptual patterns underlying the presentation of vowels embedded in /h-d/ (Peterson and Barney, 1952), in /h-t/ (Pols et al., 1969), in destressed /b-b/ (Verbrugge, et al., 1976), and in other phonetic environments and, as a result, interesting features of the psychological representation of vowels have emerged (e.g., it appears that at least three dimensions are necessary to adequately account for the perceptual stimulus space of vowel sounds). It is, however, difficult to compare and generalize from their results since they all used different methodological and analytical approaches. It would certainly be of great interest to find out whether the perceptual (psychological) spaces derived when vowels are presented in isolation share some common properties with those spaces derived when vowels are embedded in different phonetic contexts, and what these properties or features are.

The experiment described in this thesis attempted to determine

1) whether traditional vowel classification systems could be used to explain the perceptual (psychological)

dimensions defining the spaces derived from the subjects' similarity judgments when presented with pairs of synthetic speech sounds;

2) whether the perceptual representations are invariant across different phonetic environments or dependent upon them;

3) the nature and the extent of the differences, if any, between the perceptual and physical configurations, and

4) between individual differences.

The Material & Stimuli

The use of synthetic speech stimuli would be recommended at first in order to present the listeners with an invariant sequence of vowel sounds, the parameters of which can be controlled, from one context to another.

In this connection, work has recently been reported that has shown the practicality of routine text-to-speech translation (Elovitz, Johnson, McHugh and Shore, 1976). The investigators developed a set of letter-to-sound rules, to translate English text into the international phonetic alphabet (IPA), that will produce correct[1] pronunciations for about 90% of the words, or approximately

---

[1] The criterion for correctness used by Elovitz and her colleagues (1976) as well as the rules for translation from IPA to Votrax representation are explicitly described in their paper.

97% of the phonemes in an average sample of English text. Of special interest is their development of a separate set of rules to translate from IPA into a phonetic encoding compatible with a particular commercial speech synthesizer (Federal Screw Works Votrax VS-6.0) available at the Concordia University's Computer Science department.

Three sets of speech stimuli were generated using the Votrax speech synthesizer. One consisted in isolated vowels, and two consisted of the same set of vowels embedded in /h-d/ and /p-p/ phonetic environments. The isolated speech stimuli were 12 vowels and three diphthongs: /i, I, e, ɛ, æ, a, ɔ, o, U; u, ʌ, ə, ai, aU, ɔI/ such as the ones respectively occurring in the following words: beet, bit, gate, get, fat, father, lawn, lone, full, fool, about, but, hide, how, toy. The diphthongs were added to make more perceptual alternatives available to the listeners and to obtain a more comprehensive representation of vowels, and to shed some light on the perceptual representation of diphthongs.

The Dependent Measure

A 100mm continuous line was used as a scale ranging from very similar (1) to very dissimilar (100). The listener indicated the extent to which a pair of presented sounds were dissimilar by marking off the line.

The distance between the origin of the line and the mark made by the subject was used as a dissimilarity estimate. A large distance indicated that a pair of speech stimuli were judged to sound relatively dissimilar to each other and share little similarity. A small distance indicated that a pair of speech stimuli were judged to sound relatively similar to each other and exhibit little dissimilarity.

Technique of Analysis: Multidimensional Scaling

In a multidimensional analysis, the dimensionality of the space in which the recovered points are constrained to lie must be specified in advance. This dimensionality may vary from one set of data to another. A convenient feature of the MULTISCALE program and of maximum likelihood estimation is that they allow the investigator to choose the dimensionality that best represents the data on the basis of statistical hypotheses tests. Another advantage of this multidimensional scaling package refers to the specification of a more complex type of provision for individual variation where it is assumed that subjects share a common perspective on the stimulus objects, or "group" perceptual structure, but that they vary with respect to the proportions in which they combine the dimensions of this spatial representation. The distance between points in model M3 that provides for interindividual variation is defined by the modified Euclidean distance formula

$$d_{ijr} \approx d^{*}_{ijr} = v_r \left[\sum_{m=1}^{k} w_{rm} \left(x_{im} - x_{jm} \right)^{2} \right]^{p_r/2}$$

where d_{ijr} = particular dissimilarity measure, and d^{*}_{ijr} = corresponding approximating distance.

The dimensionality of the spatial configuration is indicated by k and a particular dimension is referred to as m. The location of a particular point in this space is x_{im} - the coordinate of the i^{th} point on the m^{th} dimension.

The multiplier v_r, with r=1,...., N, is a regression coefficient and allows the possibility that the data from any two replications or subjects may differ from each other by a scale factor.

The coefficient w_{rm} is a dimension weight. Each dimension for the r^{th} subject is multiplied by this coefficient which defines the "strength" or the relative salience of the m^{th} dimension in contributing to the distance which approximates this subject's data.

The exponent p_r indicates the degree to which a subject uses the upper extremes of the response scale. When p_r is very low (e.g., p_r 0.3) it may indicate a subject whose responses are inconsistent with those used by the other subjects or with the MDS model being used Ramsay (1978).

Contextual Effects

In order to prevent the occurrence of contextual effects, the 15 stimuli used during an experimental session were presented all together, in a string, once at the beginning of the session and a second time in the middle.

In a scaling experiment dealing with vowels, stop consonants and tones, Vinegrad (1970) required his subjects to listen to two sequences of stimuli and then to make a judgment. In the case of vowels (they formed a smooth series of 13 vowel sounds running through the American English vowels /i/, /I/, and /ɛ/) the first sequence invariably contained the same seven stimuli, viz., 1-3-5-7-9-11-13, and the second sequence always contained three stimuli, viz., 1-X-13, where X represents any of the 13 vowel sounds. After an aural presentation, the subject was to indicate how similar or close X seemed to 1 or to 13 by marking a point on a seven inch line. This procedure was used to remove or reduce any contextual effects that Vinegrad observed in preliminary experiments. In practice, subjects tended to make judgments in relation to each other with the consequence that one estimate was partly influenced by the preceding one.

In the present study, the introduction of the string of 15 stimuli was intended to provide subjects with a reference space and to avoid the contextual effects observed by Vinegrad. Due to the differences in the nature and in the task between the present experiment and Vinegrad's, the string of stimuli was only presented twice in each session instead of before every experimental trial as in Vinegrad's.

Method

Subjects

The subjects were 12 native speakers of English, six males and six females, with no reported hearing deficit. They were aged 23-31 years with a median age of 26. With the exception of one male subject, all were students enrolled in the Psychology graduate program at Concordia University. Only one subject had previously participated in an experiment using synthetic speech. Subjects were paid $5.00 to participate in the study.

Material and Apparatus

The stimuli used in this study consisted of three sets of speech sounds: vowels presented in isolation, vowels embedded in /h-d/, and vowels embedded in /p-p/.

Isolated vowels: the stimuli were 12 English vowels and three diphthongs drawn from the IPA. They were generated on a Votrax VS-6.0 voice generator, produced by Federal Screw Works. The voice generator synthesized human speech with unlimited vocabulary and was similar to that used by Elovitz and her colleagues (1976). A computer program was designed (using the Editor program OTXED available at Concordia University's computer center) to translate Elovitz et al.'s representation of IPA vowels to Votrax representation using their rules for translation (pp. 453-458). The 15 IPA stimuli as well as their Votrax and "Elovitz" Latin letter representations are listed in Table 1. The isolated vowel sounds produced varied in duration and were recorded on a Sony TC-2520 tape deck with input control only. Speech rate, pitch and audio levels on the Votrax unit were adjusted manually and remained set at the same levels throughout the recording session. The levels were set at 2 for speech rate, 3 for pitch, and 2.5 for audio levels.

Vowels in /h-d/ (HVD) and in /p-p/ (PVP) contexts: computer generated /h-d/ and /p-p/ phonetic environments were produced following Elovitz, et al.'s rules of translation. The 15 isolated IPA vowels were embedded in these two phonetic contexts by means of another computer program (using the same Editor program). The stimuli thus produced were recorded in the same fashion and with the same equipment as the vowels in isolation. A Sony TC-270 was used to present all stimuli.

Design

The subjects participated in all three conditions: vowels presented in isolation (VPI), vowels embedded in /p-p/ (PVP), and vowels embedded in /h-d/ (HVD). With six possible orders of presentation

Table 1

Latin Letter Representation of IPA Stimuli Used in the Study

I.P.A.	ELOVITZ et.al.	VOTRAX	Ex.
i	IY	E	beet
ɪ	IH	I	bit
e	EY	A AY	gate
ɛ	EH	EH	get
ae	AE	AE	fat
a	AA	AH	father
ɔ	AO	AW	lawn
o	OW	O1 U1	lone
U	UH	OO	full
u	UW	IU U	fool
ə	AX	UH2	about
ʌ	AH	UH	but
ai	AY	AH E1	hide
aU	AW	AH O1	how
ɔI	OY	O1 E1	toy

VPI - HVD - PVP, HVD - VPI - PVP
PVP - HVD - VPI, VPI - PVP - HVD
HVD - PVP - VPI, PVP - VPI - HVD

two subjects, one male and one female, were tested under each order. In all, each subject contributed to 225 observations per condition.

Procedure

Subjects were individually tested over three sessions of the experiment each held on different days and lasting approximately one half-hour. The subject was seated in a partially sound-attenuated room. The stimuli were presented through a Sony loudspeaker located 80cm. in front of the subject. The participants were first told that the taped sounds they were about to hear had been produced by a speech synthesizer controlled by a digital computer. They then received the following instructions (adapted from Ainsworth and Millar, 1973; and from Christian, 1976):

> I am now going to present you with a series of speech stimuli taken two at a time. For each of these pairs your task is to decide how similar or how different these two stimuli are. Indicate your decision using the scale in front of you by marking off the line bounded on the left by the phrase "very similar" and on the right by "very dissimilar". For instance, if

> you decided that those two stimuli were very much alike you would mark off the line near the very similar end. If, on the other hand, you felt the two stimuli had nothing in common then you would mark off the line near the very dissimilar extremity. The scale is provided to allow you to indicate the degree of similarity you feel exists between two stimuli, so please try to use the full range of the scale in making your judgments. Do not hesitate more than a few seconds before making a decision. It is your immediate response which, for whatever reason, reflects the degree of similarity between each pair of sounds that we wish to record. Each pair of speech sounds will be preceded by the sound /ʃ/ which will act as a warning that the stimuli are about to be presented. Please give your full attention on all trials and remember, there are no right or wrong answers.

The experimenter then introduced the sequence of taped events in the following order:

All 15 stimuli used in a condition were first presented in a string with an interstimulus interval (ISI) of 730msec. There was then a 12 sec pause followed by six practice trials. Subsequently, a pause ensued in which all points of confusion, if any, were clarified followed by the experimental trials. The practice and experimental trials began with the sound /ʃ/ warning the subject 500msec in advance that the stimulus pair was about to be presented.

The sound stimuli were presented with an interval of 700 msec between each sound. An ISI of 4 sec separated the pair of sounds from the next occurrence of /ʃ/.

