features salience and bias

LOUIS GOLDSTEIN

UCLA-WPP 39 OCTOBER 1977

Three studies in speech perception:

Features, relative salience and bias

Louis Goldstein

UCLA Working Papers in Phonetics

October 1977

University of California, Los Angeles

For my mother

And to the memory of my father

TABLE OF CONTENTS

ACKNOWLEDGMENTS

There are many people who have helped me in asking the questions I wanted to ask, and in setting a course to answer them. They have taught me how to look at the world, or how to use its tools; they have inspired me with their own work, or they have been there to do what needed doing. I have compiled a list of such people, but I'm sure it is incomplete. If you do not see your name, my apologies: Ava Berinstein, Cathe Browman, Ron Carlson, Sandy Disner, Vicki Fromkin, Jack Gandour, Merrill Garrett, Steve Greenberg, Richard Harshman, Eric Holman, Jean-Marie Hombert, Volker Huss, Leon Jacobson, Hector Javkin, Dennis Klatt, Jim Lackner, Peter Ladefoged, Mona Lindau, Wendy Linker, Ian Maddieson, Willie Martin, John Ohala, George Papcun, Lloyd Rice, Stefanie ShattuckHufnagel, Sandy Thompson, Ginny Valian, Marcel van den Broecke, Diana Van Lancker, Roger Wales and Tom Wickens. Thanks, guys!

This dissertation consists of three separate papers that were prepared, in camera-ready form, for UCLA Working Papers in Phonetics. I would like to thank Renee Wellin and Sue Vanderbrook for doing the typing.

I am grateful to the institutions that provided financial support for me as a graduate student. I was supported in part by a University of California Chancellor's Fellowship, and in part by an NIH grant to the UCLA phonetics laboratory.

ABSTRACT OF THE DISSERTATION

Three Studies in Speech Perception:
Features, Relative Salience, and Bias

by

Louis Mark Goldstein
Doctor of Philosophy in Linguistics
University of California, Los Angeles, 1977
Professor Peter Ladefoged, Chairman

This dissertation consists of three chapters, each of which is a separate paper dealing with a particular issue in speech perception. They are tied together by a concern for integrating the investigation of speech perception, in the narrow sense, into some larger frameworks.

Chapter One compares speech perception and speech production for evidence of categorical phonological features. Multidimensional scaling analyses of three types of English consonant confusions are reported: consonant substitutions in spontaneous speech errors, CV perceptual confusions and VC perceptual confusions. Two data sets of each type are analyzed to assess reliability. Three reliable dimensions emerge in all data sets, corresponding to voicing, stop/fricative and place of articulation. Representation of consonants in terms of

categorical phonological features describes what is common to the configurations of different data types, even though there is reliable detail within each data type that is not captured by categorical features. Such features can be viewed as the components of the internal representation of speech sounds that is common to various perception and production processes.

Chapter Two attempts to relate the results of research on the perception of nonsense syllables to the perception of larger linguistic units, ie., words and phrases. This involves the notion that certain segments or certain syllables in the input signal are more perceptually salient than others; they constrain higher-level decision processes more than others. An experiment is reported in which listeners identify words and short phrases that were excised out of spoken sentences. More errors were made in the unstressed syllables of these excerpts than in the stressed syllables, for both initial consonants and vowels. This supports the hypothesis that stressed syllables are more perceptually salient than unstressed ones. In addition, the relative error rates for different consonants in this experiment agree quite well with the relative error rates in nonsense syllable recognition. This indicates that the relative phonetic ambiguity of a particular segment is one determinant of its perceptual salience in word and phrase recognition.

Chapter Three examines the role of bias in perceptual confusions of consonants. Two different models for extracting response bias from a confusion matrix are compared--a metric and a nonmetric model. The metric model is found to be appropriate for consonant confusions and is

analyzed in detail. Reliable biases are found in the perception of consonants in VC syllables. This bias correlates well with the frequency of the consonants in English words. The bias can also be shown to reflect the phonological naturalness of the consonants. Biases for consonant perception in CV syllables are less reliable than in VC syllables and the correlations with frequency and phonological naturalness are not as good. A model of nonsense syllable recognition is proposed to account for the difference between CVs and VCs. This model claims that lexical mechanisms are invoked by listeners, even when recognizing nonsense syllables.

Chapter 1:

Categorical features in speech perception and production

To appear in:
V. Fromkin (ed). Proceedings of the workshop on slips of the tongue and ear, Vienna, 1977.

Categorical features in speech perception and production[1]

Louis Goldstein

INTRODUCTION

It is a common assumption in research in speech perception that a listener's internal representation of the speech signal is organized, at least in part, in terms of the phonological features employed in linguistic analysis (cf. Studdert-Kennedy, 1976; Fant, 1967; Stevens and Halle, 1967). In fact, the process of speech perception has been summarized as follows (Studdert-Kennedy, 1976, p. 253):

> 'In short, perception entails the analysis of the acoustic syllable, by means of its acoustic features, into the abstract perceptual structure of features and phonemes that characterize the morpheme.'

It will be shown that the kind of evidence previously suggested in favor of the organization of perception in terms of features is inconclusive. The relationship between speech perception and speech production which is demonstrated in this paper provides more conclusive evidence of the importance of features in the internal representation of speech.

