CHAPTER FOUR

Psychological Representation of Speech Sounds

ROGER N. SHEPARD*

Introduction

Since few of us have the time or interest to read more than a small fraction of the printed information with which we are increasingly bombarded, speech still carries the greatest burden of human communication. Under normal circumstances, then, communication is a primarily oral-aural transaction. For, as our great reliance upon the telephone attests, such communication can be effectively carried on in the complete absence of all but purely auditory cues, whereas (except in the special case of highly skilled lip readers) it begins to deteriorate as soon as the auditory signal is degraded by attenuation, distortion, or noise.

* The analyses reported here owe much to my former coworkers S. C. Johnson for the development of the algorithm for hierarchical clustering and Mrs. J.-J. Chang for programming the exponential method of proximity analysis [1]. Extensive assistance in carrying out the analyses and preparing the figures was provided by Carolyn Brown, Maureen Sheenan, and, particularly, Christine Feng. This work was supported by Bell Laboratories and the National Science Foundation (Grants GS-1302 and GS-2283).

If, now, we turn to a study of this auditory signal upon which most communication so critically depends, we find that it, in turn, can be analyzed into a sequence of distinguishable, psychologically elementary components—the *phonemes.* These are the smallest units, the individual vowels and consonants, that we intuitively recognize as separately producible speech sounds. In terms of the articulatory organs (lips, tongue, glottis, etc.), they correspond to more or less identifiable configurations or simple changes (openings, closures, etc.). In terms of the resulting acoustic signal, they correspond to more or less identifiable energy patterns in the time-frequency domain.

Basic to human communication, then, is the ability to recognize or identify these individual phonemes as they follow one another in the auditory stream of speech. This is not to say that the understanding of connected discourse consists solely in the separate recognition of each phoneme as it occurs. For, unlike the identification of isolated phonemes in the laboratory, the recognition of those very same sounds in the course of connected speech is typically aided by additional processes of a rather different kind, viz., processes of segmentation and utilization of semantic and syntactic cues from context. Still, the process of recognition of individual phonemes must be regarded as fundamental in the sense that it can be shown to operate in the absence of the context upon which these higher-level processes depend; whereas any context that is in fact present consists largely, itself, of at least partially recognized phonemes.

For these reasons, investigators concerned with the perception of speech have devoted considerable effort to the study of the human listener's ability to identify vowel and consonant phonemes—even when these are presented without context. Typically in these studies, individual phonemes (i.e., phonemes that have been either uttered in isolation or excerpted in some way from context) are presented in random order to listeners who indicate, after each presentation, which phoneme they think has been presented. The data from such an *absolute-identification experiment* are most conveniently cast in the form of an $n \times n$ matrix showing, for each of the n phonemes studied, how often it was identified correctly and how often it was misidentified as each of the $n - 1$ other phonemes. The hope is that the numbers in such a *confusion matrix* will reveal something about how these speech sounds are processed within the listener.

Further information can be gained by repeating an experiment of this kind while systematically masking or filtering the acoustic signal in some way. Indeed, the capabilities for processing speech have attained such a level of perfection in humans that, under favorable conditions, confusions between phonemes occur only rarely—even when they are presented without context. In order to ensure a rate of confusion that is sufficiently revealing of the underlying perceptual mechanism, therefore, it is often helpful to degrade the stimulus deliberately, e.g., by the addition of noise. Moreover, by examining the changes in the patterns of confusion that result when we selectively mask or filter out particular ranges of audio frequency, we can obtain more direct evidence on how the process of identification depends upon the physically measurable parameters of the stimuli.

It is important to recognize, though, that any one confusion matrix is based solely upon the responses of the listeners and can, therefore, be constructed without any knowledge of the physical properties of the stimuli. In this respect any structure or pattern that may be discerned in such a matrix of numbers represents the interrelations among the stimuli as they are perceived psychologically—not as they are measured physically. Once such a purely psychological structure has been identified in this way, however, we are in a favorable position to turn to the *psychophysical* problem of then relating this structure to any independently measured physical parameters.

If we are lucky, the results might even lead to more sophisticated devices for the automatic recognition or for the efficient compression and transmission of speech (technological goals that are discussed in Chapter 11). Here, however, we shall focus primarily on the more purely psychological problem, viz., the problem of discovering the psychological structure of a set of speech sounds as that structure is revealed solely in the responses of listeners.

This substantive problem will also serve as a vehicle for demonstrating the advantages of some recent methodological innovations, namely some computer-based techniques for transforming patterns hidden in large matrices of empirical data into a readily assimilable pictorial form. In this respect, there will be a close connection with some of the later chapters (e.g., Chapters 7 and 8) that are also concerned with the use of machines to convert raw data into a more usable form.

Confusion Data as a Source of Information about Psychological Structure

Let us turn now to a more detailed consideration of the nature of a confusion matrix of the sort we have been considering, and to the problem of discovering and characterizing whatever pattern or structure may lay hidden in its numbers. Following the usual convention, let us suppose that these are arranged in such a way that the entry at the intersection of the *i*th row and *j*th column tells us how often the *i*th stimulus led to the response that would be correct for the *j*th stimulus. Thus "degenerate" cases [22, p. 240], nothing need be known about the form of the function relating the given data to the Euclidean distances to be recovered (except, of course, that this function is monotonic). This initially unknown form can nevertheless be essentially recovered, and to the extent that it is recovered, the spatial solution is determined to within the "extended" group of similarity transformations [26]. Since the "degenerate" exceptions have arisen only rarely in practice, much can be accomplished without assuming more than monotonicity.

Even so, any additional knowledge that we may have about the functional form of the underlying relation can be used to increase the precision of the spatial representation and to decrease its susceptibility to the sometimes troublesome "degeneracies" [27, 28]. It is therefore of some potential import that in the specific case of confusion data, the relation in question—even though assumed to be no more than monotonic—has in fact consistently turned out to have a very particular, well-defined form. With the exception of occasional degenerate cases, that is, the relation between the frequencies with which stimuli are confused and the corresponding distances among points in the recovered spatial solution has invariably been found to be closely approximated by a simple exponential decay function.

The particular method for the "proximity analysis" of confusion data to be illustrated here takes advantage of this empirical result and—instead of seeking the spatial representation that best fits the merely monotonic hypothesis—seeks the representation that best fits the stronger, exponential hypothesis. Although an analysis of this sort was originally proposed over 15 years ago [12, 29–32], it was not until the more recent implementation of appropriate numerical methods on a digital computer [21, 22, 24, 25] that it has become clear just how such an analysis should be achieved. The procedure described here follows Kruskal's [24] lead of using a variant of the method of "steepest descent" to minimize an explicitly defined measure of departure from the desired relation—in this case an exponential decay function. The computer program itself was developed primarily by Mrs. J.-J. Chang in collaboration with the present author [1].

We start with the empirically given confusion matrix in which the entry p_{ij} is the relative frequency with which the ith stimulus led to the response belonging to the jth stimulus. In this raw form, however, the entries p_{ij} are not entirely suitable as measures of proximity, for whereas we require that the proximity between i and j (like the distance between any two points) be the same in both directions, it will not in general be the case that $p_{ij} = p_{ji}$. Moreover, whereas the proximity (or distance) between a point and itself always takes on the same limiting value, it will not usually happen that $p_{ii} = p_{jj}$ for different stimuli i and j.

Before proceeding to the analysis itself, then, we first define for every pair i and j a derived estimate of the psychological proximity or similarity S_{ij} between them in terms of the four relevant numbers from the empirically given confusion matrix, viz., p_{ij}, p_{ji}, p_{ii}, and p_{jj}. A definition that has been found serviceable for this purpose is, simply, the total number of confusions between i and j divided by the total number of correct responses to these same two stimuli or, formally,

$$S_{ij} = \frac{p_{ij} + p_{ji}}{p_{ii} + p_{jj}}$$

Although there are theoretical arguments for other, slightly different definitions [30, 31], the formula just given has typically led to essentially the same results in practice and, moreover, has seemed to be somewhat less affected by statistical fluctuations in the four p components. In any case we have, by means of this formula, achieved the desired condition of symmetry, $S_{ij} = S_{ji}$, and equality of the diagonal entries, $S_{ii} = S_{jj} = 1$ for all i and j.

The iterative process itself is applied to the resulting symmetric matrix of these proximity measures, S_{ij}, to find a configuration of points and a set of values for the parameters of the postulated exponential that provide, jointly, an optimum fit to the S_{ij}. Specifically, we seek to minimize

$$\sum_{i>j} \{S_{ij} - (ae^{-bD_{ij}} + c)\}^2$$

where D_{ij} = distance between points i and j in recovered configuration
a = intercept of recovered exponential decay function
b = slope of recovered exponential decay function
c = asymptote of recovered exponential decay function

The D_{ij} are in turn computed from the coordinates for the points recovered during each iteration by the Euclidean formula

$$D_{ij} = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$$

where x_{ik} is the coordinate for the ith point on the kth orthogonal dimension of the underlying space.

Actually, in the analyses of proximity measures derived specifically from confusion matrices in the manner indicated above, we always have $S_{ii} = 1$ for all i. Since the distance D_{ii} between a point and itself is zero, it is natural to ensure that the fitted exponential function intercepts the S axis at unit height. In the analyses to be reported here, therefore, the parameter a is not allowed to vary along with b and c but is constrained instead to the fixed value 1. Also, in order to facilitate convergence (and to avoid merely local minima), we have found it helpful to add two further refinements to this method: (1) a procedure for choosing starting values for the x coordinates and the variable parameters b and c on a rational rather than a purely arbitrary basis; and (2) a second-order variant of the gradient method in which the adjustments during each iteration are based upon estimates of the second, as well as the first, partial derivatives of the expression to be minimized with respect to the variable coordinates and parameters. Additional refinements have also been introduced to ensure that the variable parameters b and c will end up with appropriate signs. These and other details of the computing algorithm are somewhat tangential to our present focus on substantive problems of speech perception and so will be reserved for fuller presentation elsewhere.

Spatial Representation Based on Confusions Among 16 Consonant Phonemes

The data with which we shall be most extensively concerned, here, are from the classical study of confusions among 16 English consonants by Miller and Nicely [33]. Basically, the listeners in their experiments attempted to identify each of 16 different syllables as these were pronounced from a randomized list. All syllables terminated in the same vowel /a/ (as in f*a*ther). They differed only in the consonant that preceded this vowel, which was always one of the 16 following: /p/, /t/, /k/, /f/, /θ/, /s/, /ʃ/, /b/, /d/, /g/, /v/, /ð/, /z/, /ʒ/, /m/, and /n/.

Altogether, Miller and Nicely obtained 17 complete 16 × 16 confusion matrices, each under a different condition of filtering or signal-to-noise ratio, and, very fortunately for our present purposes, these were presented in full in their published report. To start with, however, we shall confine our attention to just one matrix, obtained by pooling their tables I through VI for those six conditions in which bandwidth was maintained at 200 to 6,500 Hz and in which deviations from the best listening condition were imposed only by manipulating signal-to-noise ratio (S/N). Combining data in this way permits the recovery of a spatial representation of greater stability without entailing any untoward biases. For, as we shall later see, although variations in S/N do affect the overall *number* of confusions, they have little or no effect on the internal *pattern* of those confusions. Table 4.1, then, presents the below-diagonal half of the 16 × 16 symmetric matrix of proximity measures obtained from the pooled confusion matrix by means of the formula for S_{ij} set forth above.

The problem is to convert into explicit form whatever pattern already exists implicitly in the data of this table. This is essentially achieved

Table 4.1 Measures of Confusion Among Sixteen Consonants (Based upon Miller and Nicely's data for their six unfitered or "flat" conditions I–VI.)

p	—															
t	.229	—														
k	.432	.241	—													
f	.101	.057	.077	—												
θ	.124	.079	.084	.423	—											
s	.052	.050	.063	.066	.157	—										
ʃ	.038	.050	.047	.030	.048	.115	—									
b	.022	.013	.018	.046	.045	.024	.012	—								
d	.025	.022	.020	.025	.041	.031	.033	.058	—							
g	.013	.016	.030	.015	.039	.033	.021	.069	.342	—						
v	.016	.022	.020	.035	.040	.023	.020	.210	.059	.054	—					
ð	.028	.016	.018	.032	.031	.026	.018	.145	.094	.120	.338	—				
z	.025	.023	.025	.018	.033	.035	.017	.055	.106	.139	.080	.161	—			
ʒ	.019	.017	.019	.007	.017	.022	.012	.027	.089	.125	.029	.033	.136	—		
m	.025	.022	.021	.016	.019	.017	.012	.038	.024	.032	.030	.034	.021	.016	—	
n	.017	.018	.020	.012	.018	.013	.011	.024	.032	.030	.022	.028	.016	.030	.151	—
	p	t	k	f	θ	s	ʃ	b	d	g	v	ð	z	ʒ	ɱ	ŋ

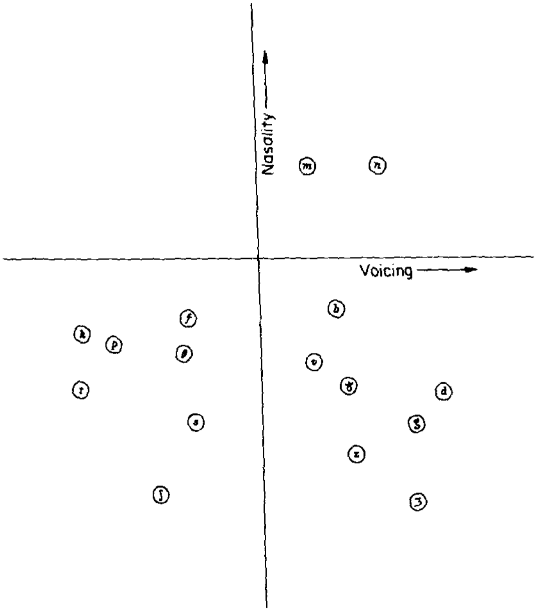

Fig. 4.1. Two-dimensional spatial representation for 16 consonant phonemes (based on the pooled data from Miller and Niceley's six flat-frequency-response conditions).

in Fig. 4.1, which presents the two-dimensional spatial representation obtained by applying the method of "exponential analysis of proximities" to the empirical proximity measures of Table 4.1.