The subjects were required to indicate on a 100mm. line (ranging from very similar (1) to very dissimilar (100)) the degree to which they felt each two sound stimuli to be similar or dissimilar. They were provided with sheets with fifteen 100mm lines horizontally spaced out. There were no labels or marks to guide the subjects in their judgments. They were told to use one line per judgment and to turn over to new sheets as necessary. Apart from the presentation of the 15 stimuli, both at the beginning and in the middle of each session, no description of the stimuli was given and the subjects were left to form their own frame of reference. A three minute pause, halfway through the testing session separated the randomized presentation of the stimuli in the upper half matrix (above the diagonal) from those in the lower half (below the diagonal). Before the session was resumed, however, the string of 15 stimuli was presented again in the same order as at the beginning of the session. The order of presentation of the speech stimuli remained the same for all subjects and from one condition to the next.

Method of Analysis

The 225 dissimilarity judgments of pairs of sounds were analyzed for all subjects for each condition by multidimensional scaling techniques using model M3 of the MULTISCALE program (Ramsay, 1978).

To test whether the fit in k dimensions is significantly better than in k-1 dimensions, Ramsay (1977, 1978) suggests to use the quantity

$$2 \left(\log L_k - \log L_{k-1} \right) \tag{1}$$

which has an approximate chi-square distribution with $N + n - 2$ degrees of freedom for model M3; where N = number of subjects, n = number of stimulus objects, and $\log L_k$ is the log likelihood obtained by fitting k dimensions. Ramsay (1977) has shown that a solution containing too many dimensions occurred in 7 out of 28 cases when the quantity in (1) was used as a chi square variate. Simply doubling the chi square criterion led to the retention of too many dimensions in only one case out of 28 (Ramsay, 1977; Christian, 1976).

Results and Discussion

The data were analyzed in an effort to determine whether the dimensions underlying the group's perceptual representation remain the same across the three phonetic contexts and can be accounted in terms of a traditional classification system of vowels. Differences, if any, between the physical and psychological representations are evaluated next. Finally, the problem of individual differences in the representation of the underlying structure of these speech stimuli is examined.

Psychological Representation of Vowel Sounds

Each subject contributed 225 observations per condition. The data for all subjects were input at once, a condition at a time, as an unbroken set of numbers (vector input). The diagonal entries in the dissimilarity matrix were not input as part of the vector, thus only 210 observations per subject were analyzed. Preliminary inspection of the raw data (dissimilarity estimates) revealed that all subjects assigned to all 15 pairs of identical stimuli dissimilarity ratings of zero or near zero values. All subjects in the study produced acceptable data as can be seen in Table 2: within-subject statistics provide no unacceptably high standard error estimates (1.3) or unacceptably low exponents (.3)

Table 2

Within Subjects Statistics
for Final Configuration

Subj.	Exponents			Unbiased Standard Error		
	Vowels in Isolation	p-p Context	h-d Context	Vowels in Isolation	p-p Context	h-d Context
1	.96	1.11	.68	.74	.84	.69
2	.85	1.07	.97	.72	.93	.95
3	1.00	.50	1.40	1.06	.90	1.06
4	1.10	.74	1.29	.99	1.06	.96
5	1.42	1.17	.84	.97	1.00	.97
6	.71	.70	.48	.76	.84	.97
7	.85	1.17	.67	.78	.83	.64
8	1.08	1.40	1.32	.80	.89	1.01
9	2.07	1.43	2.12	.98	.94	1.00
10	1.56	1.47	1.68	.93	1.01	.94
11	1.80	.93	1.86	1.08	1.02	.92
12	1.52	1.34	2.19	.86	.82	1.05

that would suggest subjects produced ratings which were at either end of the scale or exactly in the middle (Ramsay, 1978).

The multidimensional scaling analyses for the whole group of subjects yielded configurations of the vowels in a Euclidian space that reflected the underlying structure for the intervowel distances of similarity. This structure is interpreted as reflecting the properties and invariances of the vowels that permit their identification. The space can be viewed as an approximation to an auditory-perceptual structure and as providing information about the properties of this structure. The assumption is made that distances of judged similarity relate meaningfully to psychological distances of similarity. Tables 3, 4, and 5 show the dissimilarity matrices derived from the subjects' dissimilarity estimates.

Table 6 shows the significant dimensionality retained for the final configurations. The stopping rule of dimensionality adopted to establish the number of dimensions underlying the group's psychological auditory space was twice the critical chi square value $X^2 = 37.65$, with d.f. = 25 and $p = .05$ (Christian, 1976; Ramsay, 1977). Applying this stopping rule, four dimensions provide a significantly better fit of the data than three for vowels presented in isolation (VPI) ($\log L_k - \log L_{k-1} = 77.8 > 2X^2 = 75.3, p < .05$) and in PVP ($\log L_k - \log L_{k-1} =$

Table 3

Distance Matrix for the Configuration of Isolated Vowel Sounds in Four-Dimensional Space

ɪ	145													
e	29	132												
ɛ	212	179	194											
æ	217	138	423	94										
a	266	245	64	229	210									
ɔ	289	112	197	234	225	47								
o	330	299	328	281	350	294	289							
ʊ	265	236	270	237	286	193	184	121						
u	319	284	318	271	339	305	302	29	136					
ə	242	185	247	191	218	73	65	240	129	242				
ʌ	240	186	243	302	200	81	65	237	129	236	23			
ai	275	286	273	263	270	194	239	316	273	329	226	222		
aʊ	317	258	313	262	268	165	184	156	169	220	170	167	156	
ɔɪ	353	335	292	333	361	318	342	65	230	199	293	291	269	207
	i	ɪ	e	ɛ	æ	a	ɔ	o	ʊ	u	ə	ʌ	ai	aʊ

Table 4

Distance Matrix for the Configuration of Vowel Sounds
Presented in /p-p/ Context in Four-Dimensional Space

I	21													
e	2	23												
ɛ	37	35	35											
æ	35	18	40	7										
a	50	38	40	35	38									
ɔ	51	39	49	36	34	4								
o	57	54	57	56	58	38	40							
U	46	46	30	47	48	28	28	13						
u	61	57	62	55	57	36	38	8	14					
ə	48	40	47	39	38	12	8	39	26	38				
ʌ	51	30	47	38	37	3	6	38	25	35	2			
ai	50	48	49	51	49	32	34	54	47	56	35	35		
aU	55	50	54	50	49	19	21	37	30	37	22	21	22	
ɔI	51	59	51	64	67	47	50	35	37	41	48	47	38	35
	i	I	e	ɛ	æ	a	ɔ	o	U	u	ə	ʌ	ai	aU

Table 5

Distance Matrix for the Configuration of Vowel Sounds

Presented in /h-d/ Context in Five-Dimensional Space

	i	ɪ	e	ɛ	æ	a	ɔ	o	ʊ	u	ə	ʌ	ai	aʊ
ɪ	164													
e	23	160												
ɛ	214	69	213											
æ	233	98	237	47										
a	277	239	267	233	249									
ɔ	273	251	269	242	261	50								
o	285	286	277	288	294	253	248							
ʊ	264	257	269	251	249	187	174	92						
u	291	274	292	269	275	255	234	39	100					
ə	265	240	265	229	241	70	50	207	130	209				
ʌ	264	238	262	229	245	55	39	210	141	210	23			
ai	281	261	277	291	304	176	214	255	251	281	204	210		
aʊ	319	271	322	267	264	127	165	252	201	258	139	147	147	
ɔi	269	281	267	304	306	267	283	159	184	184	245	203	190	217

Table 6

Summary Results of Final Configuration

	3 dimensions	4 dimensions[1]	5 dimensions
Vowels in Isolation			
Unbiased standard error	.511	.500	.488
log Likelihood	475.367	553.164	613.468
Number of iterations	281	406	494
Vowels in /p-p/ Context			
Unbiased standard error	.504	.489	.467
log Likelihood	508.062	595.943	657.41
Number of iterations	349	209	300

	4 dimensions	5 dimensions[1]	6 dimensions
Vowels in /h-d/ Context			
Unbiased standard error	.483	.469	.453
log Likelihood	625.493	714.010	731.679
Number of iterations	413	674	123

[1] Final dimensionality retained.

$87.9 > 2X^2 = 83.1$, $p < .02$). Five dimensions were found in HVD ($\log L_k - \log L_{k-1} = 88.5 > 2X^2 = 83.1$, $p < .02$). The results thus indicate that the dimensionality reached for the final configurations depends upon the context in which the speech sounds are presented.

Physical Representation of Vowel Sounds

In order to determine the physical space of the speech stimuli, spectrograms of the vowel sounds were made using a spectrum analyzer (Vibralyzer 7030A, from Kay Electronics). The first three lower formant center frequencies (F1, F2, F3) were recorded and are presented in Table 7. In terms of a geometrical model, we can say that analysis of the sound spectrograms of the 15 speech stimuli yields a set of 15 points in a three dimensional space. Representations of the points in the F1 by F2 and F2 by F3 planes of the physical space are given in Figure 5. The Pearson product moment correlations between these variables (the formant frequencies) were determined in order to provide a more general description of the stimuli and to get an idea about the interdependency of these variables. The correlation coefficients obtained are -.32 for F1 and F2, -.02 for F1 and F3, and .80 for F2 and F3. From these coefficients it may be concluded that, for this group of speech sounds, F1 is independent of, or orthogonal to the other two formants while F2 and F3 ap-

Table 7

Frequencies (Hz) of First Three Lower Formants of Vowel Sounds Used in the Study

	F1	F2	F3
i	240	2100	2650
I	350	1900	2500
e	400	1800	2400
ɛ	500	1650	2400
æ	600	1700	2450
a	700	1150	2400
ɔ	600	900	2350
o	400	800	2400
ʊ	480	1000	2300
u	250	900	2300
ə	600	1100	2300
ʌ	620	1200	2350
ai	600-250	1100-2100	2600
aU	600-400	1200-900	2500
ɔI	450-300	1000-2000	2600