Features in phonology

Features have a number of different functions, at different levels, within the context of phonological theory. One, the phonological function, is to classify the segments of a morpheme with respect to universal phonological categories such as voiced, voiceless, stop, etc. This categorical representation is relevant to the application of language-specific phonological rules. Another function, the phonetic, is to specify these segments as to their language-specific (but not speaker-specific) values on a number of potentially continuous phonetic parameters. In the case of some features, the possible values are essentially categorical even in the context of their phonetic function. Consider the feature [voice]. Phonologically, English obstruents can be classified simply as [+voice] or [-voice]. Phonetically, some more detailed specification needs to be made, so that the English

voice/voiceless distinction can be differentiated from similar distinctions in other languages in which there is a systematically different way of realizing the distinction. English obstruents will have to be assigned two modal values with respect to some phonetic parameter of voicing. For example, considering voice-onset time (VOT) as the relevant voicing parameter (Lisker and Abramson, 1964), voiced stops in English might be assigned a value of + 15 msecs and voiceless stops a value of +80 msecs. The fact that /p/, /t/ and /k/ differ systematically from one another in VOT (as do /b/, /d/ and /g/) would *not* need to be explicitly represented, however, since it is likely that this variation is universally predictable (Ladefoged, 1975). Thus, even on the phonetic level, voicing in English can be seen as categorical with the categories representing modal values of eg. VOT, rather than fully abstract classifications.

In this paper, the claim that a speaker/listener's internal representation of speech sounds is organized in terms of features, is taken to mean in terms of *categorical* features. As suggested above, this does not necessarily imply phonological, as opposed to phonetic features. Furthermore, demonstration of psychological validity for some categorical features does not rule out the possibility of there being some non-categorical (i.e. continuous) features that play a role in speech perception and production, as well. In fact, a model proposing a rather simple relationship between continuous and categorical features will be outlined below.

Evidence for categorical features in perception

Evidence for features in perceptual representations have come primarily from two basic types of experiment: selective adaptation (see Cooper, 1975, for review) and phoneme confusion and similarity judgment (see Singh, 1975, for review). This paper will primarily concern itself with the latter type of evidence. Evidence from consonant confusions or from subjects' judgments of consonant similarity has been often presented as evidence for features of consonants (Miller and Nicely, 1955; Singh, 1966; Singh and Black, 1966; Wang and Bilger, 1973; Peters, 1963; Singh, Woods and Becker, 1972). The logic that provides for internal representation in terms of features from these experiments is as follows: when consonants are divided into groups according to features, the consonants within the group are, on the average, more confusable with each other than with consonants *not* in the group. (A statistical formulation of this procedure has been described by Klatt (1968), who applied it to memory confusions). This grouping is then taken to reflect the psychological organization of the consonants in question. A similar argument can be made for studies employing similarity judgments. The consonants within a group can be shown to be judged more similar to one another than to consonants not in the group. The problem with this approach, however, is that detailed examination of the similarity or confusability structure in these experiments may yield reliable patterns that are richer and more detailed than can be accounted for by hypothesizing an internal representation in terms of feature categories. In particular,

there may be reliable within - group patterns. Since some kind of auditory or cognitive variables will have to be proposed to account for this microstructure in the data, it is conceivable (and more parsimonious) that these variables also account for the division of consonants into groups.

A common technique for revealing similarity or confusability structure (and the technique that is employed in the present study) is multidimensional scaling. This family of mathematical techniques (Torgerson, 1958; Shepard, 1962; Kruskal, 1964; Harshman, 1970; Carroll and Wish, 1974) models the stimuli of a similarity or confusion experiment as points in an n-dimensional space, such that the distances between pairs of points in the space can be related to the similarity or confusability between pairs of stimuli by some simple function. The nature of this function differentiates different forms of multidimensional scaling. The number of dimensions present in the configuration must be determined empirically. The greater the number of dimensions, the better the fit of the configuration to the data will be. However, since one would like to include in the configuration only the reliable properties of data, not the noise, one should choose a dimensionality that stops short of accounting primarily for noise. (Practical approaches to this problem are outlined in the methods and results sections, below).

Having derived a spatial configuration that represents the similarity structure inherent in a confusion matrix in some satisfactory number of dimensions, one can examine this structure for the presence of groups that are meaningful from a linguistic or perceptual point of view. This examination requires the researcher to choose a particular set of reference axes in the configuration. Any pair of axes in the space can be rigidly rotated (maintaining 90° between these two axes) without altering the interpoint distances in the configuration. While there are certain techniques for deriving non-arbitrary rotations of the axes (Harshman, 1970; Carroll and Chang, 1970), such procedures are not applicable to all experimental situations. Thus, in the experiments to be reviewed below, as well as in the new analyses presented in this paper, reference axes are rotated so as to make the resulting configurations maximally interpretable. That is, axes are rotated so that the stimuli seem to be grouped meaningfully.

As an example of the application of multidimensional scaling to perceptual confusions, consider the re-analysis by Shepard (1972) of the Miller and Nicely (1955) consonant confusion data. Miller and Nicely presented subjects with CV syllables under different levels of noise, and different conditions of band-pass filtering. The syllables included the consonants shown in Fig. 1 before the vowel /ɑ/. Shepard analyzed the combined data from all noise conditions, with no filtering. He calculated a proximity measure for each pair of consonants, based on the number of times the two consonants were confused with one another. (This procedure is outlined in the method section, below). He fitted these proximities to a spatial configuration using an exponential function. Two dimensions accounted for 99.4%