The recovered configuration has been rigidly rotated so that the vertical and horizontal axes of the figure partition the 16 points into those representing the unvoiced stops and fricatives /ptkfθsʃ/, the corresponding voiced stops and fricatives /bdgvðzʒ/, and the (voiced) nasals /mn/. Now, by an entirely different method, Miller and Nicely had already demonstrated that among the so-called "distinctive features" that linguists have invoked as a basis for classifying consonant phonemes, those of voicing and nasality go the furthest toward accounting for these data. The emergence of the three-way grouping noted in Fig. 4.1 does not therefore provide in itself any new insight into these data. Still, it does at least attest to the potential validity of the gross features of spatial representation of this type.

Moreover, in view of the fact that these consonants were considered by Miller and Nicely to differ with respect to as many as *five* different distinctive features (viz., affrication, duration, and place of articulation, as well as voicing and nasality), it is noteworthy that 99.4 percent of the variance of the data in Table 4.1 can be accounted for solely on the basis of the two dimensions of Fig. 4.1. The residual departures of the original proximity measures S_{ij} from the reconstructed exponential decay function of the distances D_{ij} in this recovered two-dimensional space are displayed in Fig. 4.2. (Each point in this plot corresponds, of course, to a pair of the consonants.)

The fact that we can obtain such a good fit in two dimensions does not, however, mean that these dimensions must be interpreted as reflecting variations in two distinctive features only. On the contrary, as we move from left to right across the lower half of Fig. 4.1, we encounter four discernible vertically oriented groups consisting, successively, of the unvoiced stops /ptk/, the unvoiced fricatives /fθsʃ/, the voiced fricatives /vðzʒ/, and, not entirely separated from the preceding group, the voiced stops /bdg/. So the distinction of affrication is at least partially preserved. Further, the parallelism between the unvoiced and voiced fricatives with respect to place of articulation (and possibly duration) is maintained in the parallel vertical ordering of the two series of corresponding points for /fθsʃ/ and /vðzʒ/. Evidently it would be an oversimplification, then, to describe this space solely in terms of the distinctive features of voicing and nasality.

Of course, we can always seek a new best-fitting solution in a space of higher dimensionality, and the fraction of variance accounted for

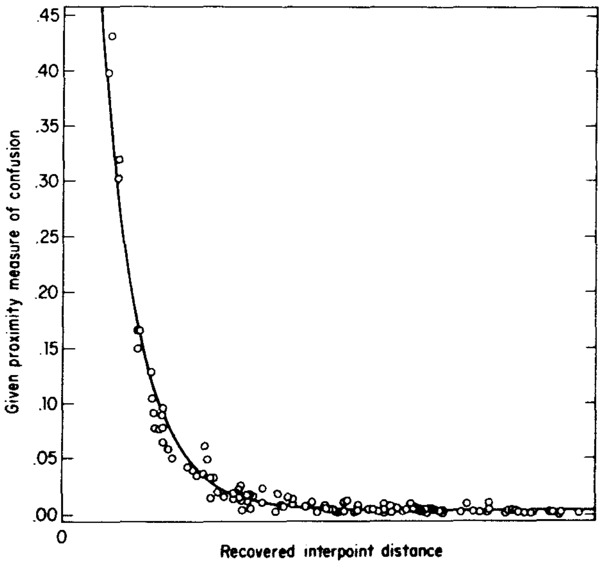

Fig. 4.2. Goodness of fit of the confusion data for 16 consonants to an exponential decay function of interpoint distance in the two-dimensional representation of Fig. 4.1.

in the given proximity data can only be increased thereby. Indeed, previous attempts have been made to extract as many as four spatial dimensions from these same data [20]. So the two-dimensional results shown in Figs. 4.1 and 4.2 do not fully settle the question of the true psychological dimensionality of these stimuli. Still, with less than 1 percent of the variance remaining to be accounted for, relatively little psychological importance could be attached to any additional spatial dimension that might be extracted. Indeed the error variability inherent in the S-measures themselves may well amount to as much as 1 percent. If so, the determination of the projections of the points on a third axis—with a concomitant increase, by half again, in the total degrees of freedom of the configuration—is not likely to possess much reliability.

A more promising alternative for evincing further structure in these data will be described in a subsequent section. An attempt will then be made to show how that alternative technique can be used in combination with the spatial type of solution just discussed to reveal further regularities in the data of Miller and Nicely. Already, however, it should be clear that the spatial solution by itself can furnish a useful reduction of the original confusion data. It is a reduction in that all 120 of the S-values originally given in Table 4.1 can now be reconstructed (except, of course, for the approximately ½ percent residual variance) from just 34 numbers (viz., the 2 × 16 spatial coordinates together with the two parameters of the fitted exponential). It is a *useful* reduction in that the information or structure, which was only implicit in Table 4.1, has been converted in Fig. 4.1 into an immediately accessible, explicit form.

Spatial Representation Based on Confusions Among 10 Vowel Phonemes

Before presenting the second of the two methods of analysis to be described here, it may be helpful to consider a second illustrative application of the first method—this time to vowel rather than consonant phonemes. In a well-known study by Peterson and Barney [34], ten monosyllabic words differing only in the vowel were presented aurally in random order to a group of listeners who were requested to indicate after each presentation which of the ten words they thought had been pronounced. The words ("heed," "hid," "head," "had," "hod," "hawed," "hood," "who'd," "hud," and "heard") were alike with respect to the initial and terminal consonants /h/ and /d/ but differed with respect to the interpolated vowel, /i/, /I/, /ε/, /æ/, /a/, /ɔ/, /ʊ/, /u/, /ʌ/, or /ɝ/, respectively. The difficulty of the task was partly determined by the further circumstance that in different presentations the same word was pronounced by different speakers (who varied in both age and sex).

A 10 × 10 symmetric matrix of proximity measures was computed from the resulting confusion matrix [34, table I], and was then subjected to the exponential analysis of proximities as before. Best-fitting configurations were sought in both two-dimensional and three-dimensional space. The resulting three-dimensional solution is shown in Fig. 4.3, where an attempt has been made to encode variations in the third dimension (depth) by variations in the sizes of the spheres representing the ten points. The obtained three-dimensional structure has some face validity. At least, words that seem similar in sound (like "hod" and "hawed") are close together, whereas words that seem relatively different (like "heed" and "hud") are far apart.

The extent to which the distances among the ten points can be used to account in this way for the actual frequencies with which these words were confused is indicated in Fig. 4.4. This time the exponential decay function, with which the distances were brought into a mutual best fit, accounts for 99 percent of the variance of the given proximity data. (Actually, however, because so many of the S-values are essentially zero, this is somewhat less impressive than the 99.4 percent reported in connection with Fig. 4.2.)

Further evidence for the validitiy of such a spatial representation can be sought in its relations with other, external variables. In this connection, it should be noted that it is only the internal structure of the configuration itself that is uniquely determined by the data; the orientation of the recovered configuration with respect to the reference axes is wholly arbitrary. There is, therefore, no reason to suppose that any given external variable (such as the physically measured frequency of a particular formant) will be especially related to the projections

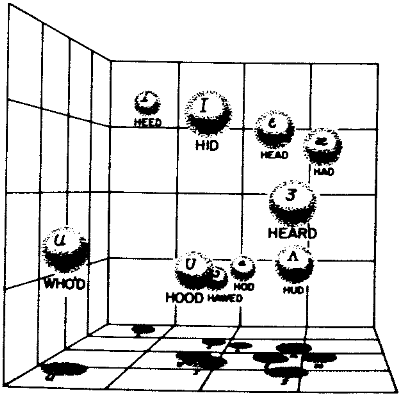

Fig. 4.3. Three-dimensional spatial representation for 10 vowel phonemes (based on the data of Peterson and Barney).

in the given proximity data can only be increased thereby. Indeed, previous attempts have been made to extract as many as four spatial dimensions from these same data [20]. So the two-dimensional results shown in Figs. 4.1 and 4.2 do not fully settle the question of the true psychological dimensionality of these stimuli. Still, with less than 1 percent of the variance remaining to be accounted for, relatively little psychological importance could be attached to any additional spatial dimension that might be extracted. Indeed the error variability inherent in the S-measures themselves may well amount to as much as 1 percent. If so, the determination of the projections of the points on a third axis—with a concomitant increase, by half again, in the total degrees of freedom of the configuration—is not likely to possess much reliability.

A more promising alternative for evincing further structure in these data will be described in a subsequent section. An attempt will then be made to show how that alternative technique can be used in combination with the spatial type of solution just discussed to reveal further regularities in the data of Miller and Nicely. Already, however, it should be clear that the spatial solution by itself can furnish a useful reduction of the original confusion data. It is a reduction in that all 120 of the S-values originally given in Table 4.1 can now be reconstructed (except, of course, for the approximately ½ percent residual variance) from just 34 numbers (viz., the 2 × 16 spatial coordinates together with the two parameters of the fitted exponential). It is a *useful* reduction in that the information or structure, which was only implicit in Table 4.1, has been converted in Fig. 4.1 into an immediately accessible, explicit form.

Spatial Representation Based on Confusions Among 10 Vowel Phonemes

Before presenting the second of the two methods of analysis to be described here, it may be helpful to consider a second illustrative application of the first method—this time to vowel rather than consonant phonemes. In a well-known study by Peterson and Barney [34], ten monosyllabic words differing only in the vowel were presented aurally in random order to a group of listeners who were requested to indicate after each presentation which of the ten words they thought had been pronounced. The words ("heed," "hid," "head," "had," "hod," "hawed," "hood," "who'd," "hud," and "heard") were alike with respect to the initial and terminal consonants /h/ and /d/ but differed with respect to the interpolated vowel, /i/, /I/, /ɛ/, /æ/, /a/, /ɔ/, /ʊ/, /u/, /ʌ/, or /ɝ/, respectively. The difficulty of the task was partly determined by the further circumstance that in different presentations the same word was pronounced by different speakers (who varied in both age and sex).

A 10 × 10 symmetric matrix of proximity measures was computed from the resulting confusion matrix [34, table I], and was then subjected to the exponential analysis of proximities as before. Best-fitting configurations were sought in both two-dimensional and three-dimensional space. The resulting three-dimensional solution is shown in Fig. 4.3, where an attempt has been made to encode variations in the third dimension (depth) by variations in the sizes of the spheres representing the ten points. The obtained three-dimensional structure has some face validity. At least, words that seem similar in sound (like "hod" and "hawed") are close together, whereas words that seem relatively different (like "heed" and "hud") are far apart.

The extent to which the distances among the ten points can be used to account in this way for the actual frequencies with which these words were confused is indicated in Fig. 4.4. This time the exponential decay function, with which the distances were brought into a mutual best fit, accounts for 99 percent of the variance of the given proximity data. (Actually, however, because so many of the S-values are essentially zero, this is somewhat less impressive than the 99.4 percent reported in connection with Fig. 4.2.)

Further evidence for the validitiy of such a spatial representation can be sought in its relations with other, external variables. In this connection, it should be noted that it is only the internal structure of the configuration itself that is uniquely determined by the data; the orientation of the recovered configuration with respect to the reference axes is wholly arbitrary. There is, therefore, no reason to suppose that any given external variable (such as the physically measured frequency of a particular formant) will be especially related to the projections

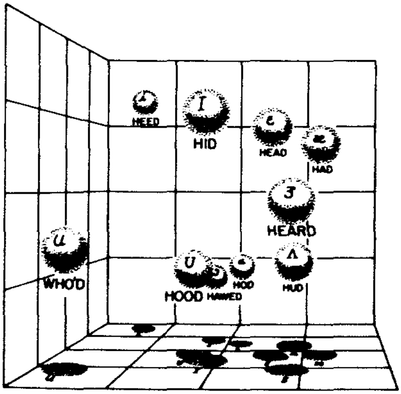

Fig. 4.3. Three-dimensional spatial representation for 10 vowel phonemes (based on the data of Peterson and Barney).