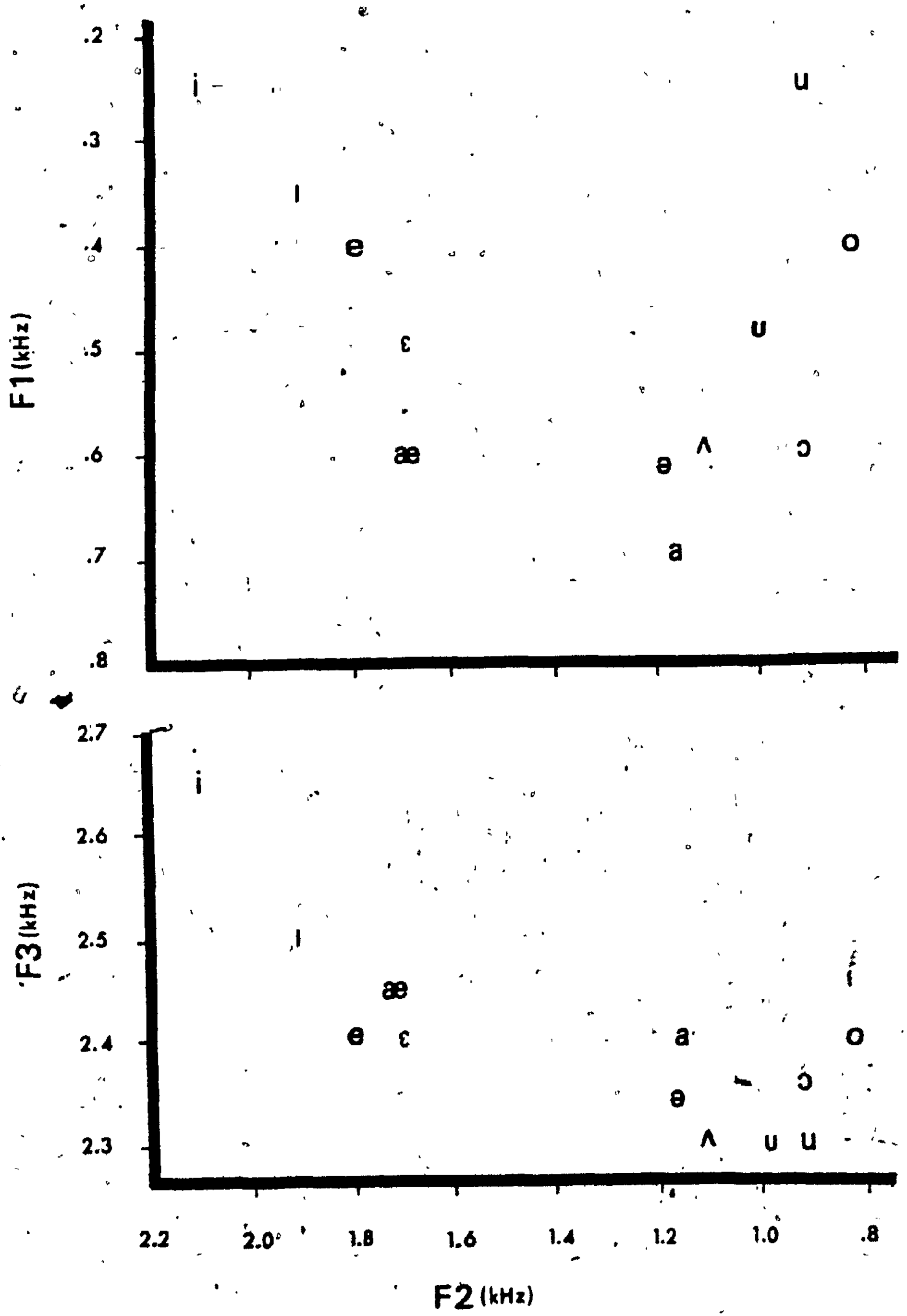

Figure 5. Physical representation of vowel sounds in the planes defined by F2 (horizontal axis) and F1 and F3.

pear to be dependent ($p < .01$). Fig. 4 shows that there is a monotonic relationship between F2 and F3: F3 increases with F2 and is determined by the shape of the vocal tract (Lindblom, 1972).

Psychological and Physical Representations Compared

The evaluation of the correspondence between the physical and psychological representations of the speech stimuli is approached next. The problem here is to determine which, and to what extent, physical aspects of the stimuli correspond to their perceptual representation. Intimately fused with this problem is the question of invariance of dimensions across contexts.

Preliminary inspection of the multidimensional scaling plots of the stimuli used in the study revealed that the points in the space defined by Dimensions I and II appeared to be ordered along the F2 and F1 continua. The projections of the twelve vowel points (excluding the three diphthong signals) on the axes corresponding to Dimensions I and II in each of the three phonetic contexts were thus correlated with formant frequencies of F2 and F1 respectively. The Pearson product moment correlation coefficients between the physical values of the stimuli (F1 and F2 frequencies) and the values derived from the subjects' similarity estimates were .69 for F2 with Dimension I and .77 for F1 with Dimension II in

PVI, .95 and .92 in PVP, and .97 and .79 in HVD. Figures 6, 7, and 8 represent the mapping of the speech stimuli in the perceptual space defined by Dimensions I and II. All these correlation coefficients are significant at the .01 level and suggest a strong correspondence between the physical and psychological dimensions. Furthermore, in each context, these two dimensions account for the greater part of the variance. Table 8 indicates the percentage of variance accounted for by each dimension within each context.

Diphthongs are quite different from vowel sounds in that they represent the unisyllabic utilization of two otherwise different vowels in the language. Therefore separate analyses were conducted with and without diphthongs to assess their contribution in the perceptual representation of speech sounds. In addition to the variance calculated when all 15 stimuli are taken into account, Table 8 also shows the variance accounted for by each dimension when only the twelve vowel signals are considered.

When all fifteen speech signals are considered, Dimensions I and II together account for 69.63% of the variance in VPI, 73% in PVP, and 68% in HVD. Of the variance produced by the twelve vowel sounds, these two dimensions now account for 77.6% in VPI, 84.28% in PVP, and 75.36% in HVD. Although the increases in percentage

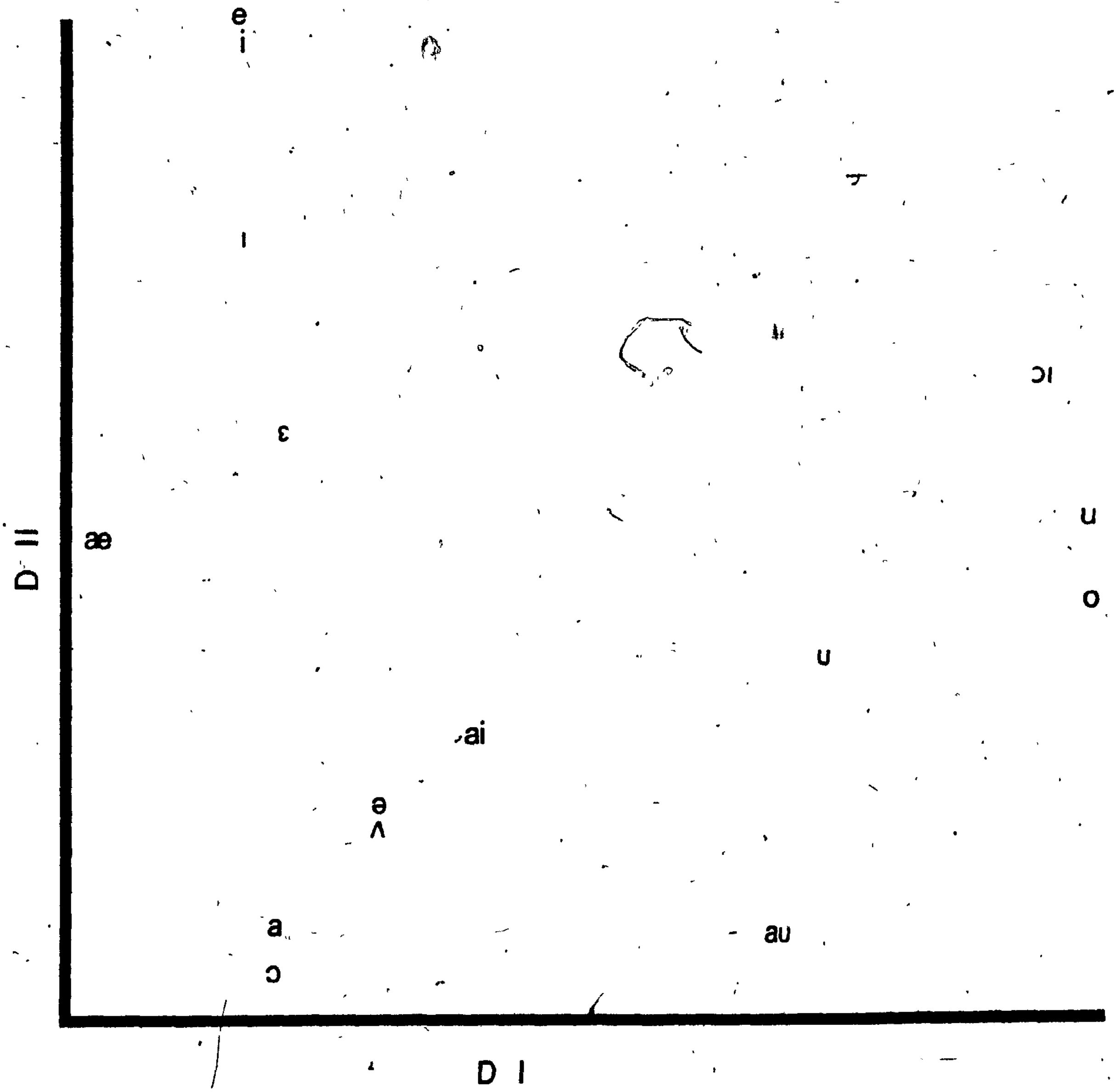

Figure 6. Plot of Dimension II (vertical axis) against Dimension I for the configuration of isolated vowel sounds.

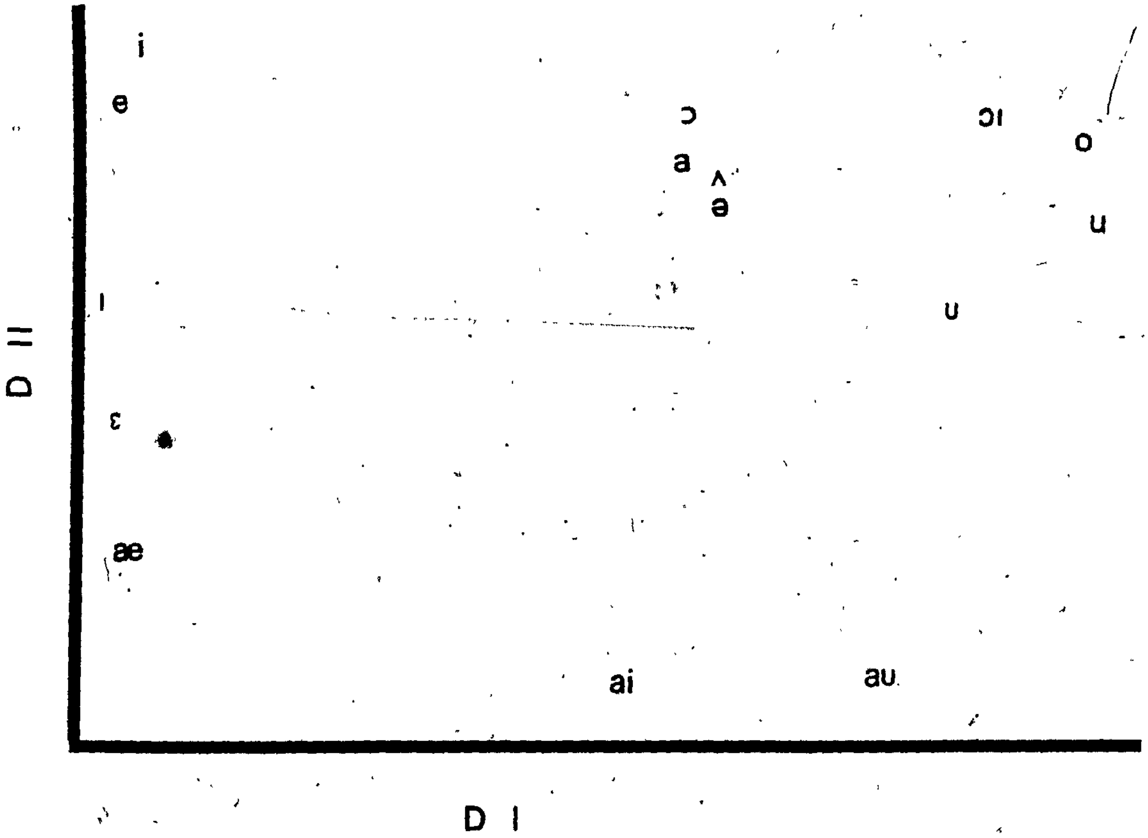

Figure 7. Plot of Dimension II (vertical axis) against Dimension I for the configuration of vowel sounds embedded in /p-p/.