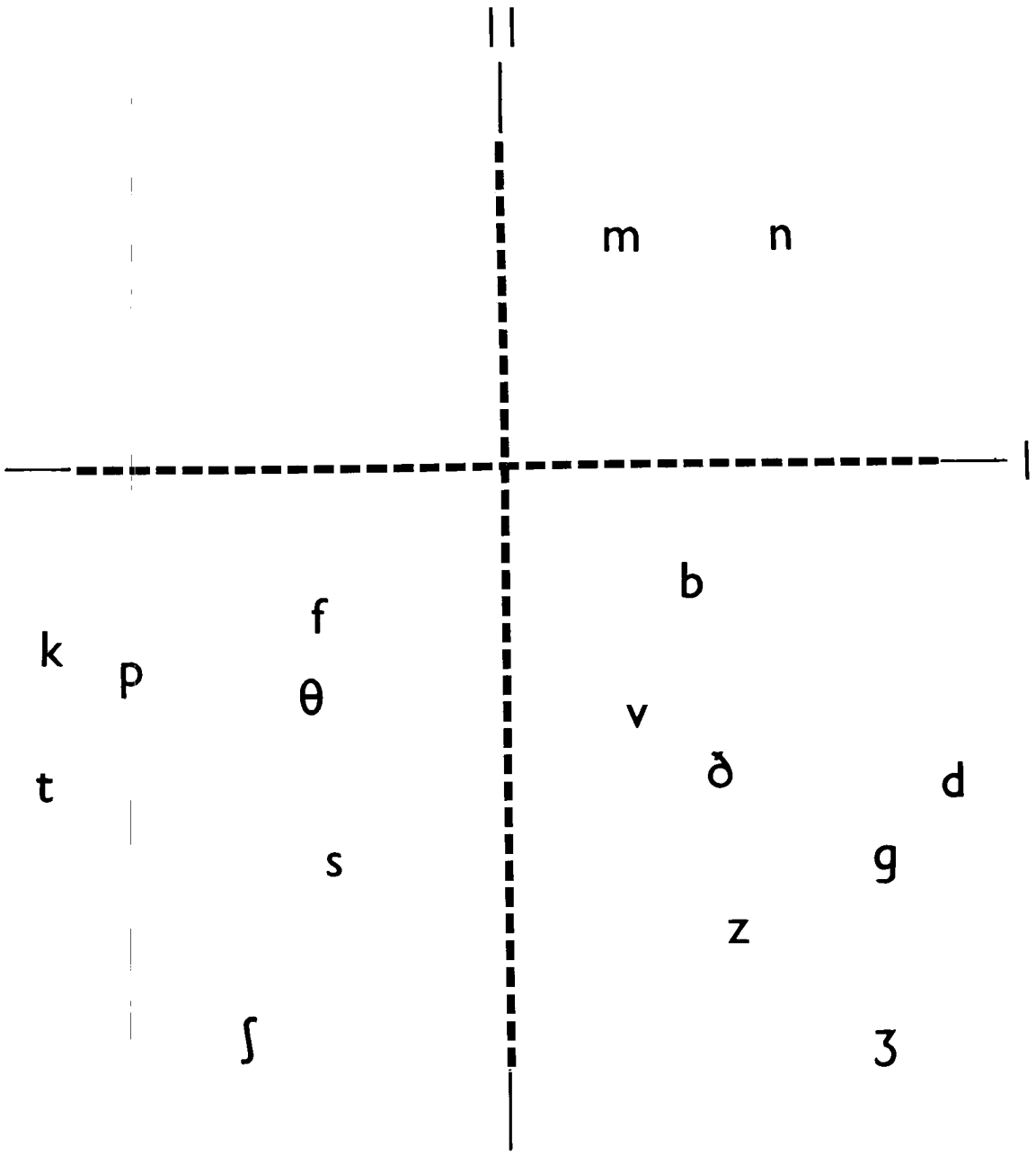

Figure 1. Two-dimensional configuration of Miller-Nicely (1955) consonant confusion data, as analyzed by Shepard (1972). Dotted lines separate voiced from voiceless (I) and nasals from non-nasals (II).

of the variance in the proximities. Fig. 1 shows the positions of the 16 consonants in the two-dimensional space, using the rotation of reference axes chosen by Shepard. Dimension I clearly divides the consonants into two large groups -- voiced vs. voiceless. Dimension II divides consonants into nasals and non-nasals. Since these groups correspond to phonological features, this could be taken as evidence of the importance of these features in perceptual organization. However, there is no evidence in Fig. 1 that these groups are categorical -- at least in the voiced, voiceless and non-nasal groups there is substantial structure within the groups.

How is this within-group structure to be accounted for? Before attempting to account for it at all, we need to know whether or not this structure is reliable. That is, a comparable independent data set should be examined to determine if this structure is observed in both sets. It is not known whether or not this configuration of the Miller-Nicely data is reliable in this sense. However, we assume that it is, in order to continue the example. Clearly we can see the influence of other features within these groups. For example, dimension I divides voiceless consonants into stops and fricatives. Both voiced and voiceless fricatives are distributed along dimension II according to place of articulation. Thus, the reliable within-group variance might be due to other categorical features, perhaps indicating that too few dimensions were extracted in the analysis. It is possible to use statistical procedures to remove the potential effects of other categorical features on these dimensions. One could then determine whether there was any reliable within-group structure still present after removing the effect of other categorical features. If such an analysis were to find that residual variance was reliable, after removal of all categorical features, this would indicate that some non-categorical properties of the stimuli or the perceptual system were being tapped in this experiment. One would then need some way of determining whether the major groupings themselves (nasality and voicing, in this case), were a function of internal perceptual categories or the same continuous properties that account for the within-group structure.