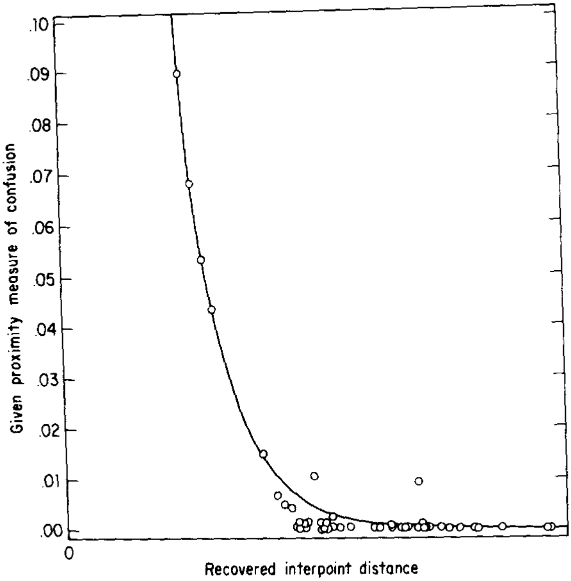

Fig. 4.4. Goodness of fit of the confusion data for 10 vowels to an exponential decay function of interpoint distance in the three-dimensional representation of Fig. 4.3.

x_{ik} of the recovered points on any one reference axis k. This consideration has often been overlooked in interpreting multidimensional scaling solutions (e.g., see Ref. 35). Actually, for the purpose of interpretation, it is quite permissible—indeed preferable—to seek for each such external variable a new axis on which the projections of the points are optimally correlated with that variable [36, 37].

In general, of course, to provide a specification of a speech sound that is complete *physically* would require a very large number of parameters. Even if we disregard phase relations, we would still need to partition the time-frequency plane into a vast number of small compartments and then specify the amount of acoustic energy contained within each such compartment. In the case of vowels, however, it is only the center frequencies of the lowest two or three formants (peaks in the energy spectrum) that have been thought to be critical *psychologically*—i.e., for recognition (see, e.g., Refs. 38 and 39). Accordingly, the method reported by Miller, Shepard, and Chang [37] was used to find three new axes through the configuration shown in Fig. 4.1 such that each would best agree with the average center frequency of a different one of the first three formants of these vowels as measured by Peterson and Barney [34, table II]. In the average formant frequencies used for this purpose, the values given by Peterson and Barney, separately, for each of their three types of speakers (viz., men, women, and children) were weighted according to the number of each type in their total sample of 76 speakers (viz., 33, 28, and 15, respectively).

The resulting product-moment correlations with the projections on the new, rotated axes were .88, .98, and .82 for the first, second, and third formants, respectively. As is to be expected, these relations between the purely physical variables and the recovered psychological structure were appreciably nonlinear. To this extent, the true strengths of these relations are even greater than indicated by the obtained linear correlations.

The angles between the new directions corresponding to these three formants were 99° for one and two, 104° for one and three, and 49° for two and three. That the first axis is nearly orthogonal (close to 90°) to the other two indicates that the psychological effect of the first formant is nearly independent of the other two. The substantially smaller angle between the second and third directions, on the other hand, suggests that the psychological effects of these two higher formants are to some extent interdependent.

In Fig. 4.3 the three-dimensional configuration is pictured as viewed normal to the plane of the first two rotated axes. Indeed, horizontal and vertical axes through the pictured structure would, as nearly as possible, agree with the frequencies of the first and second formants, respectively. This is why the picture reproduces something resembling the traditional "vowel loop" (starting with the "high front" vowel /i/ at the upper left and proceeding around the U-shaped curve through the "lower" vowels at the right and back to the "high back" vowel /u/ at the lower left). The rotated axis that best relates to the frequency of the third formant points both away (into the depth of the picture) and upward (since it is also correlated with the second).

Because the 10 recovered points fall roughly on a U-shaped "loop" and because the axes corresponding to the second and third formants are appreciably correlated, a good approximation to the original data can also be achieved in a reduced space of just two dimensions. In fact, 97 percent of the variance of the given S-values can still be accounted for by the best-fitting two-dimensional solution. In appearance this reduced configuration is rather similar to the projection of the three-dimensional configuration on the plane of the first two axes. That is, it is not very different from the pattern already shown in Fig. 4.3—but with all ten spheres reduced to the same size.

In either case, a rather clear relationship emerges between certain physically measured properties of these stimuli and a "psychological" structure that was obtained entirely independently of those physical measurements. The existence of this relationship provides a different kind of support for the psychophysical notion that it is the frequencies of the first three formants—and particularly of the first two—that are critical for the identification of spoken vowels. At the same time, it furnishes further evidence of the validity of such spatial representations of psychological structure.

Hierarchical Representation of Confusion Data by Cluster Analysis

The representation of stimuli as points in an underlying continuous space seems particularly natural when the psychologically relevant physical variables are themselves inherently continuous, as in the case of the formant frequencies of the vowels. In cases of this kind there is reason to suppose that the psychological space is a continuous (and probably differentiable) deformation of the space defined by the basic physical variables [30, 31]. It is therefore understandable that, when we have recovered such a psychological space, the existence of relevant and measurable physical variables can, as in the case of the vowels, greatly facilitate the interpretation of that space.

Sometimes, however, as in the case of the consonants, we do not yet have a good hold on the relevant physical variables. Perforce, in attempting interpretations of any psychological structure that may be recovered in this latter case, we tend to fall back on the traditional "distinctive features," which (like the earlier linguistic classifications of vowels as "high" or "low" and "front" or "back") derive more from qualitative considerations of how these sounds are articulated by the speaker than from quantitative measurements of how the resulting acoustic wave actually strikes the ear of the listener. Moreover, each such distinctive feature, being merely qualitative, defines not a quantitative spacing of the consonants on some underlying continuum but only a discrete classification into two (or sometimes three) separate groups.

In cases of this type, it is not wholly obvious that the most valid psychological representation of the stimuli will be as points in a continuous space. Perhaps, instead, they should simply be represented as grouped into a certain number of discrete "clusters," each of which contains those stimuli that are most frequently confused with each other. Accordingly, we turn now to a second general type of method that does achieve a clustering representation of this general sort. In fact it divides the stimuli, not only into clusters, but also into subclusters and sub-subclusters, according to that overall hierarchical scheme that best agrees, in a certain sense, with the matrix of symmetrized confusion measures S_{ij} as a whole.

Pictorially, the recovered hierarchical clustering takes a form that is known among combinatorial mathematicians as a "rooted tree." The stimuli, instead of appearing as points in a continuous space, now appear as the discrete, terminal nodes of their associated tree. The inclusion relations among the clusters and subclusters of these stimuli are represented by the way in which the branches of the tree converge from these terminal nodes toward the base or "root."

Interestingly, one method for yielding this sort of representation was proposed over 20 years ago for application to problems of biological taxonomy [5, p. 181; 40]. However, the present exposition and results are based on a more elegant formulation and computer algorithm recently developed at Bell Laboratories by S. C. Johnson [41]. Actually, Johnson's general approach (which is based on the unifying concept of an "ultrametric") subsumes two somewhat different methods: one that constructs clusters that are in a certain well-defined sense optimally "compact," and another that constructs clusters that are in a similar sense optimally "connected" (i.e., free of gaps). Here, however, we shall be concerned with only the former ("compactness") method. So far that method has seemed to provide the most uniformly interpretable results in the analysis of data on confusions among speech sounds.

The type of tree representation recovered by this method is best explained by reference to a specific example. Figure 4.5, then, presents the structure that was recovered when Johnson's program was applied to the matrix of S-values already computed from Miller and Nicely's [33] data and presented in Table 4.1 above. The 16 consonants are listed across the top of the tree. At that level, each stimulus is regarded as constituting its own separate "cluster." But as we move down through the tree, what were previously separate clusters successively join together to form fewer, larger clusters until we finally reach the base of the tree, where all 16 stimuli have finally merged into one grand cluster.

Any horizontal section through this tree defines a particular partitioning of the 16 stimuli into a certain number of mutually exclusive and exhaustive classes. Moreover, the total system of distinct classifications obtained by taking such sections at all possible levels in the tree is necessarily strictly hierarchical in the sense that classifications can always be obtained, at any higher level, merely by splitting and, at any lower level, merely by combining the component classes or "clusters."

Figure 4.5 has been constructed such that the precise height at which any particular cluster splits into two or more branching subclusters (as indicated by the numerical scale at the left of the tree) reflects the internal cohesiveness or "strength" of that particular cluster. Trees constructed with this feature have sometimes been called "dendrograms"

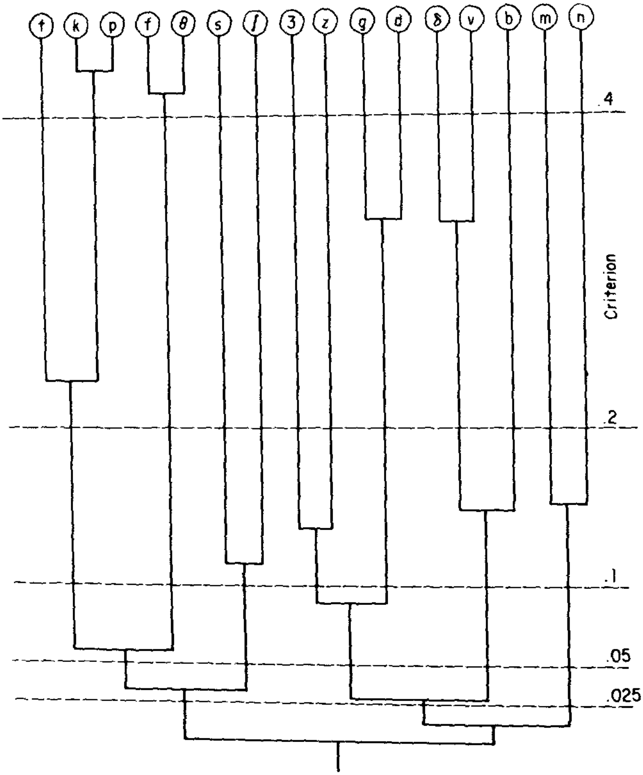

Fig. 4.5. Hierarchical clustering representation for 16 consonant phonemes (based, again, on the pooled data from Miller and Niceley's six "flat" conditions).

[5, 42]. In the present application, the numerical strength of any cluster is always the *smallest* S-value for any pair of stimuli included in that cluster. It thus corresponds, in a sense, to the weakest link holding that cluster together.

The algorithm for finding such a representation is quite simple. At each stage of the process, any two (or more) objects that are connected by the largest remaining value in the S-matrix are simply combined into a new object or "cluster." A new S-matrix is then constructed for the reduced set of objects in which the S-value connecting a new, combined object to any other object is defined to be the *smallest* of the S-values that previously connected the components of this new object to that other, external object. The process is then repeated until (after not more than $n-1$ stages) all n of the original objects have been incorporated into one final object. The resulting overall hierarchical scheme has the property that, for the clustering defined by the horizontal section corresponding to a specified number of clusters, the within-cluster S-values are all kept above the highest possible bound. Since the S-values are inversely related to "ultrametric" distances [41], this amounts to minimizing the "diameters" of the clusters and hence to maximizing "compactness."

The resulting hierarchical clustering (as it is revealed in the topological structure of the tree) is strictly invariant under monotonic transformations of the given S-values. The numerical values of the clusters as reflected in the vertical heights of the branch points, however, do not possess this degree of invariance. Still, we usually have some confidence in more than just the rank order of the given values of S; and, to this extent these S-values can guide us in selecting particular horizontal sections through the tree. Certainly, the clustering associated with a cut (at .34 in Fig. 4.5) that just leaves /g/ and /d/ clustered while just separating /ð/ from /v/ would not be very reliable. The most stable clusterings would be those resulting from a cut (like that shown at .2) that can be moved up and down over an appreciable range without changing the clustering.

For some purposes it is convenient to plot the numerical values associated with different cuts through the tree as a function of the number of resulting clusters, as shown in Fig. 4.6. Here, the more reliable

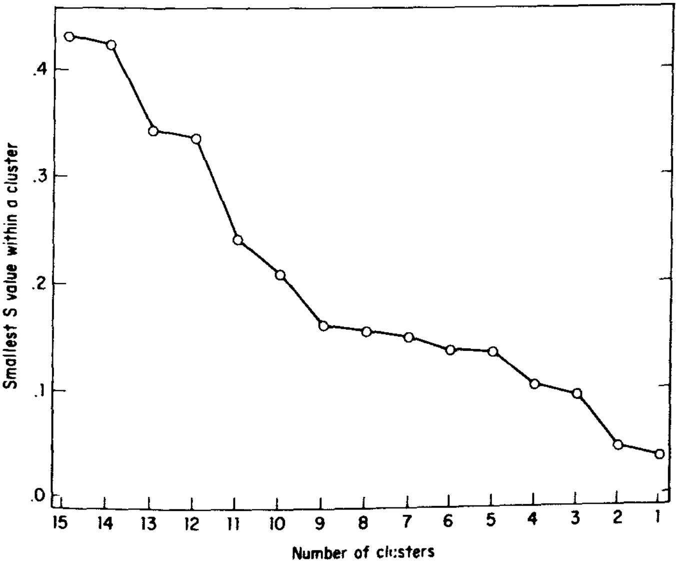

Fig. 4.6. Dependence of the smallest level of confusion within a cluster upon the number of clusters (for the pooled data from the six "flat" conditions).

clusterings correspond to points on this graph immediately followed by an abrupt drop in the curve. Thus we see, for example, that the division of the 16 consonants into 11 clusters is probably more reliable than their division into 10 clusters.