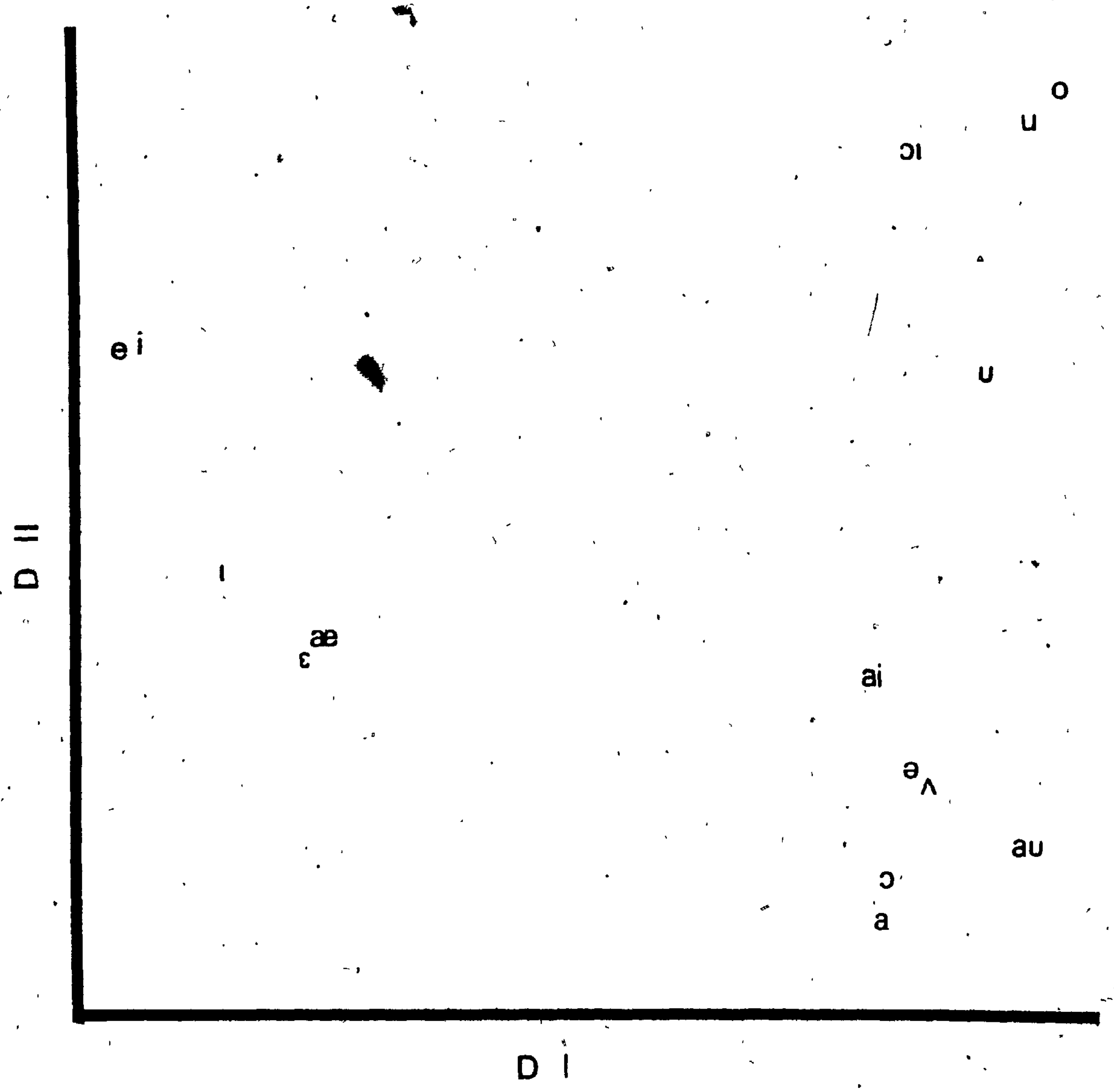

Figure 8. Plot of Dimension II (vertical axis) against Dimension I for the configuration of vowel sounds embedded in /h-d/.

Table 8

Percentage of Variance Accounted for by Each Dimension

	Isolated Vowels		Vowels in /p-p/		Vowels in /h-d/	
	Total[1]	Partial[1]	Total	Partial	Total	Partial
Dimension I	40.68	42.96	49.5	59.5	38.86	46.06
II	28.95	34.10	23.5	24.78	27.14	29.30
III	20.03	8.74	20.77	6.99	16.99	4.01
IV	10.33	14.19	6.33	8.71	13.04	17.84
V					3.96	2.78

[1] Total: vowels and diphthongs
Partial: vowels only

of variance accounted for are not large, they do reflect on the particular characteristics of diphthongs. The classification system of vowels, in terms of the articulatory features of tongue advancement (F2), tongue height (F1) and other features such as tenseness, cannot be directly applied to diphthongs because of the complexities involved in their production. When approached from the unitary point of view, i.e., from the view that diphthongs are independent phonemic units rather than a combination of two different units (vowels), they require a more complex phonetic classification because the initial vowel is different from the final one. Table 9 shows the tongue movements necessary for the realization of the three diphthongs used in the study and Figure 9 depicts the direction of these movements. The direction of the change in tongue movement is somewhat predictable. For example, in terms of tongue height, although the diphthongs start at either the mid or low tongue positions, they all terminate in the high tongue position. Subjects might be taking this particular characteristic of diphthongs into account when making their similarity judgments.

It seems, then, that in diphthongs the direction of the movement of the tongue in terms of its height and advancement is determined by rules and that an appreciable amount of tongue body movement is involved within

Table 9

Tongue Movements Necessary for the Production of the Three Diphthongs Used in the Study

Diphthong	Tongue Movement
/ai/	(a) from front to front (b) from low to high
/au/	(a) from front to back (b) from low to high
/ I/	(a) from mid to high (b) from back to front

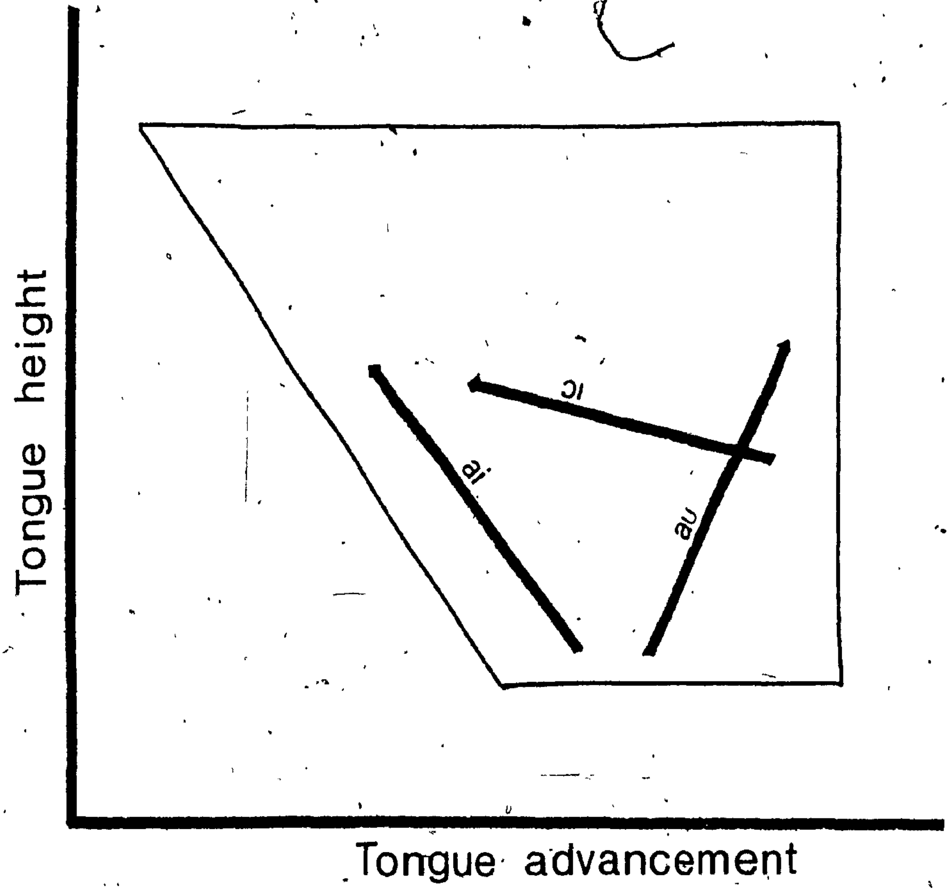

Figure 9. Tongue movements necessary for the production of diphthongs.

the perimeter of one syllable (as indicated by the changing formant frequencies on a spectrogram). For these reasons listeners can be expected to classify diphthongs in a class of their own. Indeed Dimension III appears to discriminate between the perception of diphthongs and that of the corpus of vowel sounds. Table 9 indicates the percentage of variance accounted for by Dimension III when all 15 speech signals are considered and when the three diphthongs are excluded. The percentages are, respectively, 20.03 and 8.74% in VPI, 20.77 and 6.99% in PVP, and 16.99 and 4.01% in HVD. In other words, Dimension III accounts for a larger percentage of the variance when diphthongs are included in the set of stimuli. Figs. 10, 11, and 12 represent the configurations of speech signals along Dimensions I and III.

What can one say about the placement of diphthongs along Dimensions I, II and III in all three phonetic contexts? In the case of Dimension I (cf., Figs. 6, 7 and 8), diphthongs share the same perceptual space as the back vowels. In fact, the initial vowels of the three diphthongs /ai, aU, ɔI / are also back vowels. This suggests that, at least on Dimension I, the information carried by the initial vowels is responsible for their being assigned to that particular end of that dimension. Subjects, then, perceive diphthongs as back vowels.