An example of a scaling analysis in which the dimensions were correlated with continuous properties of the stimuli can be found in Ingram (1975). He had subjects rate similarity among 12 consonants (shown in Fig. 2) in the environment before the vowel /ɑ/, by means of triadic comparison. The two-dimensional configuration resulting from a non-metric multidimensional scaling (which requires only that the relationship similarities and distances in the space be monotonic) is shown in Fig. 2. Dimension I divides consonants into three phonological categories -- stops, resonants and sibilants. However, Ingram shows that the values of the consonants on this dimension correlate highly (.94) with the measured durations of the consonant stimuli used in the experiment. Similarly, he finds that the values of dimension II can be shown to correlate highly (.92) with a weighted difference of energy in two selected frequency bands. Thus, we are faced, once again, with an ambiguity in interpreting the results.

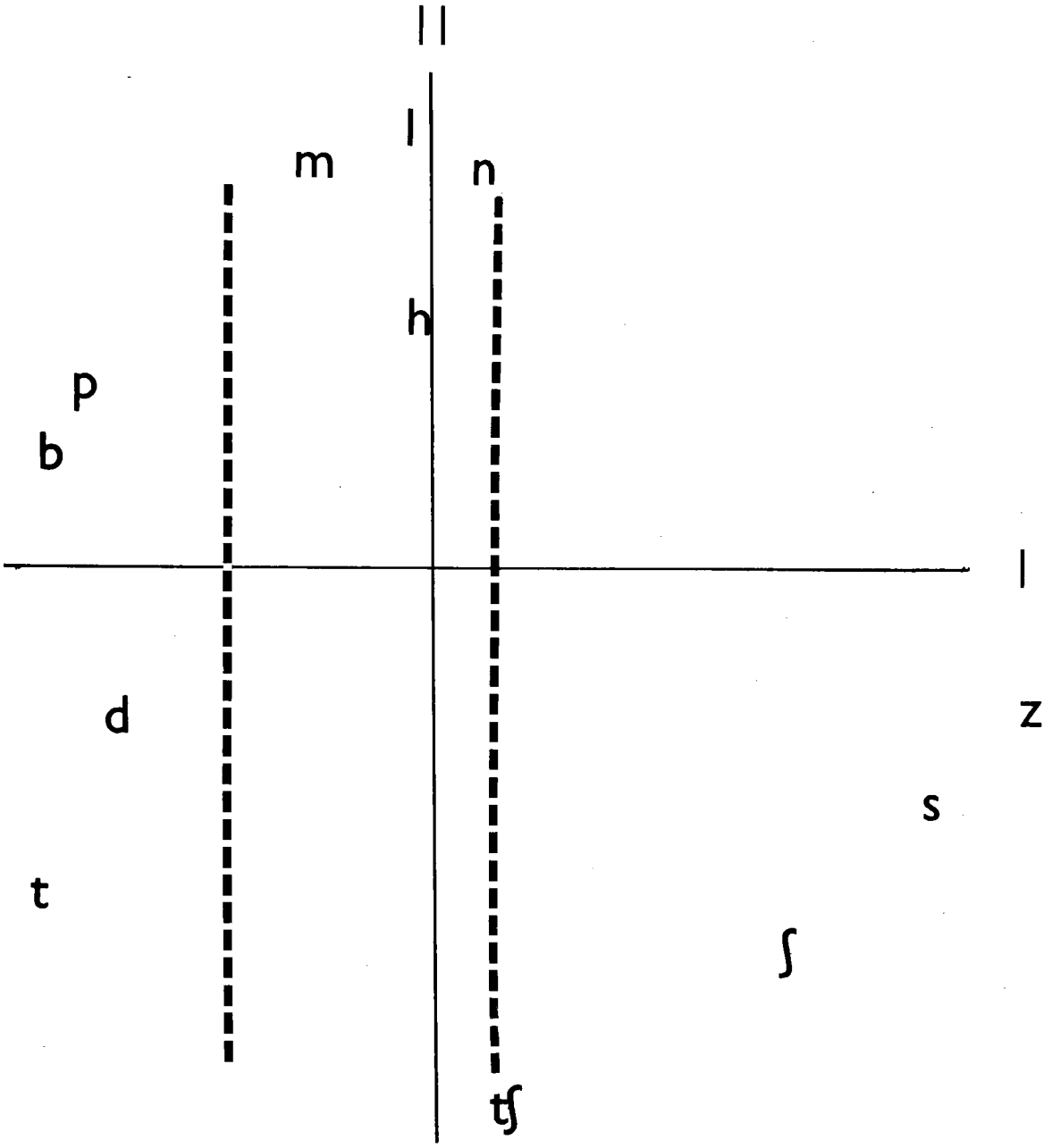

Figure 2. Two-dimensional configuration of /Ca/ similarity judgement data, after Ingram (1975). Dotted lines divide dimension I into stops, resonants and sibilants.

Do they provide evidence for categorical features in the perceptual representation of speech sounds, or are the similarity judgments explicable in terms of low-level auditory properties of the stimuli?

The analyses to be described below attempt to resolve some of interpretive problems outlined above and to establish the importance of categorical features in speech perception and production. Specifically, we view categorical features as representing what is common to perception and production processes. Our procedure is to find pairs of sets of data in perception and production that can be considered to be close replications of one another. It will then be possible to determine the extent to which there is reliable non-categorical variance in perceptual and in production spaces. It will also be possible to compare dimensions of perception with those of production. It is hypothesized that the common variance between perceptual and production spaces will be categorical in nature.