Combined Hierarchical-Spatial Representation and Its Relation to Distinctive Features

Figure 4.5, despite its radically different appearance from Fig. 4.3, actually contains much of the same information. The consonants /p/ and /k/, for example, are first to join in Fig. 4.5 and are also closest together in Fig. 4.3. Next comes the pair /f/ and /θ/ in both figures. Likewise, at a level near the bottom of the tree we find a division into the three major groups already noted in Fig. 4.3; namely, the groups /ptkfθsʃ/, /bdgvðzʒ/, and /mn/, which are distinguished solely on the basis of voicing and nasality.

That this agreement holds up throughout the whole hierarchical structure can be verified by representing that structure on top of the earlier spatial solution itself. This has been done in Fig. 4.7 by drawing closed curves around those points (of the earlier Fig. 4.3) that are grouped together by taking horizontal slices through the tree (in Fig. 4.5). In

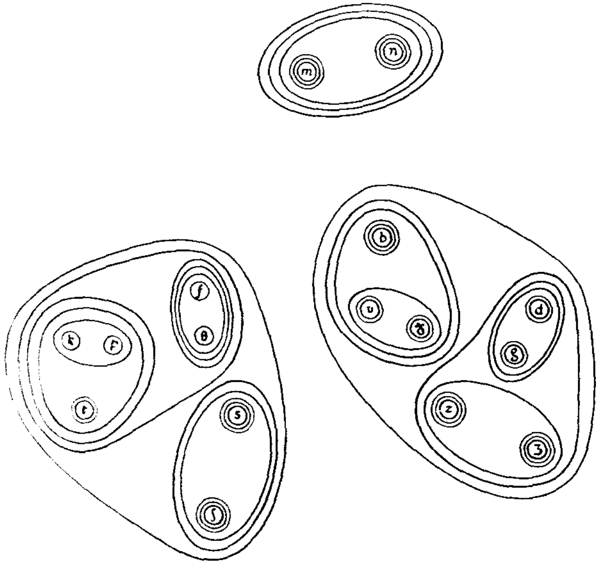

Fig. 4.7. Combined spatial and hierarchical representation (in which the hierarchical clusters of Fig. 4.5 are embedded into the spatial configuration of Fig. 4.1).

order to keep the combined figure from becoming too crowded with these added curves, the horizontal slices have been taken only at certain representative heights; specifically, at the geometrically decreasing levels .4, .2, .1, .05, and .025 (as indicated by the dashed lines in Fig. 4.5).

The lowest level (.025) divides the consonants into just the three major clusters indicated by the three outermost curves. The next level (.05) then subdivides the consonants into the five clusters indicated by the five curves just inside the earlier three. Finally, for the highest level (.4), there are 14 separate clusters indicated by the 14 innermost curves, and only the pairs /pk/ and /fθ/ still remain enclosed together. The overall agreement between the spatial representation and the hierarchical clustering consists in the fact that points that are enclosed together within more curves are generally closer together too. It is this fact, of course, that permits these curves to assume such simple, convex forms.

By embedding the hierarchical clustering scheme within the two-dimensional spatial solution in this way, we obtain a picture that seems particularly revealing of the underlying structure of the original confusion data. Certainly Fig. 4.7 is more immediately informative than the matrix of numerical data with which we began (Table 4.1). Moreover, the two types of representations that have been combined in Fig. 4.7 are to some extent complementary. The spatial representation contains some information that is not preserved in the hierarchical clustering, e.g., that within the cluster /bvð/ it is /v/ and /ð/ that are closest to the nearby cluster /zʒ/. The clustering, on the other hand, may be more readily related to other discrete classifications such as those based upon the traditional distinctive features. (Then, too, the clustering solution can serve as a partial confirmation that the spatial solution corresponds to the absolute optimum, not just to one of several merely local optima.)

An important feature of Fig. 4.7 is that it was obtained solely on the basis of the empirical data (Table 4.1). We have not, that is, provided any opportunity for theoretical preconceptions about the structure or grouping of these 16 consonants to insinuate themselves into the final picture. This approach is, in this sense, purely psychological. It differs from more *psychophysical* approaches, in which one starts with a definite preconception as to what variables are relevant and then tries to quantify the relation of the data to just these variables.

Miller and Nicely [33], in particular, began with a system of five distinctive features (voicing, nasality, affrication, duration, and place of articulation), and then analyzed their data specifically with respect to these five variables. This approach did enable them to assess the relative importance of each of these five variables in accounting for the confusion data. But as we shall see, it did not provide a very sensitive test of possible departures of their data from the structural requirements of their distinctive-feature system as a whole.

Figure 4.8 is a pictorial representation of the particular system of distinctive features considered by Miller and Nicely [33]. The three orthogonal dimensions of either the upper or the lower rectangular box are used to distinguish the consonants with respect to the three features of voicing, affrication, and place of articulation. The further feature of nasality, then, is encoded in the separation of the upper from the lower box. The final feature, duration, is not explicitly represented in the picture. It simply distinguishes the four "longer" fricatives /sʃzʒ/, at the lower front corners of the box, from the other 12 consonants. (Without this last feature there would, of course, be no basis for distinguishing /θ/ from /s/ or /ð/ from /z/.)

This particular system of distinctive features is just one of several that have been proposed for the consonants. Halle [43], for example, has advanced a system of eight features that are all purely dichotomous, while Wickelgren [44] has advocated a system of only four features that, however, can take on as many as four values on one feature (place). Nevertheless, these three systems are essentially alike in the properties that will be of primary interest here. They all distinguish the consonants in exactly the same way with respect to voicing and nasality and share the further important property of a complete structural parallelism between the unvoiced consonants /ptkfθsʃ/ and their corresponding voiced analogs /bdgvðzʒ/. That is, for each consonant

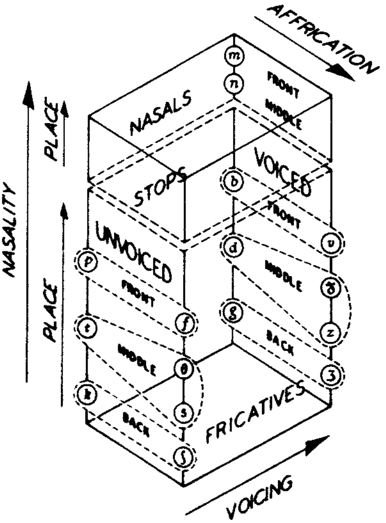

Fig. 4.8. Representation of the 16 consonants in terms of five distinctive features.

in the unvoiced group, the corresponding consonant in the voiced group has exactly the same value with respect to all features except (of course) voicing. In Fig. 4.8 this shows up in the isomorphism between the arrangement of the unvoiced and voiced consonants on their two opposing faces of the lower box.

Let us turn then to a comparison of these common properties of the various distinctive-feature systems with the independently determined empirical structure exhibited in Fig. 4.7. There is, first of all, the important agreement (already noted) between the unanimous classification of these consonants on the basis of the distinctive features of voicing and nasality and the clear-cut division of these consonants into the three major groups shown in Fig. 4.7. On the other hand, the unanimously postulated parallelism between the voiced consonants and their corresponding unvoiced consonants does *not* hold up in this figure. Instead, the unvoiced consonants tend to split into three rather strong subgroups /ptk/, /fθ/, and /sʃ/, whereas the corresponding voiced consonants tend to split, quite differently, into a strong subgroup /bvð/ and a separate group consisting of the two weakly linked subgroups /dg/ and /zʒ/. None of the proposed distinctive-feature systems appear to provide any basis for understanding why /b/ should group so strongly with /v/ and /ð/, while /p/ groups equally strongly—not with /f/ and /θ/—but with /t/ and /k/.

Of course, no evidence has yet been adduced to show that the empirical groupings exhibited in Fig. 4.7 are statistically reliable. In the ensuing sections, however, we shall find rather compelling evidence of their reality in the fact that these same groupings emerge again and again when the individual and independently collected confusion matrices reported by Miller and Nicely are analyzed separately.

We seem to have, then, an empirical basis for constructing a classificatory scheme for the consonants. In addition to its firmer foundation in data, such a scheme would come closer to representing these sounds as they are actually heard by a listener (rather than as they are thought to be articulated by a speaker). Presumably, such a listener-oriented scheme would be more useful for attempts to build automatic devices for the recognition of speech or for the compression and transmission of speech.

Effect of Variations in Signal-to-Noise Ratio

Figure 4.7 was based on the pooled data for the first six of Miller and Nicely's 17 different conditions. Since these first conditions differed only with respect to signal-to-noise ratio, it is of some interest to analyze the data for each condition separately and, thus, to determine the effect

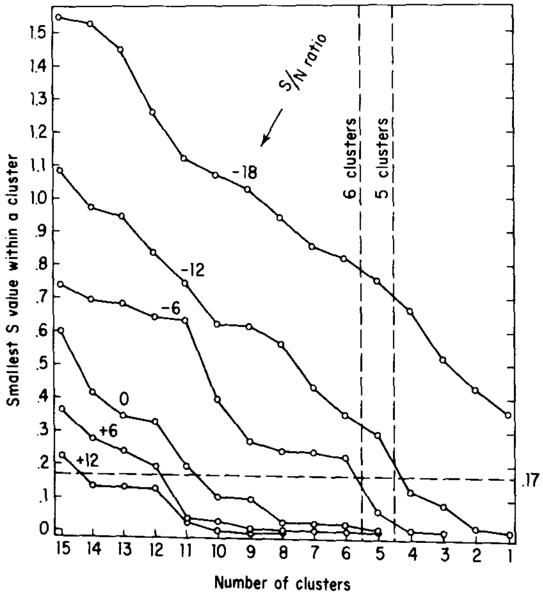

Fig. 4.9. Dependence of the smallest level of confusion within a cluster upon the number of clusters, plotted separately for each of the six "flat" conditions. (*Miller and Niceley, tables I–VI.*)

of that variable on the pattern of the resulting confusions. As a first step toward this end, Johnson's clustering algorithm was applied to a matrix of S-values (like the present Table 4.1) computed separately for each of Miller and Nicely's Tables I through VI.

Figure 4.9 includes, for each of the six resulting clustering solutions, a curve of the type previously introduced in Fig. 4.6. The curves show how the level of confusion within clusters (the smallest within-cluster S-value) decreases as the clusters are merged into fewer and fewer clusters. The signed number over each curve indicates the *S*/*N* ratio of the condition to which that curve applies. Naturally, lower *S*/*N* ratios lead to higher frequencies of confusion and hence to higher curves.

Now an informative way of looking at the clustering results is to specify some fixed level of confusion and then look at the clustering defined by the slice through each of the six trees at that specified level. For this purpose, the S-value of .17 (indicated by the horizontal dashed line in Fig. 4.9) seemed to establish a level with two desirable properties: (1) it cuts through as many of the six curves as possible, and yet (2) it cuts through each at a point of relatively abrupt decline in that curve. We can see from the figure that, at the *S*/*N* ratio of —18 dB, the curve remains above a confusion level of .17 even when we get down to a single cluster; evidently all 16 consonants are confused with each other at this level. Under the optimum condition of +12 dB, at the other extreme, we find that there are 15 different clusters that can be distinguished from each other at this level. (Only /f/ and /θ/ are still confused with an *S*-value of .17.)

The specified level of .17 thus defines six distinct clusterings—one for each of the six *S*/*N* conditions. Given a spatial representation of the consonants as 16 points in a plane, then, we can embed any or all of these six clusterings in this plane in the manner previously illustrated in Fig. 4.5. Perhaps the most reasonable spatial representation to use for this purpose is the one that was based on the combined data for all six conditions; that is, the one already presented in Figs. 4.1 and 4.5. This has been done in Fig. 4.10 where the signed number associated with a closed curve designates the *S*/*N* ratio of the condition upon which that curve is based. Thus for —18 dB, all 16 consonants form one confused group, as we already noted. For —12 dB, then, this group divides into five different groups that can be discriminated from each other at the specified level (.17), and so on for the higher *S*/*N* ratios.

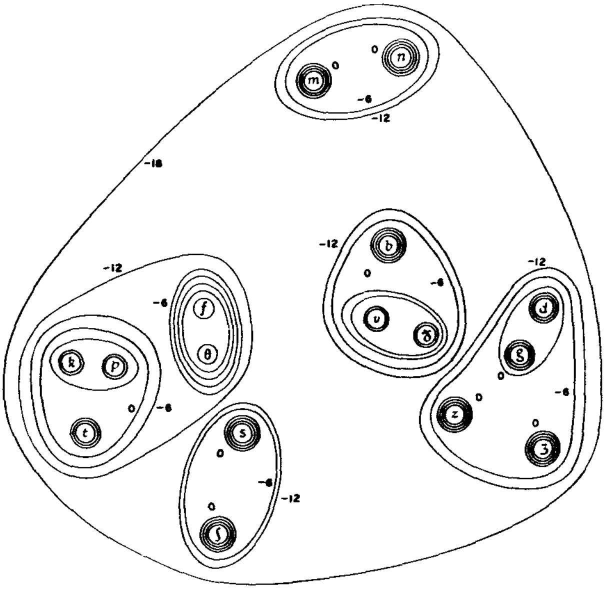

Fig. 4.10. Representation of the effect of S/N ratio on confusion among the 16 consonants. For each of the six *S*/*N* conditions, a closed contour has been drawn around those points in Fig. 4.1 that represent consonants that were confused together at the criterion level of .17.