In Dimension II, in all three contexts, /ɔɪ/ is

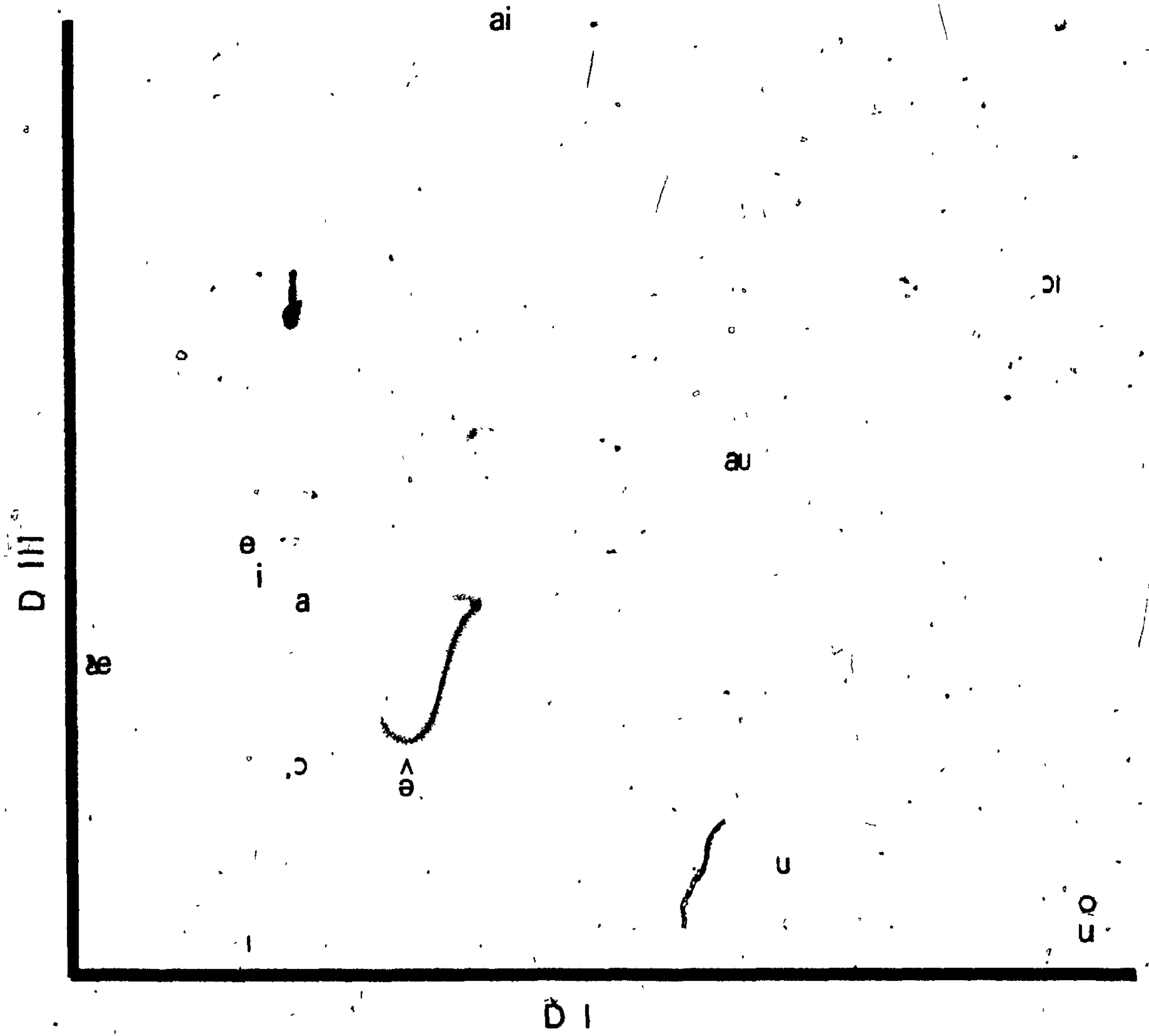

Figure 10. Plot of Dimension III (vertical axis) against Dimension I for the configuration of vowel sounds presented in isolation.

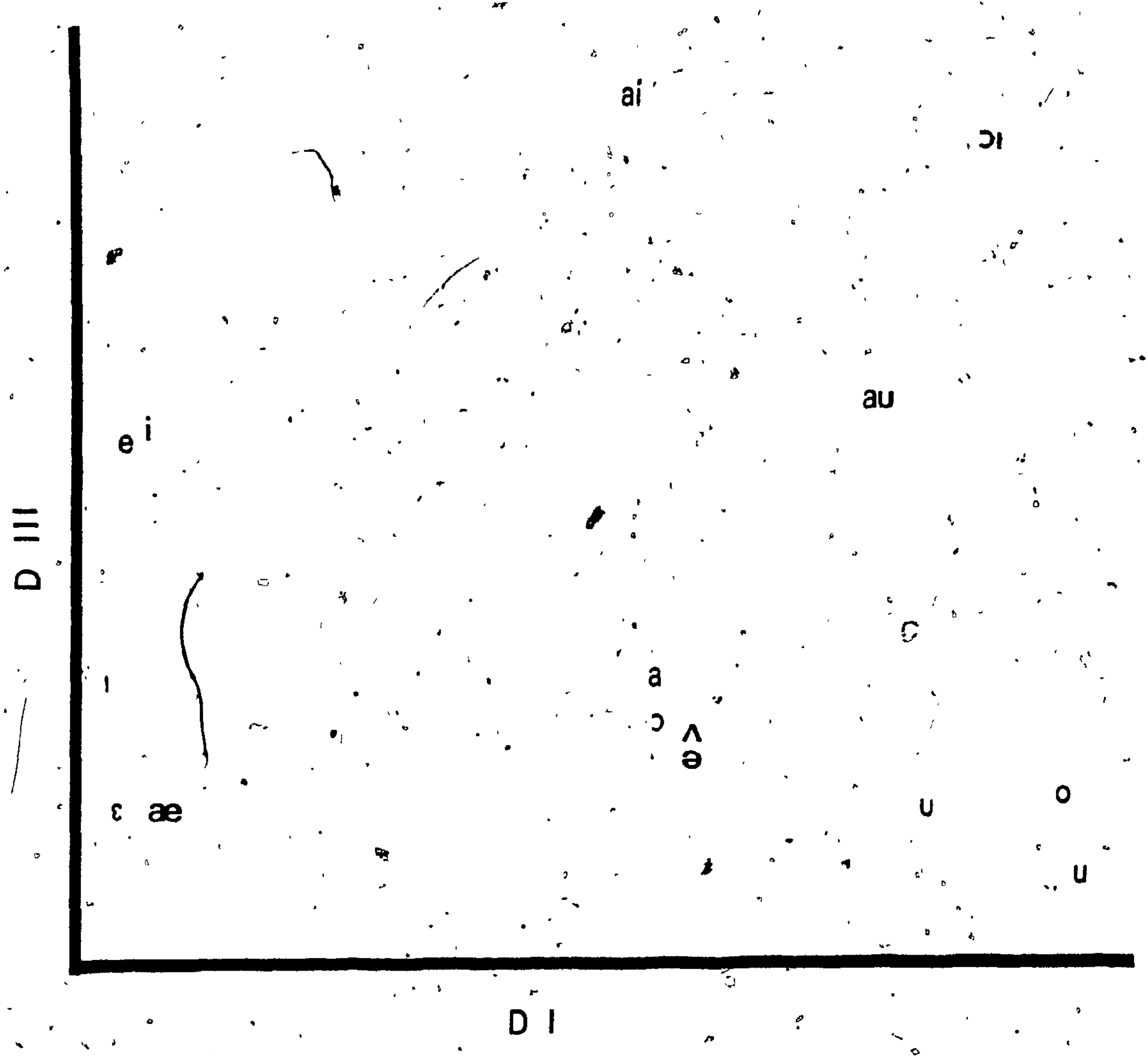

Figure 11. Plot of Dimension III (vertical axis) against Dimension I for the configuration of vowel sounds embedded in /p-p/.

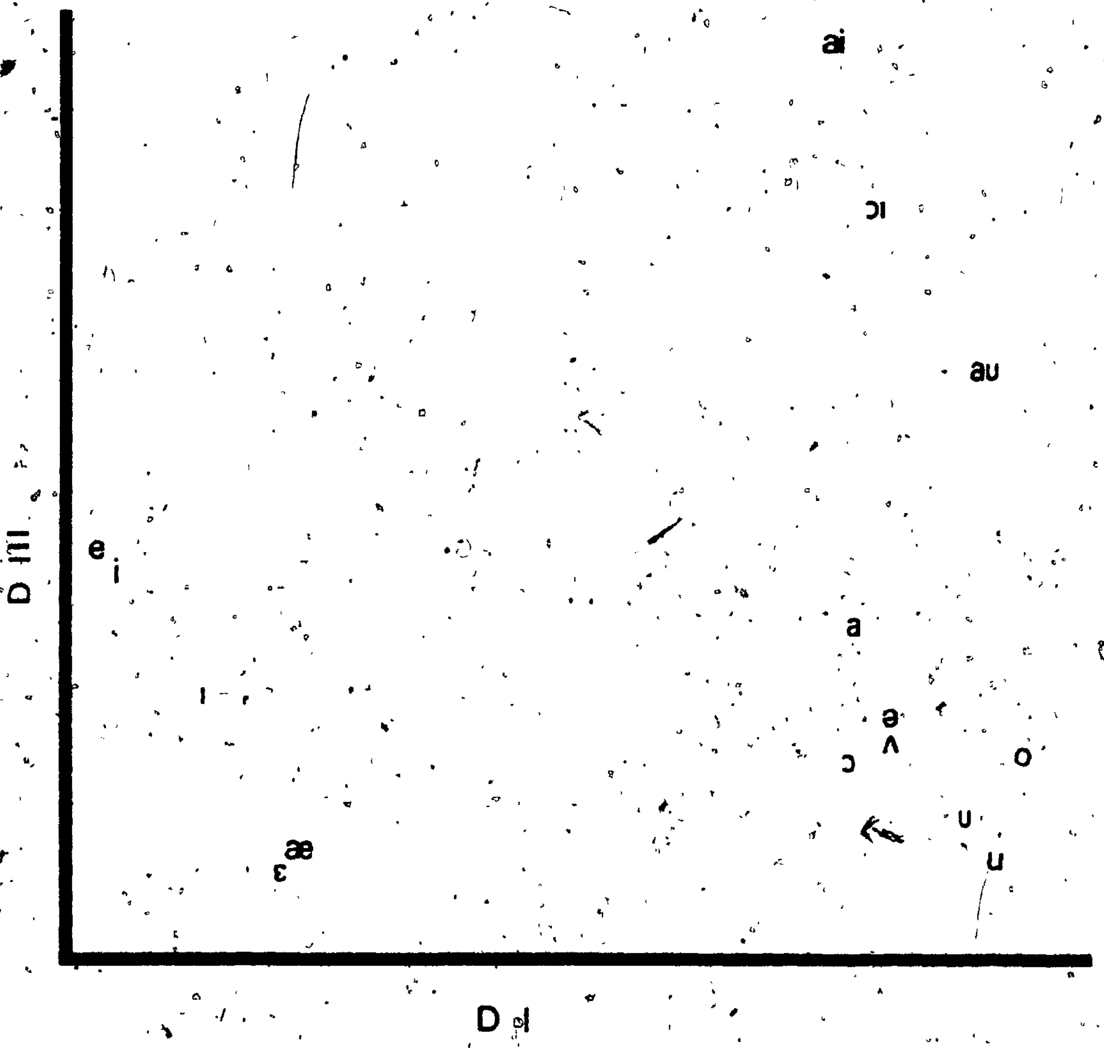

Figure 12. Plot of Dimension III (vertical axis) against Dimension I for the configuration of vowel sounds embedded in /h-d/.

located in the "high" area of the vowel space, while both /ai, aU/ are located in the "low" area of that space. The initial F1 frequencies of the diphthongs appear to be responsible for this distribution. Both /ai, aU/ have the same F1 initial frequency (i.e., 600 Hz) and are located at about the same height along Dimension II. With a lower F1 initial frequency (i.e., 400 Hz) /ɔI/ is placed with the high vowels in the perceptual space. The information carried by the initial F1 frequencies of the diphthongs determines their location in the perceptual representation of the vowel space along Dimension II.