Obtaining experimental confusion data for speech production is more difficult than obtaining the comparable data for speech perception. Therefore, analysis of speech production is based on errors in spontaneous speech production (Fromkin, 1973; Shattuck, 1975). In particular, single sound changes -- cases in which one segment is replaced by another -- are used as a measure of consonant similarity with respect to the production system. Earlier analyses of this type of error have shown that consonants that share features tend to be substituted for one another more than consonants that do not. MacKay (1970) has shown that for 56% of the consonant reversals he analyzed, the consonants differed in only one feature. Thus, there is reason to conclude that speech errors are constrained by some measure of phonological similarity, and, therefore, it is reasonable to compare them with perceptual confusions.

METHOD

Data: Production

Two independently collected sets of spontaneous speech errors were analyzed, so that reliability of the analysis could be assessed. The corpus of over 7,000 spontaneous errors described by Fromkin (1973) constituted the basis for one set. These were examined for errors in which a single consonant in the intended utterance is replaced in the actual utterance by a different consonant. Such errors belonged to one of four categories :

(1) substitutions: a consonant is replaced by another that does not appear anywhere in the immediate environment.

example: a milk shake → a bilk ...

(2) anticipations: a consonant is replaced by one that occurs later in the utterance.

example: also share → alsho share

Table 1. Consonant confusion matrix for data set SOT1. Rows represent intended consonant and columns replacing consonant.

	p	t	k	b	d	g	m	n	f	v	θ	ð	s	z	ʃ	ʒ	l	r	w	j	h	tʃ	dʒ
p		20	25	12	3		15	1	23	2	1		4				3	2			4	3	1
t	26		16	7	10	1	2	10	7		3		10	1			12	7	2			3	5
k	22	18		7	4	19	1	2	7	1			5		2		1	1	1		5	3	
b	18	2	11		11	15	21		7	5			2		2		2	4	3		1		2
d	3	6	7	15		4		7	2	1	1		1				10	3	2		1		9
g		1	8	12	4		2	1	2	1			1	2	1		1	2					
m	9	5	1	26	1	1		27	4	1			2	1			12	5	5		3		3
n	3	4	4		12	1	23		1	2			2	1			13			3			2
f	25	4	6	7	1	1	10	1		8	1		9		2		2	1	1		4	2	
v	1			2	2		2	2	8		1	2	1	1			6	1	2				1
θ		1							5			1	10		1		1				1	1	
ð															1								2
s	12	11	5		2		4	4	11	2	2			5	32		3	2	2		4	3	
z					1	2	1	2		4						2	1						2
ʃ			2	1	1	1			3				9			1	1	1	1		1	1	
ʒ															1								1
l	4	6	1	2	8		8	11	2	4			1	1	1			42	6	5	5	1	3
r	6	3	1	5	3	2	4	2	3				4		1		37		10	3	4		5
w	3	1	1	5	2		4	1	1	5			1		1		3	8		1	2		1
j		1						3		1							5	2	1				1
h	2		6	3	2		4	1	7				4		2		4	4	2				3
tʃ	3	1	3	1									2		2		1						2
dʒ		2			3		4	3		1					2		4	2	1			2	

(3) perseverations: a consonant is replaced by one that occurred earlier in the utterance.

example: give a boy → give a goy

(4) spoonerisms: two consonants in an utterance replace each other.

example: keep a tape → teep a cape

Errors belonging to any of these classes were tabulated into a confusion matrix, indicating the total number of times each target consonant was replaced by each of the other possible consonants. Each spoonerism was included twice in this matrix, as each of the reversing consonants can be considered as both target and replacement. The resulting matrix, referred to as SOT1 (for slips of the tongue), includes 1,369 entries. (See Table 1).

Shattuck (1975) presented a target-replacement confusion matrix constructed according to these same principles, from her corpus of speech errors. This matrix constituted the other set of speech error data for the present analysis. It includes 1,057 entries, and is presented as SOT2 in Table 2.

Data: Perception

Perceptual confusion data from published results of Wang and Bilger (1973) were analyzed. They presented listeners with series of both CV and VC syllables in two types of listening conditions: *noise* -- signal mixed with noise at various S/N levels; and *quiet* -- signal played without added noise at various low levels. The consonants /p, t, k, b, d, g, f, θ, s, ʃ, v, ð, z, ʒ, tʃ, dʒ/ served in both CV and VC syllables. The vowels /ɪ, ɑ, u/ were employed throughout, but subjects were only required to identify the consonant. Wang and Bilger present four separate confusion matrices for these consonants. Each has data for a given syllable type (CV or VC) in a particular experimental condition (noise or quiet). Each matrix is summed over subjects, vowels, and S/N level. These four confusion matrices were used in the present analysis. CVs and VCs were both analyzed in order to examine possible systematic differences in perceptual dimensions in different syllable positions. Noise and quiet conditions were both analyzed so as to allow the reliability of CV and VC data to be assessed. The noise and quiet conditions are clearly not exact replications. However, examining what is reliable across these conditions isolates those perceptual relationships that are at least robust enough to recur under minor variations in experimental conditions.