This combined picture reveals a gratifying degree of consistency in Miller and Nicely's data, for in this figure the different curves (i.e., curves for different S/N ratios) are based upon independent sets of data. One consequence is that Fig. 4.10—unlike the earlier Fig. 4.7—is not necessarily constrained to a hierarchical pattern. It could, for example, happen that the cluster /fθ/ would group with the cluster /ptk/ at −12 dB, but would group with the cluster /sʃ/ at −6 dB. If this happened, Fig. 4.10 could not be constructed without one curve actually cutting across another. It is the remarkable consistency of the data obtained under the different S/N ratios that permits the curves to assume this strictly nested or hierarchical structure.

This in turn permits us to think of these curves as level curves like the elevation contours on a topographical map. Concretely, we might think of the 16 consonants as situated at the lowest points of 16 depressions in an uneven terrain. Added noise, then, might picturesquely be likened to muddy water welling up in these depressions. As we infuse more and more noise, more and more of the previously separate puddles lose their identity by merging into fewer, larger pools. At −12 dB just five separate "pools" (i.e., distinguishable sounds) are left; and at −18 dB everything has finally run together, so to speak, in one big muddy confusion.

An even more striking demonstration of the consistency of the data is afforded by taking vertical rather than horizontal cuts through the six curves in Fig. 4.9. Whereas a horizontal cut corresponds to a fixed level of confusion (or S-value), a vertical cut corresponds to a fixed number of clusters. We can, that is, move down each hierarchical tree until we find the consonants grouped into some prespecified number of clusters and then see just what those clusters are. In the present case, cuts of this sort were taken at both five and six clusters, as indicated by the two vertical dashed lines in Fig. 4.9. We shall not consider the two most extreme conditions in any detail; at +12 dB there are insufficient confusions to define a grouping into fewer than eight clusters, and at −18 dB the data are to nearly random to define *any* very stable groupings. For all four of the intermediate conditions (−12, −6, 0, +6 dB), however, the data are completely consistent. In each case the grouping into five clusters leads to exactly the *same* five clusters. These are indicated by the dashed curves in Fig. 4.11. Moreover, in all four cases, the change from five to six clusters comes about by a split of the same cluster, /ptkfθ/, into the same two subclusters, /ptk/ and /fθ/, as indicated by the solid curves in the figure. Evidently the patterns to which attention was called in the earlier Fig. 4.7 are indeed reliable.

Indeed, although the +12-dB condition did not lead to enough confusions to define fewer than eight clusters, the eight-cluster solution for that extreme condition is completely consistent with the six-cluster solutions exhibited in Fig. 4.11 and, in fact, can be derived from that six-cluster pattern merely by splitting /ʃ/ off from the cluster /sʃ/ and by splitting /ʒ/ off from the cluster /dgzʒ/. Hence, /bvð/ emerges once again as a coherent grouping. Even in the almost random data of the very noisiest, —18-dB condition, /b/ was more often confused with /v/ than with any other phoneme. This initially unexpected grouping of the voiced stop /b/ with certain voiced fricatives rather than with the other voiced stops has thus recurred, now, in each of Miller and Nicely's first six independent sets of data.

The remarkable regularity summarized in Fig. 4.11 has an interesting consequence. It suggests that, although *S/N* is a powerful determiner of overall level of confusion (as indicated in the overall heights of the curves in Fig. 4.9), it has little or no effect on the internal pattern of confusion, for over a range from —12 to +6 dB we find that the consonants that are most confused together constitute exactly the same groups; and this is true even though the absolute level of confusion associated with these groups changes over a 50-fold range from .352

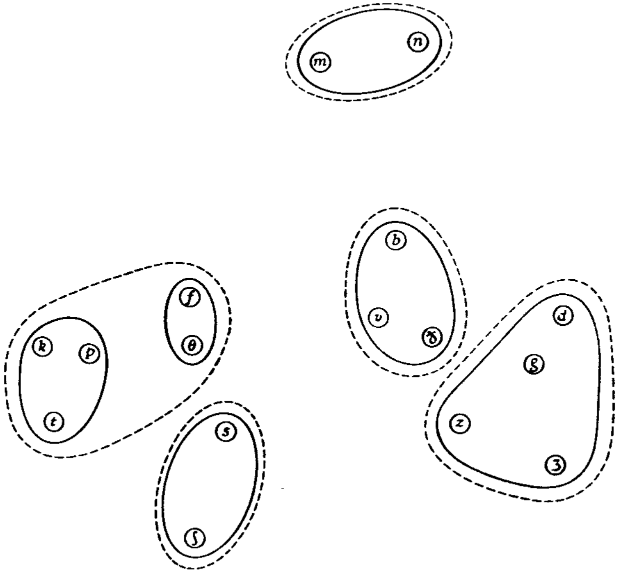

Fig. 4.11. Representation of the invariance of the pattern of confusion over different S/N ratios. The broken contours represent the five-cluster solutions and the solid contours represent the six-cluster solutions for all but the most extreme *S/N* conditions.

(for six groups at −12 dB) to .006 (for the same six groups at +6 dB). The strictly nested pattern of the contours in the earlier Fig. 4.10 (for all six *S/N* conditions) can of course be adduced, too, as evidence for such an invariance under wide shifts in *S/N* ratio.

This same invariance can also be demonstrated by obtaining a spatial solution like that already shown in Fig. 4.1—but for each of the six *S/N* conditions separately. For this purpose the assumed exponential relation between probability of confusion and interpoint distance has the nice property that the "slope" parameter *b* enters into a purely symmetric, multiplicative relation with the distances D_{ij}. Thus we are free to fix *b* at any arbitrary value (e.g., unity), and differences in overall level of confusion should then be accommodated merely by changes in the overall size of the resulting spatial configuration.

Essentialy, this is what in fact happened when spatial solutions were separately obtained for each of the individual matrices. For the worst (−18-dB) condition, all 16 points were crowded together in one tight little clump. With improvements in *S/N*, however, each consonant became perceptually more distinct or "distant" from all others, and so the 16 points spread further and further apart. During this overall expansion of scale, moreover, the relational structure (as defined by the *relative* distances among the 16 points) remained quite stable—at least until +6 or +12 dB. (At these highest *S/N*, the spatial solution tended to become somewhat indeterminate owing to the small number of nonzero entries in the resulting confusion matrices.)

In order to evaluate this conjectured invariance throughout the entire range of *S/N* ratios, we evidently need more stable estimates of the frequencies of confusion—particularly for the most dissimilar pairs at high *S/N*. Toward this end, the 120 pairs of consonants were partitioned into four groups on the basis of the *S*-measure previously computed for the average data for Miller and Nicely's Tables I through VI. The high-similarity group contained the four pairs (/pk/, /fθ/, /dg/, and /vð/) of highest average similarity ($.250 < S \leq .500$); the medium-high group contained the nine pairs of next highest similarity ($.125 < S \leq .250$); the medium-low group contained the twelve pairs of next highest similarity ($.075 < S \leq .125$); and, finally, the low group contained the remaining 95 pairs of lowest similarity ($.000 < S \leq .075$).

Figure 4.12 shows the mean *S*-value for each of these four groups as computed for each of the six *S/N* conditions separately. As is to be expected, the mean psychological similarity declines as *S/N* improves for all pairs (whether of high or low average similarity). However, the shapes of the curves for the four levels of average similarity are quite different, and so the invariance we are looking for is not very evident in the untransformed *S*-measures of confusion themselves.

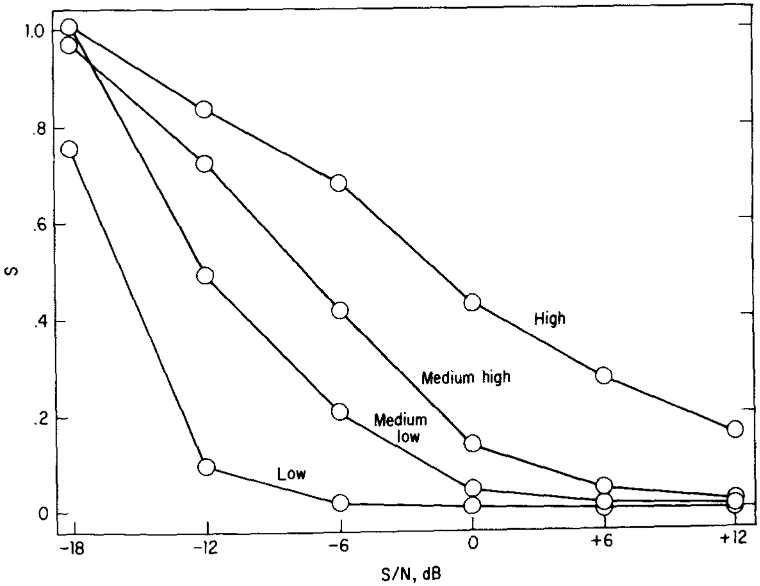

Fig. 4.12. Mean level of confusion (S) as a function of S/N ratio, for pairs of consonants of low, medium-low, medium-high, and high degrees of similarity.

The spatial model, however, suggests that the invariance should emerge in clearer form when the similarities are converted into distances. According to the exponential assumption, moreover, this is to be accomplished by applying a logarithmic transformation to the empirical S-values. In particular, since the asymptote c of the exponential will generally be greater than zero [30], the desired distance estimates should be computed as $D = -\log (S - c)$. In the present case, if we take $c = .0003$, we find the kind of invariance we are looking for. Approximately, the distances for each S/N condition then differ from the distances for any other S/N condition merely by a constant factor. Indeed, as shown in Fig. 4.13, we can then find a slightly adjusted spacing of the six S/N conditions such that, after the logarithmic transformation, the four curves which in Fig. 4.12 were highly nonlinear now radiate nearly linearly from an approximately common point (perhaps somewhat to the left of —18 dB).

We seem to have, then, a possible way of separating the intrinsic structure of the 16 consonants, which is reflected in the relational pattern of the points, and the extrinsic effect of added noise, which is primarily reflected in a compression of the overall scale of this pattern. The possibility of such a separation undoubtedly depends upon the fact that the noise added by Miller and Nicely was spectrally flat or "white." By

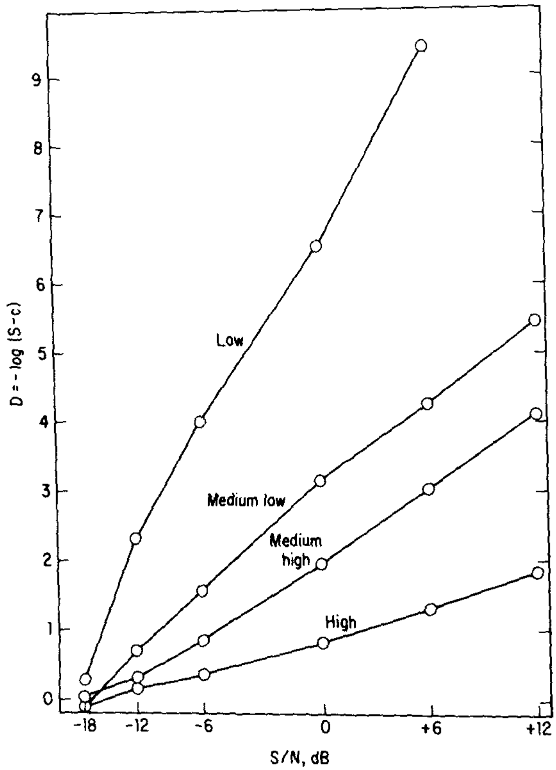

Fig. 4.13. The data of Fig. 4.12 logarithmically transformed to reveal the invariance of structure (proportionality of interconsonantal distance) over different S/N ratios.

choosing suitably band-limited noises one presumably could mask various cues or features of these sounds *differentially* [2] and thus induce nonscalar deformations of the spatial configuration.

Effects of Variations in Low-pass Filtering

The next six of the conditions reported by Miller and Nicely [33] were designed to clarify the effects of filtering out the higher frequencies. The signal-to-noise ratio (before filtering) was fixed at +12 dB, and a high-pass cutoff was maintained at 200 Hz. The low-pass cutoff, however, was varied and, in particular, took on values of 300, 400, 600, 1,200, 2,500, or 5,000 Hz (in Miller and Nicely's Tables VII through XII, in that order).

Figure 4.14 is constructed in the same manner as the earlier Fig. 4.9, but this time the number attached to each curve indicates the setting of the low-pass cutoff. As is to be expected, the overall level of confusion is higher for conditions in which more of the high frequencies have been filtered out. In order to see just how this filtering affects the pattern of confusion, a cut has again been made through these six curves at the same S-value (.17) used in Figs. 4.9 and 4.10. Fortunately, except in the case of the highest of the six curves, this cut (which is indicated by the horizontal dashed line in Fig. 4.14) again intersects each curve at a point of comparatively sharp drop.

Again as in Fig. 4.10, the spatial solution originally obtained on the basis of the first six matrices combined has been used as a framework on which to display the six clusterings that result from this specified level of confusion. The result is presented in Fig. 4.15. The picture is slightly different from that obtained before for variations only in S/N (Fig. 4.10) but, again, we find a perfect hierarchical nesting of the contours. Under the worst of these six conditions (low-pass cutoff set at only 300 Hz), only about four kinds of sounds can be discriminated at the .17 level of confusion. Under the best condition (cutoff raised to 500 Hz), as many as 14 consonants can be discriminated (and confusion at this level is now confined to just the two pairs /fθ/ and /vf/).