Vowels and diphthongs occupy different areas of the psychological space and are clustered on separate ends of Dimension III. There is a pattern in the ordering of diphthongs along that dimension. The diphthong /aU/ is closest to the body of vowels, followed then by /ɔI/ and /ai/. Examining the physical characteristics of the three diphthong signals we can see from Table 4 that /aU/ is the diphthong exhibiting the least amount of change in its formant frequencies, when they shift from the initial vowel to the terminal one. The shifts are respectively from 600 to 400 Hz for F1 and from 1200 to 900 Hz for F2. In other words, /aU/ is, of the three diphthongs used in the study, the most steady-state like signal. /ɔI/ is next with shifts from 450 to 300 Hz for F1 and from 1000 to 2000 Hz for F2. The diphthong /ai/ is always farthest removed from the body of vowels and indeed exhibits large shifts in both F1 and F2. On the basis of these observations, it can be concluded that Dimension III corresponds to discrimination between monophthongs and diphthongs.

Dimension I, corresponding to F2, or to tongue advancement in terms of articulatory characteristics, accounts in all three contexts for the larger part of the variance when compared to other dimensions. The reason F2 is important is probably due to the fact that it varies more than the other formants. Frequency values of F2 in the present study ranged from 800 to 2100 Hz with only two vowels /ɔ, u/ having the same F2 frequency. Frequency values of F1, on the other hand ranged from 240 to 700 Hz with several vowels having the same F1 frequency or contiguous F1 frequencies. These results are in accord with those of other investigators who found the first two lower formant frequencies, F1 and F2, to be the most important factors in vowel recognition (Cohen, et. al., 1967; Peterson and Barney, 1952; Shepard, 1972). The relationship between the physically measured properties (F1 and F2) of these speech sounds and the psychological structure that was derived independently of those physical measurements provides additional support for the psychophysical notion that it is the frequencies of the first two formants that are critical for the perception of vowel sounds. Furthermore, subjects appear to require the use of a third dimension to almost exclusively discriminate between diphthongs and vowels. This is a finding of interest for there has been no report yet on the perceptual representation of diphthongs. As was previously mentioned the results of this study suggest that diphthongs are clearly differentiated from the rest of the vowel sounds and constitute, for this group of subjects, a separate class of speech sounds.

On the basis of these results, it can be concluded that subjects took into account information carried in the physical representation of the signals about F1, F2 and the steady (vs. shifting) nature of these formants.

Examination of Dimensions IV and V

The same first three dimensions appear to be used in the perceptual representation of vowel sounds independently of phonetic environment.

Not all dimensions, however, are invariant across different phonetic contexts. The fourth dimension accounting for 10.33% of the variance in VPI could not be defined. In both PVP and HVD, however, the fourth dimension seemed to be accounted for by the protensity features tense and lax. Figures 13 and 14 show the configurations in the I-IV plane of the psychological auditory space. These features are acoustically defined as

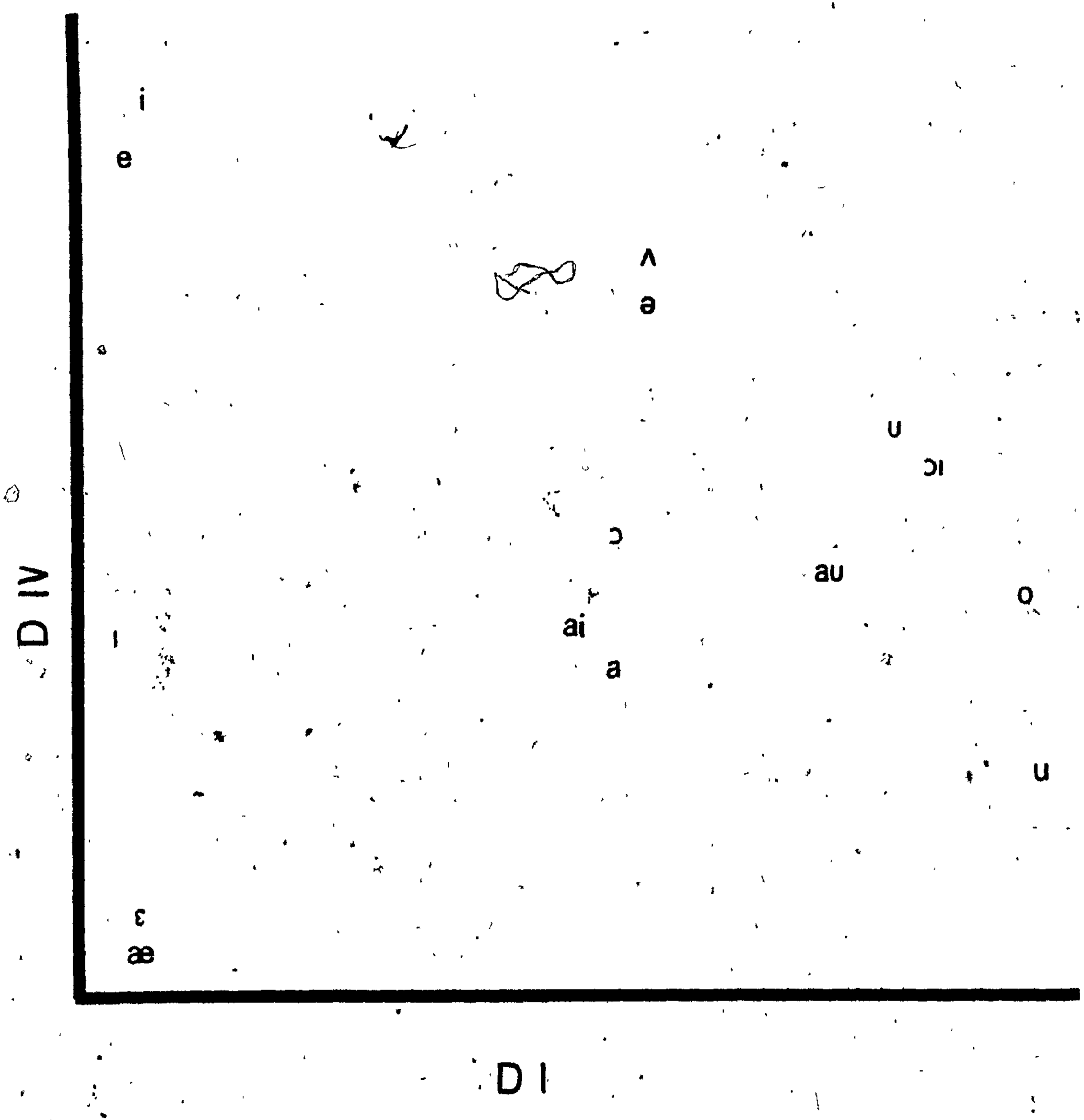

Figure 13. Plot of Dimension IV (vertical axis) against Dimension I for the configuration of vowel sounds embedded in /p-p/.

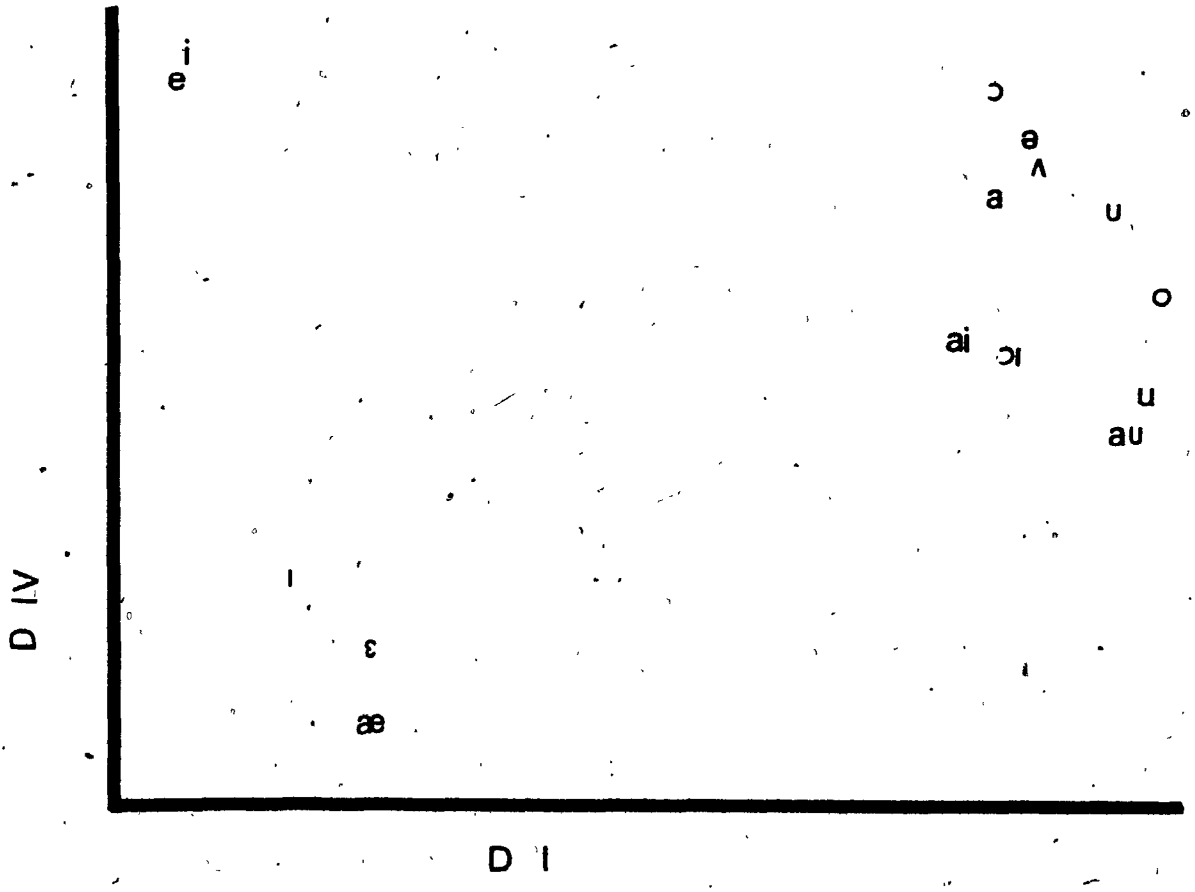

Figure 14. Plot of Dimension IV (vertical axis) against Dimension I for the configuration of vowel sounds embedded in /h-d/.

longer (vs. reduced) duration of the steady state position of the sound. Tense vowels are produced with added muscle tension while the lax vowels are produced without such tension (Jakobson and Halle, 1968). In the English vowel system, some vowels are tense, others are lax, and still others are neutral for the feature tenseness. In this study, the phonologically tense vowels are /i, e, u, o/ and the lax vowels are /I, ɛ, U, /. The vowels /ae, ɔ, a/ are neutral for the feature tenseness. In the perceptual configurations in Figs. 13 and 14, however, /o, u/ appear as lax or as neutral for the feature tenseness. This may be due to their short duration; in fact, these two stimuli are the shortest of all speech sounds used in the study. This leaves /i, e/ as the only tense vowels; at the lax end of the continuum /I, ɛ, U, / can be found; /ae, ɔ, a/ are neutral. The status of the diphthongs is less clear although both /ɔI, aU/ are considered to be lax and /ai/ is considered to be tense (Singh and Singh, 1976). In this respect, both /ɔI, aU/ are properly represented in the psychological space, but /ai/ lies at the lax/neutral end of the tense-lax continuum, suggesting that diphthongs are, in these contexts all lumped together.