To make the perception and production data sets comparable, it was decided to analyze only the speech error data relevant to the sixteen consonants in the perceptual confusion experiments. Submatrices were generated from the matrices in Table 1 and Table 2 by eliminating those rows and columns corresponding to consonants not in the confusion

Table 2. Consonant confusion matrix for data set SOT2. Rows represent intended consonant and columns replacing consonant.

	p	t	k	b	d	g	m	n	f	v	θ	ð	s	z	ʃ	ʒ	l	r	w	j	h	tʃ	dʒ
p		10	16	4			2	2	23	1		1	2						1		1		
t	15		16		5		2	7	3				13	1			2	5				7	2
k	12	16		1		7		2	6	1			2	1							8	4	
b	3		1		9	5	7		4	7			2	1			1	3	3	1			1
d		4		8		6		6	1	1		1	2	5			4					1	4
g		1	10	8	5		1	1	1	2				1	1						1		
m	3			7	1			22	1			1					5	10	12				2
n	1	2	1	1	4	1	24			1			2	2			17	1					
f	25	1	4	4		1	1			3	4		13	1	1		3		3	1	2		
v	1	1		3	3	1	1	1	3			3	1	3			3	2	1				
θ		2							1				6									1	
ð					1		1							1			1						1
s	2	13	3	1	4			2	10	1	7			2	33		2				3	4	2
z					4	1	1	2		3		1				3							2
ʃ	1	1								1	1		20				1				1	1	
ʒ														1									1
l		3		5	1		3	11	2	2	1		1					36	5	11	2		1
r		2		2	2		5	1									47		21	4	1		1
w				5	1		12	2	3	1							11	26		1	1		1
j								1									11	3					
h			5			1			2				3	1	1				1			1	
tʃ	1	3	4		1								3		2								
dʒ		1		1	7		2	1										2	1				

sets. Thus, six sets of data were analyzed--SOT1,SOT2,CVN,CVQ,VCN and VCQ. There are three types of data -- speech errors, CV perception, VC perception -- and two sets of data designed to allow reliability assessment within each type.

Analysis

Nonmetric multidimensional scaling was the main tool for the analysis (Shepard, 1962; Kruskal, 1964). The nonmetric procedure makes a minimal assumption about the function that relates the similarity data to the distances in the derived configurations. That function is constrained only to be as close to monotonic as possible. This type of scaling was chosen so that the perceptual confusions and speech errors could be analyzed using identical techniques. While it is known that an exponential function fits well for perceptual confusion data (Shepard, 1972), it was not known what function would be best for speech errors.

Before scaling the data, a single value, representing the similarity associated with each pair of consonants was calculated to submit to the scaling routine. For the perceptual confusions, the procedure to derive these similarities from the confusion matrices was as follows: First, the entries in each row of the confusion matrix (i.e. the responses to a given stimulus) were multiplied by a constant proportional to the reciprocal of the number of times the stimulus had been presented. This eliminated slight discrepancies in the number of times each stimulus was presented. These corrected confusions were then converted into similarities by using equation (1), from Shepard (1972):

$$(1) \quad Sim(i,j) = \frac{f(i,j) + f(j,i)}{f(i,i) + f(j,j)}$$

where $Sim(i,j)$ = similarity between i and j
$f(i,j)$ = number of confusions of stimulus i with response j
$f(i,i)$ = number of correct responses for i

Luce's model of choice (Luce, 1959) attributes differences between stimuli in number of correct responses to differences in bias between the stimuli. Thus, dividing the number of confusions between a pair of consonants by the number of correct responses for that pair of consonants removes the effect of bias from the resulting values. For the speech error data, bias could not be removed in this way, since there is no diagonal, i.e. no correct responses, in the speech error confusion matrices. Therefore, an estimate of similiarity was derived from speech errors simply by averaging the matrix across the diagonal, according to (2) below:

$$(2) \quad Sim(i,j) = f(i,j) + f(j,i) \ /2$$

The non-metric multidimensional scaling was performed by the KYST program (Kruskal, Young and Seery, unpublished manuscript) on the IBM 360/91 at UCLA. This program uses an iterative method to find a con-

figuration of points in an n-dimensional space that minimizes a quantity referred to as stress (Kruskal, 1964). Stress is calculated by monotonically regressing the distances between points in the configuration (D_i) onto the input data. If $\hat{d}_i$ represents the values predicted by this regression, then stress is a measure of the departure of $\hat{d}_i$ from D_i as in (3) below:

(3)
$$\text{STRESS} = \sqrt{\frac{\sum_{i=1}^{NN} (D_i - \hat{d}_i)^2}{\sum_{i=1}^{NN} D_i^2}}$$

where NN = number of pairs of points.

The starting configuration for the present analysis was the metric scaling configuration of the data (Torgerson, 1958). This starting configuration maximizes the probability of the program finding a global, rather than local, minimum of stress. The metric scaling assumes that the similarity values, themselves, represent a linear function of the true interpoint distances among the stimuli. It tries to find the best n-dimension approximation to this set of distances. The program allows a choice of a metric for the relationship between projections of points on the n-dimensions of the space and the interpoint distances. The Euclidean distance metric was chosen, which assumes the relationship between dimensions and distance as in (4) below:

(4)
$$D(x,y) = \sqrt{\sum_{i=1}^{M} (x_i - y_i)^2}$$

where $D(x,y)$ = distance between points x and y
x_i = coordinate of point x on dimension i
y_i = coordinate of point y on dimension i
M = number of dimensions

Solutions were obtained in one to six dimensions.