The fact that none of these curves are forced to cross again points to the consistency of the data. Even so, these data are not quite as homogeneous as the data from the first six conditions. In particular, there are indications of a qualitative change in the pattern of confusions

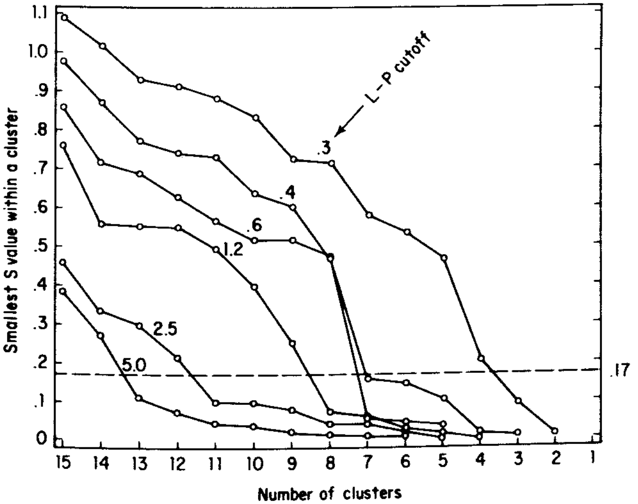

Fig. 4.14. Dependence of the smallest level of confusion within a cluster upon the number of clusters, plotted separately for each of the six low-pass conditions. (*Miller and Niceley, tables VII–XII.*)

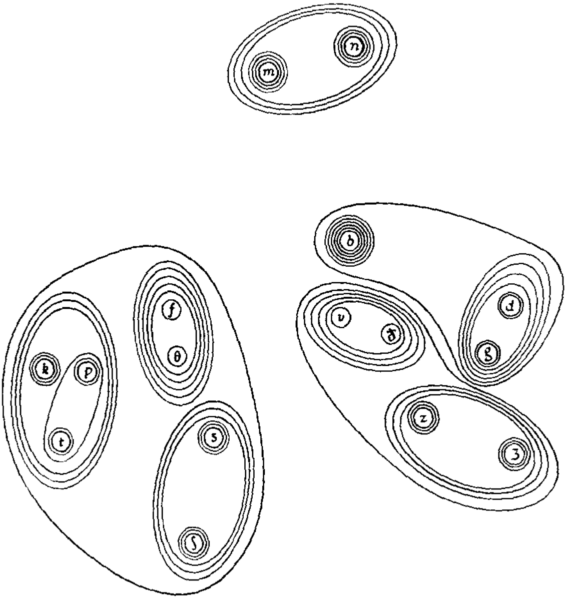

Fig. 4.15. Representation of the effect of differential filtering of high frequencies on confusions among the 16 consonants. For each of the six low-pass conditions, a closed contour has been drawn around those consonants that were confused together at the criterion level of .17.

for the condition of severest filtering—in which the low-pass cutoff is reduced to only 300 Hz. The three- and five-cluster representations for just this one condition (Miller and Nicely's Table VII) are exhibited in Fig. 4.16. (These two clustering representations both correspond to relatively sharp drops in S-value, as can be seen from the highest curve in Fig. 4.14.)

This pattern, which resulted when only the very lowest frequencies were transmitted, departs systematically from the pattern consistently obtained under all conditions of relatively "flat" transmission of frequencies (Fig. 4.11). Indeed, this lowest-frequency condition is the only one of Miller and Nicely's 17 conditions that yields a pattern in close agreement with the distinctive-feature schemes discussed earlier. As indicated by the labels added to Fig. 4.16, the clusters defined by the confusions under this one condition are precisely what we should expect on the basis of three of Miller and Nicely's distinctive features, viz., voicing, nasality, and affrication.

Of course, the spatial configuration (which was based on the six "flat" conditions) is not quite appropriate for the highly filtered condition. (This is why the clusters in Fig. 4.16 tend to be long and narrow.) If a new spatial solution is obtained for just this condition, the clusters assume the more compact form shown in Fig. 4.17. The fact that the consonants are relatively undiscriminated within each of these five inner clusters indicates that the remaining distinctive features (viz., place and duration) tend not to be preserved when all but the lowest frequencies are filtered out.

Even the remaining five of the low-pass conditions show some evidence of qualitative changes in pattern. In any event, vertical cuts (corresponding to fixed numbers of clusters) do not yield precisely the same results for all conditions (as they did for the "flat" conditions in Fig. 4.11). Still, certain of the clusters do recur in most or all of the low-pass conditions. The most ubiquitous of these are shown in Fig. 4.18. The numbers outside each contour indicate which of Miller and Nicely's conditions (VII through XII) yield that particular cluster. Thus, for example, all five of these six conditions except VII led to the cluster /bvð/, which has already been put forward as something of a puzzle for traditional distinctive-feature schemes.

Generally, the pattern resulting from low-pass filtering is remarkably like the pattern resulting from the addition of broadband noise (Fig. 4.11). (Indeed the only notable difference seems to be that /f/ and

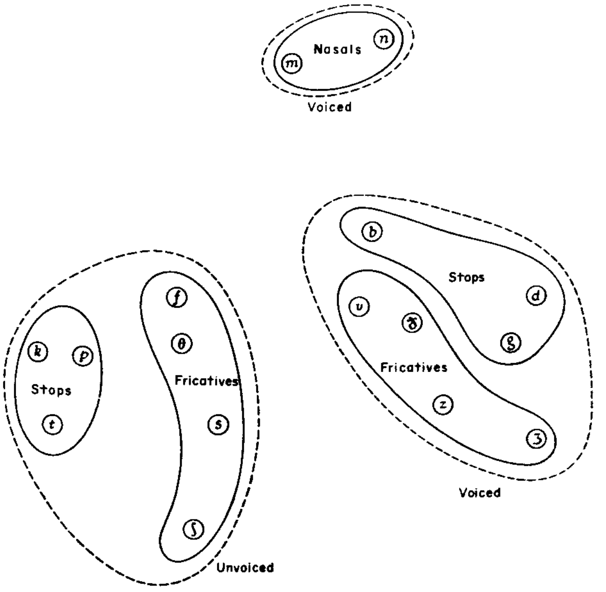

Fig. 4.16. The three-cluster and five-cluster representations for the condition in which only the lowest frequencies were passed (embedded, again, in the spatial configuration of Fig. 4.1).

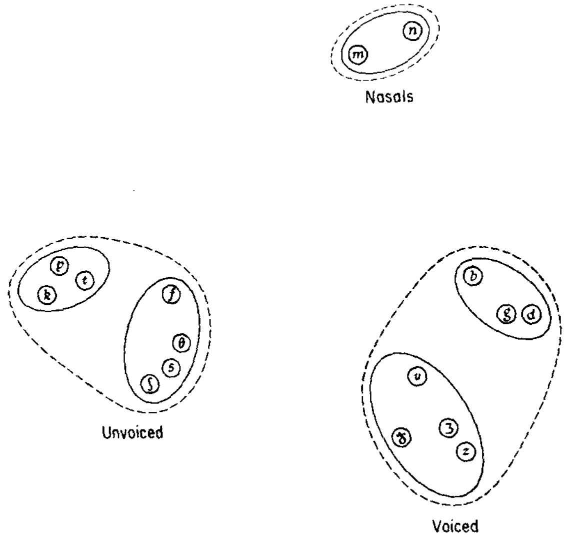

Fig. 4.17. The three-cluster and five-cluster representations of Fig. 4.16 reembedded in a new two-dimensional spatial representation based just upon the data from the condition in which only the lowest frequencies were passed.

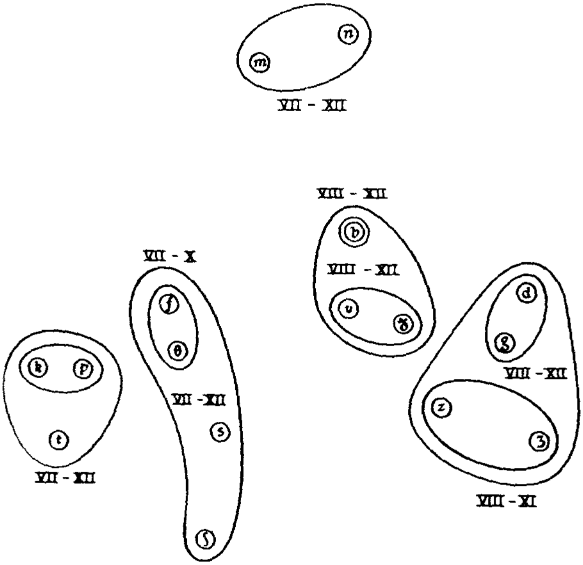

Fig. 4.18. Recurrent clusterings in the low-pass conditions (embedded, again, in the configuration of Fig. 4.1).

/θ/ group with the unvoiced stops /ptk/ in the "flat" conditions but with the other unvoiced fricatives /sʃ/ in the low-pass conditions.) Miller and Nicely also noticed the overall resemblance in the two patterns of confusion and attributed it to the relatively greater susceptibility of the high frequencies to masking by white noise [33, p. 350]. Such a differential effect of white noise needs, however, to be reconciled with the remarkable invariance in pattern already noted with changes in S/N.

Effects of Variations in High-pass Filtering

The five remaining conditions studied by Miller and Nicely are concerned with the effects of filtering out just the lower frequencies. Again the signal-to-noise ratio (before filtering) was fixed at +12 dB. But this time a low-pass cutoff was maintained at 5,000 Hz, while a high-pass cutoff took on values of 1,000, 2,000, 2,500, 3,000, and 4,500 Hz (in Miller and Nicely's Tables XIII through XVII, in that order).

Figure 4.19 shows the five resulting curves of the type displayed before (in Figs. 4.9 and 4.14). As expected, the overall level of confusion generally increases with the frequency of the high-pass cutoff. A horizontal cut has also been made through these curves at the S-value selected before (viz., .17). Conveniently, this again intersects most of the curves at a point of relatively steep decline.

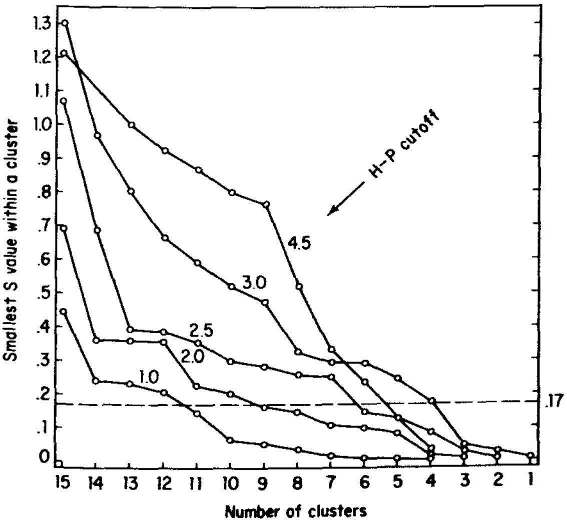

Fig. 4.19. Dependence of the smallest level of confusion within a cluster upon the number of clusters, plotted separately for each of the five high-pass conditions. (*Miller and Niceley, tables XIII–XVII.*)

In Fig. 4.20, as before, the spatial solution that was originally based upon the pooled data for the first six conditions has been used as a framework on which to display the clusterings corresponding to the .17 cut through Fig. 4.19. The contours for the least severely filtered condition (Miller and Nicely's Table XIII) are not included in the figure. Generally, these omitted contours would fall inside the others. However, as might be expected, the pattern for this nearly "flat" condition departs somewhat from the pattern for the remaining, more severely filtered conditions. At the .17 level of S, this discrepancy is confined entirely to the occurrence of the two clusters /fθ/ and /ðz/ in this relatively unfiltered condition.

With the exception of this one condition, however, Fig. 4.20 shows that the contours for the high-pass conditions form a hierarchically nested set. The pattern that thus consistently emerges for these high-pass conditions nevertheless differs radically from the patterns noted earlier for the flat conditions (Fig. 4.10) and the low-pass conditions (Fig. 4.15). This is further indicated by the most recurrent clusters, which are represented in Fig. 4.21 in the same manner as in the earlier Fig. 4.18.

Regarding these high-pass conditions, Miller and Nicely remark that

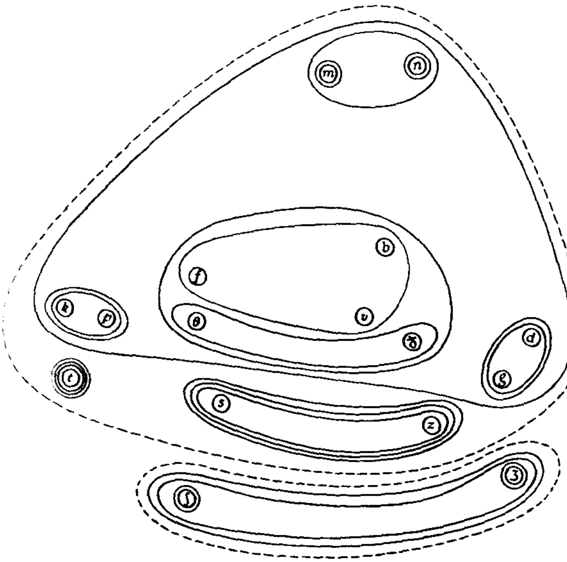

Fig. 4.20. Representation of the effect of differential filtering of low frequencies on confusions among 16 consonants. For each of the five high-pass conditions, a closed contour has been drawn around those consonants that were confused at the criterion level of .17.