The feature tenseness is directly related to the duration of vowel sounds. Except for the vowels at the base of the vowel quadrilateral/ae, a, ɔ/, all long vowels are tense vowels. Duration has been reported in some studies as being, along with F1 and F2, one of the most important factors in vowel recognition (Cohen, et. al., 1967; Pickett, 1957). These studies, however, did not include diphthongs among the set of speech presented to their subjects.

In the HVD context, the fifth and last dimension could not be defined. Its psychological contribution is probably not too important since it accounts for less than 4% of the variance.

The first three dimensions corresponding to F2, F1, and discrimination between diphthongs and vowels appear to be invariant across the three phonetic contexts considered. The fourth dimension was found to be the same in both PVP and HVD and seemed to correspond to the tenseness feature. The fourth dimension in VPI and the fifth dimension in HVD were hard to describe and may simply reflect individual differences.

Individual Differences in the Psychological Representation of Vowel Sounds

An important feature of the multidimensional scaling program used is that it allows for subject variation in the representation of data. Although the dimensions recovered in the psychological space underlie the group's perception of the speech stimuli, the pooling of the group's data may mask interesting individual differences. This program takes into account individual differences in the perception of stimuli: it assumes that the subjects share a common perspective on the speech stimuli but that they nonetheless vary with respect to their relative use of the dimensions underlying the group's perceptual representation. A glance at Table 10, which lists the final weight estimates for each subject on each dimension within each context, is sufficient to reveal the extent of individual variation in the perception of vowel sounds presented in different phonetic environments. Subjects appear to use each dimension differently from other subjects, and to a different extent from one condition to the next.

Although there is wide variation between individuals' final weight estimates the first four dimensions contain few relatively small weight estimates suggesting that most of the subjects used,albeit to a different extent, these first four dimensions. The fifth dimension in HD was literally ignored by five subjects; the rest showed wide variation in their use of that dimension.

Based on the preceding observations, it can be concluded that the first three dimensions underlying the psychological representation of vowel sounds, corresponding respectively to F2, F1, and to steady-state vs. shifting state sounds, are the same across the three

Table 10

Final Weight Estimates

Subject	1st Dimension			2nd Dimension			3rd Dimension			4th Dimension			5th Dimension
	VPI	PVP	HVD	VPI	PVP	HVD	VPI	PVP	HVD	VPI	PVP	HVD	HVD
1	.714	.526	1.224	1.618	.741	.496	1.622	1.228	.971	.047	1.505	2.266	.043
2	1.039	.599	.664	1.281	1.149	1.877	1.289	1.074	2.366	.392	1.179	.050	.044
3	1.149	1.556	.586	.708	.175	.551	1.254	2.196	.398	.888	.072	.481	2.984
4	1.219	2.611	.908	.632	.049	.754	1.065	1.210	1.340	1.083	.129	.913	1.086
5	.990	1.083	1.759	1.142	1.404	.508	.975	.489	.939	.892	1.024	1.750	.044
6	1.932	.465	.581	.381	.784	1.440	1.551	1.155	1.459	.136	1.596	1.475	.044
7	.763	.599	2.205	1.234	.603	1.356	.513	.576	.805	1.491	2.222	.448	.187
8	.204	.617	1.137	1.232	.969	1.053	.051	.048	.239	2.514	2.366	1.892	.679
9	.752	.797	.798	1.178	2.702	.965	.718	.449	.601	1.352	.053	1.579	1.057
10	1.242	.902	.794	1.247	1.382	1.385	.823	.576	.945	.687	1.141	.053	1.822
11	.994	1.051	.537	.272	.588	.510	1.156	2.312	1.247	1.578	.049	.296	2.410
12	.999	1.199	.805	1.075	1.453	1.106	.980	.692	.691	.945	.656	.795	1.604

phonetic contexts considered, i.e., vowels presented in isolation, vowels embedded in /p-p/ and vowels embedded in /h-d/. These three dimensions account for 89.67% of the total variance in the first context; 93.67% in the second; and 83% in the third. Furthermore, Dimension IV, corresponding to the feature tenseness accounted in both PVP and HVD for a small percentage of the variance. Dimension IV in VPI and Dimension V in HVD could not be defined. The percentage of the variance that can be "explained" in each context is now 89.67% in VPI, 100% in PVP, and 96.04% in HVD. Furthermore, individual differences between subjects suggest that although they share a common perceptual space (as reflected by the use of the same dimensions) they nonetheless vary in the extent to which they use these dimensions.

General Discussion

The first three dimensions defining the perceptual representation of the speech stimuli used in the study were found to be stable across three different phonetic contexts and to account for most of the variance. The first two dimensions, corresponding to F2 and F1, represent two of the most basic vowel parameters, tongue advancement and height, and form the foundation of a two dimensional vowel space that is used to describe nearly all languages. This basic two dimensional vowel space has been described in terms of the highest point of the tongue. Compared to other features required for the description of vowels, these two are the only ones that require specifications in terms of a greater number of values than a simple binary contrast (Lindau, 1978). Vowels in English are described as high, mid or low in terms of vowel height and as front, central or back in terms of tongue advancement. The maximum number of contrasts is three for these two particular features in English. As can be seen from Figures 6, 7, and 8, the points representing the speech stimuli appear to be fully contrasted along Dimension II (F1) except for a flattening of the area occupied by the back vowels in PVP, where they are also fully contrasted along Dimension I (F2), i.e., they are grouped in clusters representing front, central and back vowels. This is not the case in HVD where front vowels lie at one end of the dimension and the rest of the speech stimuli at the other end. The contrasts along the first dimension are not as sharply defined in VPI as in the other contexts. It may be that in the course of normal communication we, as speakers and listeners, are seldom exposed to isolated vowel sounds. The CVC syllables, it may be argued, constitute richer and more complex stimuli than isolated vowels, thus leading to a clearer perception of their characteristics. It is evident from reports by other investigators and from the present results that these two features, tongue advancement and height (correlated with F2 and F1), have a special status in the perception of vowel sounds.

As was previously mentioned, Pols, van der Kamp and Plomp (1969) concluded, from the multidimensional scaling analyses of triadic comparisons of the constant vowel part of normally spoken Dutch words of the type hVt, that the most important factors in vowel perception are the first and second formant frequencies.

Another team of investigators (Cohen et. al., 1967) had synthetic isolated Dutch vowel sounds identified by subjects who were able to score 87.3% correct responses. Over three quarters of the errors observed were clearly systematic: either front vowels were judged on the basis of information provided by F1, thus /e/, for example was identified as /o/ and /ɛ/ as /ɔ/, or long vowels were judged as their short counterparts or vice versa, thus /e/ was identified as /I/ or /ɑ/ as /a/. Cohen and his colleagues also concluded that F1 and F2 represented two of the components necessary in the identification of vowel sounds.

Moreover, as was mentioned above, in his analysis of the Peterson and Barney (1952) data, Shepard (1972) derived two major dimensions corresponding to F1 and F2, the contribution of which he believed to be most important in vowel recognition.

The third dimension can be conceived, in a sense, as a byproduct of the first two since it discriminates between those sounds with a stable articulation (steady F1 anf F2 frequencies) and those involving changes in articulation (shifting F1 and F2 frequencies). In other words, it differentiates between the vowels produced with the tongue in a fixed position and the diphthongs produced by moving the tongue from one position to another. Except for the present one, there has been as yet no study that has looked at the perceptual representation of diphthongs, perhaps because of the complexities involved in their production and classification. Although the classification system of vowels, in terms of tongue height and tongue advancement, is not directly applicable to diphthongs, it is of interest to note that subjects nevertheless used the information contained in the initial component of the diphthongs to represent them as back vowels or as high or low vowels as is evident from the consistent representation of these signals along Dimensions I and II across different contexts.

The fourth dimension was found to be stable across two phonetic contexts only, PVP and HVD, and appeared to correspond to the protensity features tense and lax. These features accompany long and short vowels in English and occur with concomitant differences in length. Usually, the tense vowels tend to envelop the lax ones, i.e., to occupy peripheral positions in the vowel quadrilateral with the lax vowels being more central (Lindau, 1978). In this study, two vowels normally classified as tense have apparently been perceived as lax because of their short duration. There are two contrasts in the feature tenseness. This is evident in the distribution of the front vowels along that dimension with tense vowels at the opposite end of the lax vowels. The picture is less clear for the central and back vowels which occupy a relatively compact area of the psychological space and appear to have been represented as lax.

The fourth dimension in VPI departed from the PVP and HVD patterns observed in Dimension IV and could not be related to any feature characteristic of vowels such as F3 and duration, for example. It was also difficult to determine what the fifth dimension in HVD corresponded to.

The dimensions corresponding to F1, F2, and steady (vs. shifting) state were invariant in the three phonetic contexts considered. With respect to F1 and F2, Ainsworth (1971) has investigated the effect of changing the fundamental frequency on the perception of two-formant synthetic vowels and hVd words. He found that raising the fundamental by an octave caused a rise of about 5% in the mean of F1, and of about 2% in the mean of F2, and that isolated and context vowels gave similar results except that the context vowels were recognized more consistently. The experiments were repeated with rising and falling fundamentals, but the results were indistinguishable from those obtained with the steady sounds. Ainsworth (1971) therefore noted that contextual effects in the recognition of synthetic vowels with varying fundamental frequencies were negligible.

Other investigators, using different sets of vowel stimuli and various techniques of analysis provide, by their consistent findings, an indirect confirmation of the invariance of F1 and F2 contribution in the perception of vowels presented in isolation or embedded in /h-d/ (Cohen et. al., 1963, 1967; Peterson and Barney, 1952; Pickett, 1957; Shepard, 1972). In systematically comparing the perceptual representation of vowel sounds in three different phonetic contexts, the present study helps further confirm the contribution of F1 and F2 in vowel perception to the perception of vowels embedded in /p-p/ and adds direct supportive evidence of the invariance of the first two lower formants in the contexts examined.

The third dimension, corresponding to discrimination between vowels and diphthongs, was also found to be invariant across the different phonetic environments. There exists, at this point, in the literature no contradiction or corroboration of these results. It seems, however, reasonable to speculate that, given the important role played by F1 and F2 in the perception of vowel sounds, fluctuations or shifts in F1 and F2 frequencies are likely to be taken into account and significantly contribute to the perception of speech sounds when diphthongs are included among the stimuli.