RESULTS

Dimensionality

A first task in interpreting the results of a multidimensional scaling is to determine (at least tentatively) how many dimensions are required to best model the data. In the ideal configuration, all of the systematic, reliable properties of the data are modeled but noise in the data is not. There are no automatic procedures for determining how many dimensions provide the best approximation to this ideal configuration. The most widely accepted technique is to examine the stress values at a range of dimensionalities. Any plot of stress vs. dimensionality will show stress decreasing as dimensionality increases,

because the fit to the data will always be better with more dimensions. Let us assume that at some dimensionality, n, the configuration begins to account primarily for noise rather than systematic properties in the data. The reduction in stress between dimension n-1 and dimension n would, therefore, be expected to be smaller than the reduction observed from dimension 1 through dimension n-1. If this occurs, one can see an elbow in the plot of stress vs. dimensionality at dimension n-1. Furthermore the decrease in stress for additional dimensions n+1, n+2, ... will be comparatively constant as each dimension accounts for a similar proportion of the noise.

Plots of stress vs. dimensionality for the six sets of data analyzed are shown in Figure 3. One through five dimensional solutions are plotted. The six-dimensional solutions did not achieve a minimization of stress for some of the sets of data, and will not be considered further. Clear elbows emerge for the CVN and CVQ data sets. This elbow is at three dimensions. An elbow also occurs at three dimensions for the VCN data, but the other sets do not have convincing elbows at all. The lack of any observable elbow is one problem with using this criterion for dimensionality. Moreover, Shepard (1974) has argued that interpretability and reliability of solutions are better guides to choice of dimensionality than is reduction of stress. Similar points have been made by Wish and Carroll (1974) and Gandour and Harshman (1977), who have shown that reliable, interpretable dimensions may contribute only minimally to the goodness of fit of a scaling solution. Thus, the lack of elbows in the present analysis need not be considered either as a problem in the analysis or as a barrier to further interpretation.

A theoretically more motivated procedure for determining dimensionality has been proposed by E. Holman (personal communication). This procedure requires analyzing a pair of sets of data that can be considered to be replications. One then correlates, across a range of dimensionalities, the interpoint distances derived for one data set with the raw data from the paired data set. This correlation should increase as dimensionality increases, as long as the dimensions represent reliable properties of the data that are common to both sets of data. At the dimensionality where one begins to extract data-set-specific noise, this cross-set correlation should stop increasing, or even decrease.

These cross-set correlations were calculated for the two sets of speech errors, the two CV sets and the two VC sets. Correlations were calculated for one through five dimensions, using the GAMMA program (provided by Eric Holman). The program calculates gamma, a rank order correlation (related to Kendall's tau), since the distances are fitted only on the basis of monotonicity with the data in non-metric scaling. Figure 4 plots gamma vs. dimensionality. For each plot, the labelled data set provides the interpoint distances for the correlation. For both SOTs and CVs, gamma clearly stops increasing after three dimensions. For the VC pair