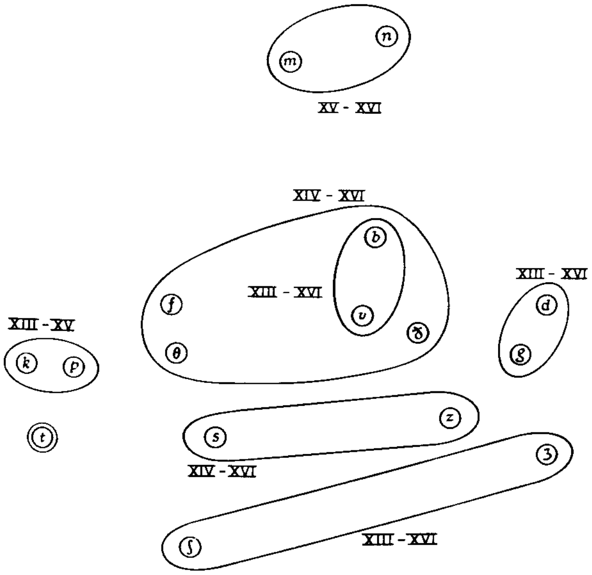

Fig. 4.21. Recurrent clusterings in the high-pass conditions.

"the errors do not cluster or fall into obvious patterns in the confusion matrix, but seem to distribute almost randomly over the matrix" [33, p. 350]. However, the present Figs. 4.20 and 4.21 do seem to reveal some definite structure in these data. Roughly speaking, we might say that, under the high-pass conditions, it is those consonants that are separated horizontally in the figure that are most confused (e.g., /s/ and /z/ or /ʃ/ and /ʒ/ in Fig. 4.21). Under the most extreme low-pass condition, on the other hand, the prevalent confusions tended to run in a roughly vertical direction (e.g., between /f/ and /ʃ/ or between /v/ and /ʒ/ in Fig. 4.16). Thus we are led to the crude interpretative conjecture that the vertical and horizontal dimensions of this space have, in part, to do with discriminations among high frequencies and among low frequencies, respectively.

Now voicing has been largely identified with the laryngeal injection of low-frequency energy. So it is not surprising that, after the low frequencies have been filtered out, the distinction of voicing is all but lost for the fricatives (as shown by the strong pairings /θð/, /sz/, and /ʃʒ/ in Fig. 4.20). Note, however, that this distinction of voicing is relatively much less dependent upon the presence of low frequencies in the case of the stops. Here, then, we seem to find a further departure from the kind of symmetry or parallelism that one tends to expect on the basis of the traditional systems of distinctive features. Moreover, we see (in Fig. 4.21) that whereas the unvoiced stops tend, as before, to group together (particularly /p/ and /k/), the voiced stop /b/ persists in its closer association with the voiced fricatives (particularly /v/).

Actually, of course, the difficulty in accounting for these aspects of the patterns in terms of distinctive features arises only to the extent that we suppose, with Miller and Niceley [33, p. 348], that "the features were perceived almost independently of one another" and, also, that just these five features were operative. The absence of parallelism in the structure of the voiced and unvoiced consonants (Figs. 4.11, 4.18, and 4.21) could, for example, be explained by assuming, instead, that whether the fricative-*vs.*-stop distinction or the place and duration distinctions take precedence critically depends upon whether the consonants are unvoiced or voiced, respectively (cf. Fig. 4.8). Likewise, the pattern in Figs. 4.20 and 4.21 could be explained by supposing simply that the salience of the voicing distinction depends strongly upon whether the consonants are stops or fricatives. The attractiveness of an account of the confusion data in terms of distinctive features is, however, somewhat reduced by the necessity of invoking such interactive complications. Perhaps the simplest and most plausible way of explaining both of these seeming anomalies in the confusion data is the one suggested by Savin [45], in which we invoke—instead of interactions among these five distinctive features—an additional, *sixth* distinctive feature, aspiration, that arises only in the case of the three syllables /pa/, /ta/, and /ka/. Such a feature would certainly account for the persistent grouping of the unvoiced stops /ptk/ even when there is no parallel grouping of the voiced stops /bdg/. At the same time, since aspiration is represented in the higher frequencies, it would account for the tendency of the unvoiced stops /ptk/ to remain distinct from the voiced stops /bdg/ even when the unvoiced fricatives /fθsʃ/ are no longer discriminated from their corresponding voiced fricatives /vðzʒ/.

In any case, the point should now be clear that it is potentially limiting to conduct one's analysis solely within the framework of any one particular a priori framework of distinctive features. Sometimes it may prove illuminating to take, as an alternative starting point, a natural grouping of the consonants (e.g., as in Fig. 4.11) determined a posteriori, purely by analysis of the empirical data.

Judged Similarities among Consonants and Their Relation to Confusion Data

So far here, we have exclusively considered, as a measure of psychological similarity, the frequency with which phonemes are actually confused with each other by human listeners. It is of course also possible to ask listeners to attempt direct subjective estimates of the similarities of pairs of stimuli. Indeed, such an experiment has been done by Peters [19] using, as stimuli, the same 16 consonants studied by Miller and Nicely. Accordingly, some comparisons can now be made between the patterns we have noted in the confusion data reported by Miller and Nicely and whatever pattern may emerge from the judgmental data subsequently collected by Peters.

In the experiment of interest here, Peters had each subject pronounce each pair of consonants aloud and then rate the pair on a nine-point scale ranging from 1 for "extreme similarity" to 9 for "extreme dissimilarity." Each consonant was pronounced as followed by the vowel /ʌ/. Thus the subject would pronounce one of the 120 possible pairs, say /pʌ-tʌ/, and then write down a number from 1 to 9 to indicate the apparent similarity (or dissimilarity) of the two consonants (e.g., /p/ and /t/). It may be a significant aspect of this procedure that the subject can respond to the consonants not merely as a listener (as in Miller and Nicely's experiments) but as a speaker as well.

For purposes of overall comparison, Johnson's [41] clustering algorithm was applied to the matrix of similarity estimates obtained simply be averaging over all nine of Peters' subjects.* Also, the spatial solution originally obtained on the basis of Miller and Nicely's "flat" conditions (Fig. 4.1) is again used to display the resulting clusters. The most salient of these have been included in Fig. 4.22. In this case many widely separated points (like /t/ and /d/) are grouped together as pairs. In order to simplify the picture such high-similarity pairs have simply been connected by a single heavy line (rather than enclosed in a surrounding curve, as before).

At the level of four clusters, we find the consonants neatly grouped into the stops /p,t,k,b,d,g/, the sibilants /s,ʃ,z,ʒ/, the remaining fricatives /f,θ,v,ð/, and the two nasals /m,n/. Then, at the level of eight clusters, we find that every voiced consonant is simply paired with its unvoiced counterpart, except for the two voiced nasals which (as in all earlier solutions) remain grouped with each other. As the greatly elongated representations of these groupings attest, the pattern here bears little resemblance to those found in the confusion data (except, perhaps, for the conditions of high-pass filtering, which did lead to such a horizontal elongation in the case of the fricatives—but not the stops).

This marked disparity between the similarity and confusion data is of some interest—particularly since earlier comparisons between the two types of data with other types of stimuli have not revealed any such

* Peters' published report does not include the matrices of data upon which his analyses were based. I am grateful to him for generously making these available to me and thus permitting me to carry out the further analysis reported here.

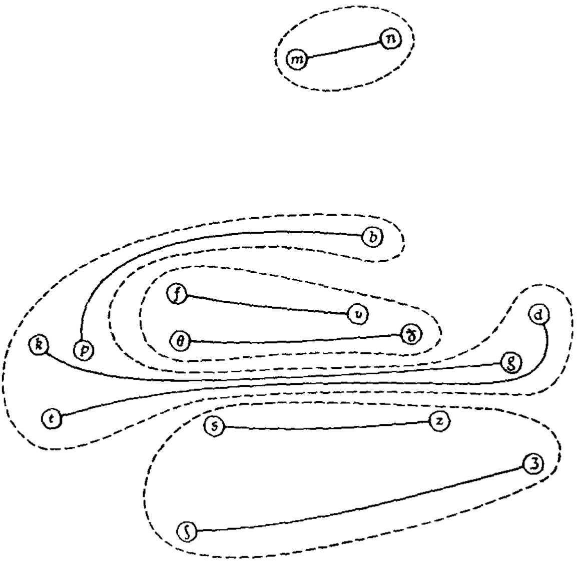

Fig. 4.22. Clusterings according to Peters' data on judged similarity of consonant phonemes.

discrepancy (e.g., Ref. 12, p. 515). In the present instance, the discrepancy may in part be due to the difference, already noted, between responding purely as listener and responding as speaker. However, a somewhat different, more cognitive hypothesis also suggests itself.

First, however, with regard to distinctive features, Peters [19, p. 1989] concluded from his analysis of these data, "In general phonemes are first sorted according to manner; voicing is next in importance, and place, for some individuals, is also important." The simplest interpretation of the very strong pairings evident in Fig. 4.22, however, seems to be that the feature of voicing was, more than any other feature, suppressed or ignored by these subjects. This is curious in view of the fact that the confusion data clearly show voicing (and, probably, nasality) to be the most easily discriminated of the distinctive features (Fig. 4.1).

One hypothesis that might explain this apparent suppression of the usually salient feature of voicing in the judgmental task is that Peters' subjects treated this task as an analogy task rather than a pure similarity task. As an illustration of the kind of possibility being suggested here, consider the 12 visual stimuli displayed in Fig. 4.23. Now the difference between the black and the white figures shown is probably at least as salient or discriminable as the differences among the various shapes. Yet subjects who are asked to judge the similarities among these figures might be most impressed by the parallelism between the structure of the set of black figures and the structure of the set of white figures. The fact, that is, that each black figure has a perfect analog among the white figures might lead them to discount the obvious difference in color between the two parallel sets and simply report that the black triangle is most like the white triangle, and so on for the other shapes.

There is, of course, no certainty that Peters' subjects treated voicing in the way the hypothetical subjects were just supposed to have treated color. Still, the parallelism in the distinctive-feature structure of the voiced and unvoiced consonants was probably apparent to Peters' subjects. At least half of his subjects had had some training in phonetics and, since the distinctive-feature schemes are based primarily upon the way the consonants are articulated, the fact that the subjects pronounced the phonemes themselves might tend to focus their attention on the distinctive-feature structure of these phonemes.

In any case, the disparity between the patterns observed in the two kinds of data argues for considerable caution in generalizing from subjective judgments of similarity to the confusions actually made by listeners. Moreover, the speculations as to the possible reason for the observed disparity suggests that similarity judgments may be quite sensitive to the particular types of instructions and training given to the subjects. The primary advantage of subjective judgments of similarity, in the present connection, is that they can be collected more rapidly than comparably stable confusion data. But they also have the disadvantage that their relation to the basic processes underlying the identification of speech sounds is, as we have seen, more tenuous and uncertain.

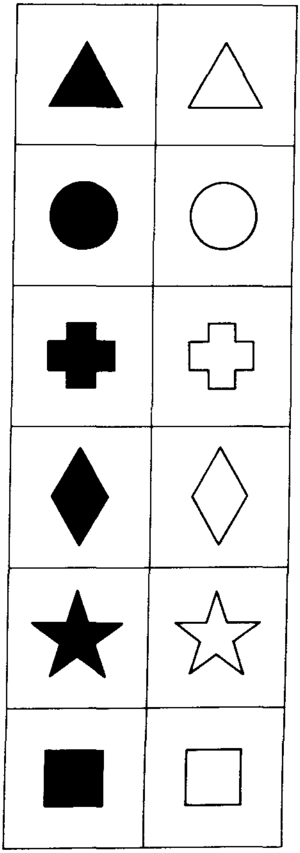

Fig. 4.23. Twelve geometrical stimuli illustrating an implicit analogical structure.

Conclusions

1. Computer algorithms of considerable elegance and power are now available for converting patterns hidden in large arrays of empirical data into a graphical form that is much more readily interpreted by the human investigator.

2. Speech sounds, in particular, can be represented as points in a Euclidean space in such a way that the frequency with which any two phonemes are confused is, to a close approximation, a simple exponential decay function of the distance between the two corresponding points (Figs. 4.2 and 4.4).

3. Moreover, a computer algorithm designed to find the spatial configuration that provides the best (exponential) account of given confusion data can provide useful information about the dimensions or features of the phonemes that underlie the recognition of these speech sounds by human listeners. Since the spatial representation is based solely upon the given confusion data, nothing need be known or assumed about the number or nature of the relevant physical dimensions of the phonemes. In short, the psychological data are allowed to speak for themselves.

4. Applications of this new method to the data of Peterson and Barney [34] and of Miller and Nicely [33] provided further support for their earlier conclusions that the recognition of phonemes is heavily dependent upon the frequencies of the first three formants in the case of the vowels, and upon the features of voicing, nasality, and (to a lesser extent) affrication in the case of the consonants (Figs. 4.1 and 4.3).

5. Although the identification of a consonant has been thought to require discriminations along as many as five different dimensions, 99.4 percent of the variance in the confusion data can in fact be accounted for on the basis of just two underlying dimensions (Figs. 4.1 and 4.2).

6. With respect to such a two-dimensional representation obtained on the basis of the broadband or "flat" conditions, the confusions under the extreme high-pass conditions (Fig. 4.21) tended to extend across the space at right angles to the confusions under the extreme low-pass condition (Fig. 4.16). Hence the two dimensions of the spatial representation may be at least roughly interpreted as reflecting discriminations among high and low frequencies, respectively.