Contextual effects were found to be significant for the fourth dimension in that it was possible to use the same characteristic or feature to account for that dimension in PVP and HVD, but not in VPI. It may be that tense-lax differences may not be interpretable as contrastive qualities in the case of isolated vowel sounds and that the domain of the tenseness feature is the syllabic structure. In the English language the feature tenseness is functional in the formation of open and closed syllables (Ladefoged, 1975; Singh and Singh, 1976). The tense vowels appear in both open and closed syllables and their lax counterparts generally appear in closed syllables only. Lindau (1978) concluded from radiographic observations that tense and lax vowels differ primarily in tongue height. Since the articulatory feature of tongue height corresponds to the acoustic domain of F1, a tentative conclusion one might draw is that information carried by F1 determines to a certain extent which vowels are perceived as tense and which are perceived as lax.

If this assumption is found to be correct, then, all the dimensions in the psychological representation that have been defined were derived from information carried in the first two lower formant frequencies: Dimension I corresponds to F2, Dimension II to F1, Dimension III to frequency shifts in F1 and F2, and Dimension IV to (apparently) some particular characteristic of F1 in PVP and HVD contexts.

A fifth dimension appeared to be significant in HVD only. Since its meaning could not be defined, it cannot be said whether this dimension relates to features or characteristics of vowel sounds in general or of vowels embedded in /h-d/.

In the context of the present study, subjects evaluated vowel sounds along four perceptual dimensions (or five, depending upon the phonetic environment considered). It has been established that at least three of these dimensions, and probably four, were derived from physical characteristics of the stimuli, namely F1 and F2. Results from most of the studies reviewed suggest that the dimensions defining the perceptual representations of vowel sounds corresponded to F1, F2, and, in some cases to duration (e.g., Cohen, et. al., 1967) and in others to F3 (Shepard, 1972). These results suggest that the physical quality of the auditory stimuli defines the perceptual representation of these stimuli. With two dimensions left unidentified, the question remains, however, to determine whether or not all the recovered dimensions of a psychological space can be related to particular aspects of the physical representation of the auditory stimuli.

One of the contributions of the present experiment is that it can be used to determine whether subjects made use of the discrete nature of distinctive features in their representation of vowel sounds or whether the dimensions were used as continua along which sounds could be ordered. If the discrete nature hypothesis is tenable, the sounds should form clusters in the psychological space such that each cluster occupies a relatively compact area of the space or marks the end of one of the dimensions. In the context of this study, there is evidence to suggest that the group of subjects as a whole contrasted the speech stimuli in a discrete fashion as is evident from the clustering of stimuli in the graphic representations. Subjects appear to fully contrast the speech stimuli in three categories along Dimensions I and II and in a binary fashion along Dimensions III and IV with, in the case of Dimension III, some evidence of continuity in the way diphthongs are represented ranging from most vowel-like to least vowel-like.

The technique used in this study can be a valuable research tool and a whole set of hypotheses may be investigated by obtaining psychological representations for the same set of speech sounds from speakers of different languages. The examination of the differences bewteen the spatial representations of a set of points (speech sounds) by different groups of subjects would probably hint at the features of a language members of a linguistic community consider most important.

Subjects listening to speech sounds in their native language may probably yield a perceptual representation of these sounds different from that yielded by listeners who would have had no prior exposure to that language. The differences between the number of dimensions used and the shape of the final configurations would tell us something about the particular features or characteristics of these sounds that "experienced" and "naive" listeners attend to. One interesting extension of this technique would concern the effect of learning a second language, or of bilingualism, on the psychological representation of speech sounds in an effort to determine in what way experience with other languages affects perception of a new language. There has been some speculation about the transfer of abilities from one language learning situation to another (cf., Segalowitz, 1977) and the results of one study (Cohen, Tucker and Lambert, 1967) suggest that bilinguals perceive phoneme sequences in a language unknown to them more accurately than monolinguals. Two further studies, however, dealing also with children failed to replicate the greater discriminative skills of bilinguals and suggest that at least five years of second language exposure may be necessary for effects to be observed (Davine, Tucker and Lambert, 1971; Lambert, Just and Segalowitz, 1970).

The problem of transfer of skills is a central theme in the psychological learning literature and deserves closer experimental scrutiny. In this respect, adult bilinguals' (or adults who possess experience with more than one language) psychological representation of speech sounds in an unknown language could be examined and compared to that of monolinguals. If there is indeed transfer of skills we would expect the psychological representation of bilinguals to better approximate that of the native speakers of the unknown language than would the monolinguals' representation.

Another line of investigation that can be pursued would deal with hemispheric differences in the perception of speech sounds. Paradigms that have shown differences between classes of sounds include laterality and dichotic masking experiments. Shankweiler and Studdert-Kennedy (1967) have found, for example, that stop consonants are recalled more accurately from the right than from the left ear, whilst the effect is shown less reliably for vowels. Laterality experiments could therefore be conducted to investigate hemispheric differences in the psychological representation of speech sounds and assess each hemisphere's contribution in speech perception. Since the perceptual configuration is a representation of a psychological space, it would be of interest to explore the spatial representations yielded by each hemisphere.

References

Ainsworth, W. A. Perception of synthesized isolated vowels and h-d words as a function of fundamental frequency. Journal of the Acoustical Society of America, 1971, 49, 1323-1324.

Ainsworth, W. A., & Millar, J. B. Methodology of experiments on the perception of synthesized vowels. Language and Speech, 1971, 14, 201-212.

Christian, J. "Psychological differentiation and definition of the self: Multidimensional scaling approach". Unpublished doctoral dissertation. Montréal: McGill University, 1976.

Cohen, A., Slis, I. H., & 't Hart, J. On tolerance and intolerance in vowel perception. Phonetica, 1967, 16, 65-70.

Cohen, S. P., Tucker, G. R., & Lambert, W. E. The comparative skills of monolinguals and bilinguals in perceiving phoneme sequences. Language and Speech, 1967, 10, 159-168.

Davine, M., Tucker, G. R., & Lambert, W. E. The perception of phoneme sequences by monolingual and bilingual elementary school children. Canadian Journal of Behavioral Science, 1971, 3, 72-76.

Darwin, C. J. The perception of speech. In E. C. Carterette and M. P. Friedman (Eds.), Handbook of Perception, Vol. VII. New York, N.Y.: Academic Press, 1976.

Elovitz, H. S., Johnson, R., McHugh, A., & Shore, J. E. Letter-to-sound rules for automatic translation of English text to phonetics. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1976, ASSP-24, 446-458.

Gerstman, L. H. Classification of self-normalized vowels. IEEE Transactions on Audio Electroacoustics, 1968, AU-16, 78-80.

Jakobson, R., & Halle, M. Phonology in relation to phonetics. In B. Malmberg (Ed.), Manual of Phonetics, Amsterdam, New York: North Holland/American Elsevier, 1968.

Ladefoged, P. A course in phonetics. New York, N.Y.: Harcourt, Brace, Jovanovich Inc., 1976.

Ladefoged, P., & Broadbent, D. E. Information conveyed by vowels. Journal of the Acoustical Society of America, 1957, 29, 98-104.

Lambert, W. E., Just, M., & Segalowitz, N. Some cognitive consequences of following the curricula of the early school grades in a foreign language. Monograph Series on Language and Linguistics. J. Alatis (Ed.). Washington, D.C.: Georgetown University Press, 1970.

Lieberman, P., Crelin, E. S., & Klatt, D. H. Phonetic ability and related anatomy of the newborn, adult human, neanderthal man, and the chimpanzee. American Anthropologist, 1972, vol. 74, 287-307.

Lindau, M. Vowel features. Language, 1978, 54, 541-563.

Lindblom, B. E. F. Phonetics and the description of language (Ms.). Proceeding of the 7th International Congress of Phonetic Sciences, Montreal.

Miller, G. A., & Nicely, P. E. An analysis of perceptual confusions among some English consonants. Journal of the Acoustical Society of America, 1955, 27, 338-352.

Mohr, B., & Wang, W. S.-Y. Perceptual distance and the specification of phonological features. Phonetica, 1968, 18, 31-45.

Peters, R. W. Dimensions of perception of consonants. Journal of the Acoustical Society of America, 1963, 35, 1985-1989.

Peterson, G. E., & Barney, H. L. Control methods used in a study of the vowels. Journal of the Acoustical Society of America, 1952, 24, 175-184.

Pickett, J. M. Perception of vowels heard in noises of various spectra. Journal of the Acoustical Society of America, 1957, 29, 613-620.

Plomp, R., Pols, L. C. W., & van de Geer, J. P. Dimensional analysis of vowel spectra. Journal of the Acoustical Society of America, 1967, 41, 707-712.

Pols, L. C. W., van der Kamp, L. J. Th., & Plomp, R. Perceptual and physical space of vowel sounds. Journal of the Acoustical Society of America, 1969, 46, 458-467.

Potter, R. K., & Peterson, G. E. The representation of vowels and their movements. Journal of the Acoustic Society of America, 1948, 20, 528-535.

Potter, R. K., & Steinberg, J. C. Toward the specification of speech. Journal of the Acoustical Society of America, 1950, 22, 807-820.

Ramsay, J. O. Maximum likelihood estimation in multidimensional scaling. Psychometrika, 1977, 42, 241-266.

Ramsay, J. O. Multiscale: Four programs for multidimensional scaling by the method of maximum likelihood. Chicago, Ill.: National Educational Resources, 1978.

Segalowitz, N. Psychological perspectives on bilingual education. In B. Spolsky and R. L. Cooper (Ed.), Frontiers in bilingual education. Rowley, Mass.: Newbury House, 1977.

Shankweiler, D. P., & Studdert-Kennedy, M. Identification of consonants and vowels presented to left and right ears. Quarterly Journal of Experimental Psychology, 1967, 19, 59-63.

Shepard, R. N. Psychological representation of speech sounds. In E. E. David and P. B. Denes (Eds.), Human communication: A unified view. New York, N.Y.: McGraw Hill, 1972.

Singh, S., & Singh, K. S. Phonetics: Principles and Practices. Baltimore: University Park Press, 1976.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. What information enables a listener to map a talker's vowel space? _Journal of the Acoustical Society of America_, 1976, _60_, 198-212.

Vinegrad, M. D. A direct magnitude scaling method to investigate categorical versus continuous modes of speech perception. _Status Report on Speech Research SR-21/22_, Haskins Laboratories, 1970.

Welch, P. D., & Wimpress, R. S. Two multivariate statistical computer programs and their application to the vowel recognition problem. _Journal of the Acoustical Society of America_, 1961, _33_, 426-434.

Wickelgren, W. A. Distinctive features and errors in short-term memory for English consonants. _Journal of the Acoustical Society of America_, 1966, _39_, 388-398.

Wilson, K. V. Multidimensional analyses of confusions of English consonants. _American Journal of Psychology_, 1963, _76_, 89-95.

Wilson, K. V., & Saporta, S. Linguistic organization. In C. E. Osgood and T. A. Sebeok (Eds.), _Psycholinguistics: A survey of theory and research problems._ Indiana University Press: Bloomington, Ind., 1965.