7. Further insight can be gained by embedding, within the spatial representation, a hierarchical clustering of the kind Sørensen [40] originally proposed for use in biological taxonomy and Johnson [41] recently supplied with an elegant mathematical rationale and computational algorithm. Closed curves are drawn around subsets of points that form natural clusters in the sense that all pairs of phonemes within each cluster exceed some criterion level of confusion (Fig. 4.7).

8. As signal-to-noise ratio ranged from −18 to +12 dB in Miller and Nicely's experiment, the overall level of confusion also changed enormously, but the recurrence of the same clusterings at each S/N level indicated that the internal *pattern* of the confusions was essentially invariant (Figs. 4.10 and 4.11). Indeed, with respect to the spatial

representation, the effect of adding a given amount of white noise seemed to be almost entirely confined to a reduction of all interpoint distances by the same, constant factor (Fig. 4.13).

9. The recurrent clusterings reveal a highly reliable difference in structure between the unvoiced consonants and corresponding voiced consonants. In particular, whereas /p/ always groups with the other unvoiced stops /t/ and /k/, /b/ groups—not with the other voiced stops /d/ and /g/—but with the voiced fricative /v/ (in 15 out of Miller and Nicely's 17 conditions) and also /ð/ (in 13 of these 15 conditions).

10. Selective removal of the high frequencies did not have much effect on the pattern of confusions until the high-frequency cutoff was lowered to 300 Hz. Above that point the only reliable difference appeared to be the tendency for /f/ and /θ/ to group with the other unvoiced fricatives /s/ and /ʃ/ rather than (as in the broadband conditions) with the unvoiced stops /p/, /t/, and /k/ (Fig. 4.18). However, when only the frequencies below 300 Hz were passed, the pattern of confusions among the voiced consonants shifted so that, for the first and only time, /b/ grouped with the other voiced stops /d/ and /g/ (Fig. 4.16).

11. Selective removal of the low frequencies, on the other hand, produced a drastic alteration in the pattern of confusions as soon as the low-frequency cutoff was raised as high as 1,000 Hz. When only the frequencies above 1,000 Hz were passed, the distinction between voiced and unvoiced consonants was largely lost for the fricatives but, interestingly, was preserved for the stops (Figs. 4.20 and 4.21).

12. The present findings of a lack of parallelism between the voiced and unvoiced consonants (see point 9, above) and between the stops and fricatives (see point 11, above) complicate attempts to account for the confusion data in terms of the traditional "distinctive feature" schemes unless, perhaps, we include the additional distinctive feature of aspiration in order to distinguish /p/, /t/, and /k/ from the other consonants.

13. Direct judgments of similarity obtained by Peters [19] for all pairs of the same 16 consonants conform to a pattern that departs radically from the patterns consistently found in the confusion data of Miller and Nicely. The fact that Peters' subjects evidently judged each unvoiced consonant to be most similar to its voiced counterpart (Fig. 4.22) suggests that certain cognitive processes (e.g., analogical reasoning) may intervene in such judgmental tasks to alter the pattern from that found in the frequencies with which listeners actually confuse these phonemes.

14. Information gained in these ways about the patterns underlying human confusions among speech sounds may prove useful for engineers

concerned with the development of devices for mechanical recognition or for compression and efficient transmission of speech.

15. Finally, the methods for extracting underlying structure, applied here primarily to data on auditory confusions among individual phonemes, may also be applicable to other quite different problems in communication science and technology. Some of the problems that currently look amenable to this approach are the following: a parallel study of visual confusions among printed letters or numbers [46, p. 90]; the study of semantic or associative structure [47, 48, 49], syntactic structure [50, 51], or, perhaps, the use of such structures for information storage and retrieval (e.g., Ref. 52); the explication of the relations among different languages [4, 53]; and the development of techniques for evaluation of the subjective quality of speech transmitted over different types of circuits [54, 55].

In the five years since this paper was completed, many of the possibilities for application of these methods that were `only suggested in the references listed here have been explored much more extensively. However none of the subsequently published reports appears to alter any of the results or conclusions presented in this chapter. There has been one important methodological development that is directly related to the material presented here. This is the perfection, by my former colleague J. D. Carroll, of a powerful new method for the simultaneous analysis of multiple matrices of the sort considered here [56] and, further, the instructive application of this new method to these very data by Carroll's present colleague, M. Wish [57]. By means of Carroll's new method, Wish has been able to obtain a spatial representation for the 16 consonant phonemes with six distinct dimensions, which he has interpreted as "voiced versus voiceless," "nasality," "sibilant versus nonsibilant," "sibilant frequency," "voiceless stops versus voiceless fricatives," and "second formant transition after voiced consonants." His results, though consistent with the findings presented here, help to bring out some aspects of the underlying patterns in the data in an especially clear-cut form.

REFERENCES

1. J.-J. Chang and R. N. Shepard, Exponential Fitting in the Proximity Analysis of Confusion Matrices, paper presented at the annual meeting of the Eastern Psychological Association, New York, April 14, 1966.
2. J. M. Pickett, Perception of Vowels Heard in Noise of Various Spectra, *J. Acoust. Soc. of Am.*, **29**:613–620 (1957).
3. J. M. Pickett, Perception of Compound Consonants, *Language Speech*, **1**:288–304 (1958).

4. C. D. Chretien, Word Distributions in Southeastern Papua, *Language,* **32:**88–108 (1956).
5. R. R. Sokal and P. H. A. Sneath, "Principles of Numerical Taxonomy," W. H. Freeman, San Francisco, 1963.
6. B. Julesz, Visual Texture Discrimination, *IRE Trans. Inform. Theory,* **IT-8:**84–92 (1962).
7. L. L. Thurstone, "Multiple Factor Analysis," University of Chicago Press, Chicago, 1947.
8. H. H. Harman, "Modern Factor Analysis," University of Chicago Press, Chicago, 1960.
9. W. S. Torgerson, Multidimensional Scaling: I. Theory and Method, *Psychometrika,* **17:**401–419 (1952).
10. C. H. Coombs, "A Theory of Data," John Wiley & Sons, Inc., New York, 1964.
11. G. Ekman, A Direct Method for Multidimensional Scaling, *Psychometrika,* **28:**33–41 (1963).
12. R. N. Shepard, Stimulus and Response Generalization: Tests of a Model Relating Generalization to Distance in Psychological Space, *J. Exptl. Psychol.,* **55:**509–523 (1958).
13. W. S. Torgerson, "Theory and Methods of Scaling," John Wiley & Sons, Inc., New York, 1958.
14. J. H. Greenberg and J. J. Jenkins, Studies in the Psychological Correlates of the Sound System of American English, *Word,* **20** (2) (1964).
15. G. Hanson, "Phoneme Perception," Almqvist & Wiksell, Upsala, 1962.
16. G. Hanson, A Factorial Investigation of Speech Sound Perception, *Scand. J. Psychol.,* **4:**123–128 (1963).
17. G. Hanson, A Further Factorial Investigation of Speech Sound Perception, *Scand. J. Psychol.* **5:**117–122 (1964).
18. R. W. Peters, Research on Psychological Parameters of Sound, *WADD Tech. Rept.* 60-249, Wright Air Development Division, Wright-Patterson Air Force Base, Ohio, 1960.
19. R. W. Peters, Dimensions of Perception of Consonants, *J. Acoust. Soc. Am.,* **35:**1985–1989 (1963).
20. K. V. Wilson, Multidimensional Analyses of Confusions of English Consonants, *Am. J. Psychol.,* **76:**89–95 (1963).
21. R. N. Shepard, The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function, I, *Psychometrika,* **27:**125–140 (1962).
22. R. N. Shepard, The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function, II, *Psychometrika,* **27:**219–246 (1962).
23. R. N. Shepard, Analysis of Proximities as a Technique for the Study of Information Processing in Man, *Human Factors,* **5:**33–48 (1963).
24. J. B. Kruskal, Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis, *Psychometrika,* **29:**1–27 (1964).
25. J. B. Kruskal, Nonmetric Multidimensional Scaling: A Numerical Method, *Psychometrika,* **29:**115–129 (1964).
26. R. N. Shepard, Metric Structures in Ordinal Data, *J. Math. Psychol.,* **3:**287–315 (1966).
27. E. T. Klemmer and N. W. Shrimpton, Preference Scaling via a Modification of Shepard's Proximity Analysis Method, *Human Factors,* **5:**163–168 (1963).
28. R. N. Shepard, Polynominal Fitting in the Analysis of Proximities, *Proc. 17th Intern. Congr. Psychol.,* pp. 345–346, North Holland Publishing Co., Amsterdam, 1964. (Abstract)

29. R. N. Shepard, Stimulus and Response Generalization During Paired-associates Learning, unpublished doctoral dissertation, Yale University, 1955.
30. R. N. Shepard, Stimulus and Response Generalization: A Stochastic Model Relating Generalization to Distance in Psychological Space, *Psychometrika*, **22**:325–345 (1957).
31. R. N. Shepard, Stimulus and Response Generalization: Deduction of the Generalization Gradient from a Trace Model, *Psychol. Rev.*, **65**:242–256 (1958).
32. R. N. Shepard, Similarity of Stimuli and Metric Properties of Behavioral Data, in H. Gulliksen and S. Messick (eds.), "Phychological Scaling, Theory and Applications," pp. 33–43, John Wiley & Sons, Inc., New York, 1960.
33. G. A. Miller and P. E. Nicely, An Analysis of Perceptual Confusions among some English Consonants, *J. Acoust. Soc. Am.*, **27**:338–352 (1955).
34. G. E. Peterson and H. L. Barney, Control Methods Used in a Study of the Vowels, *J. Acoust. Soc. Am.*, **24**:175–184 (1952).
35. R. W. Peters, Dimensions of Quality for the Vowel [æ], *J. Speech Hearing Res.*, **6**:239–248 (1963).
36. J. D. Carroll and J.-J. Chang, A General Index of Nonlinear Correlation and Its Application to the Interpretation of Multidimensional Scaling Solutions, *Am. Psychologist*, **19**:540 (1964). (Abstract)
37. J. E. Miller, R. N. Shepard, and J.-J. Chang, An Analytic Approach to the Interpretation of Multidimensional Scaling Solutions, *Am. Psychologist*, **19**:579–580 (1964). (Abstract)
38. M. Joos, Acoustic Phonetics, supplement to *Language* (Monograph No. 23), **24**:1–136 (1948).
39. R. K. Potter and G. E. Peterson, The Representation of Vowels and Their Movements, *J. Acoust. Soc. Am.*, **20**:528–535 (1948).
40. T. Sørensen, A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons, *Biol. Skrifter*, **5**(4):1–34 (1948).
41. S. C. Johnson, Hierarchical Clustering Schemes, *Psychometrika*. 32:241–254 (1967).
42. E. Mayr, E. G. Linsley, and R. L. Usinger, "Methods and Principles of Systematic Zoology," McGraw-Hill Book Company, New York, 1953.
43. M. Halle, On the Bases of Phonology, in J. A. Fodor and J. J. Katz (eds.), "The Structure of Language," pp. 324–333, Prentice-Hall, Englewood Cliffs, N.J., 1964.
44. W. A. Wickelgren, Distinctive Features and Errors in Short-term Memory for English Vowels, *J. Acoust. Soc. Am.*, **38**:583–588 (1965).
45. Harris Savin, personal communication.
46. E. J. Gibson, "Principles of Perceptual Learning and Development," Appleton-Century-Crofts, New York, 1969.
47. R. N. Shepard, Multidimensional Scaling of Concepts Based upon Sequences of Restricted Associative Responses, *Am. Psychologist*, **12**:440–441 (1957). (Abstract)
48. J. Deese, "The Structure of Associations in Language and Thought," The John Hopkins Press, Baltimore, Md., 1965.
49. C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, "The Measurement of Meaning," University of Illinois Press, Urbana, Ill., 1957.
50. E. R. Gammon, On Representing Syntactic Structure, *Language*, **39**:369–397 (1963).
51. W. J. M. Levelt, A Scaling Approach to the Study of Syntactic Relations, in

G. B. Flores d'Arcais and W. J. M. Levelt (eds.), "Advances in Psycholinguistics," North Holland Publishing Co., Amsterdam, 1970.

52. C. Watson, Computer Generation of Word Association Maps for Man-Machine Communication, *System Development Corporation Rept.* SP-1153, March 25, 1963.
53. E. R. Gammon, personal communication.
54. B. J. McDermott, Multidimensional Analyses of Circuit Quality Judgments, *J. Acoust. Soc. Am.*, **45**:774–781 (1969).
55. V. McGee, Semantic Components of the Quality of Processed Speech, *J. Speech Hearing Res.*, **7**:310–323 (1964).
56. J. D. Carroll and J.-J. Chang, Analysis of Individual Differences in Multidimensional Scaling via an N-Way Generalization of Eckart-Young Decomposition, *Psychometrika*, **35**:283–319 (1970).
57. M. Wish and J. D. Carroll, Multidimensional Scaling with Differential Weighting of Dimensions, in D. A. Kendall, R. C. Hodson, and D. Tautu (eds.), *Mathematics in the Archaeological and Historical Sciences*, Edinburgh University Press, pp. 150–167, 1971.