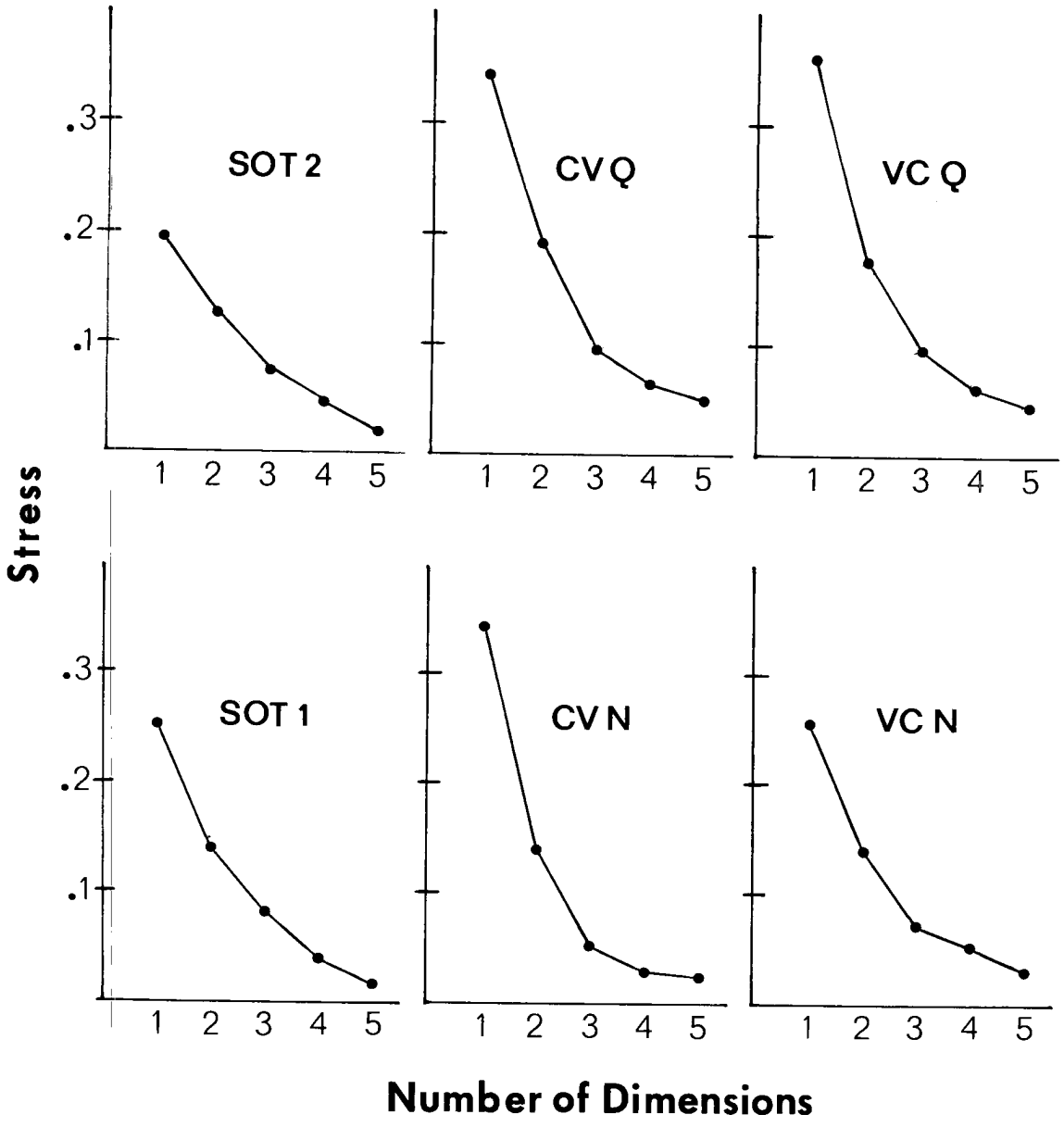

Figure 3. Plots of Kruskal's stress as a function of number of dimensions extracted, for all data sets analyzed.

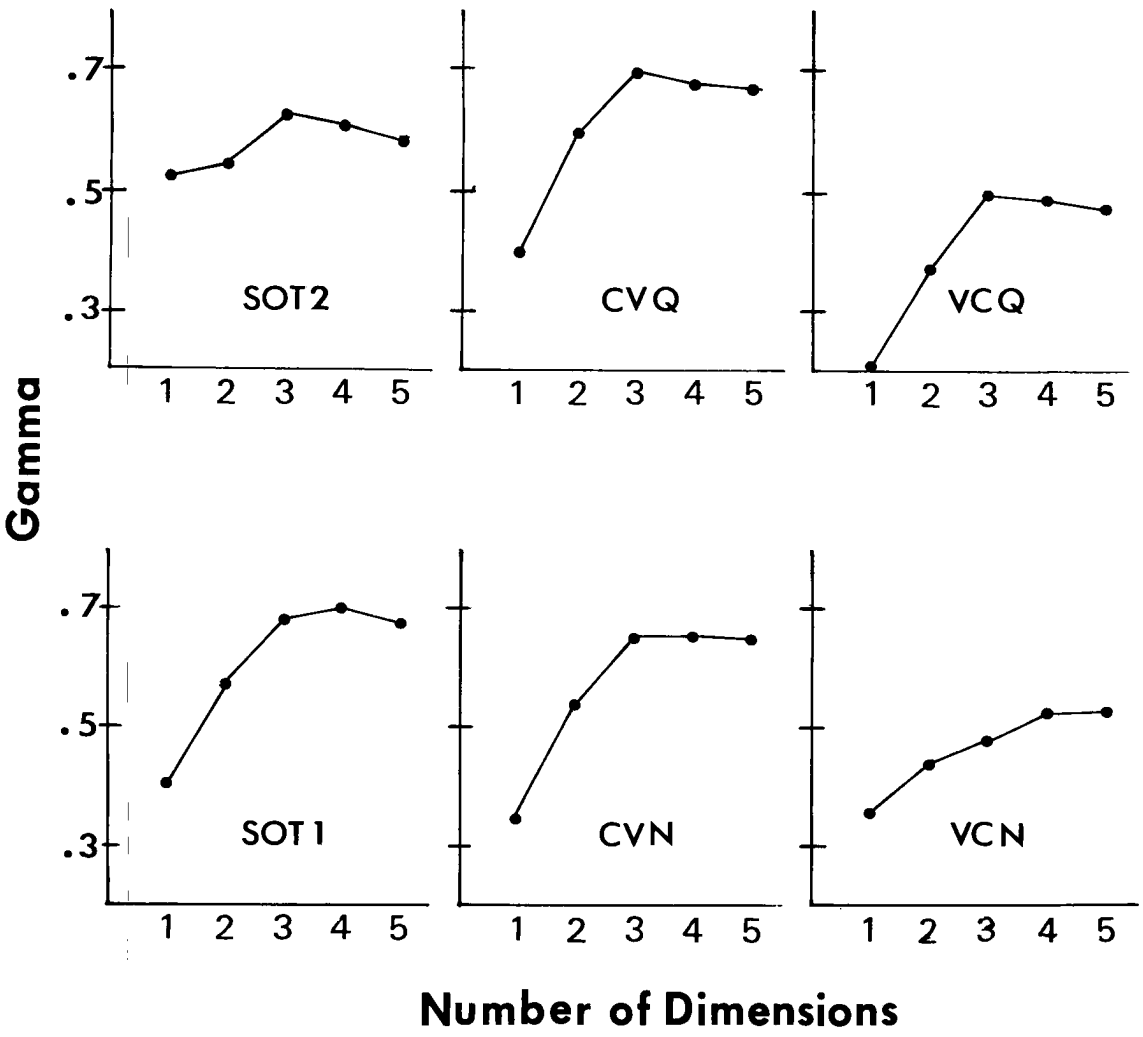

Figure 4. Plots of gamma (rank order correlation) as a function of the number of dimensions extracted. Each plot represents the correlation between the interpoint distances derived for a given data set with the raw similarity data for its paired set. Labels on the plots refer to interpoint distances used in the correlation. Paired sets are SOT1-SOT2, CVN-CVQ, VCN-VCQ.

of data sets, gamma peaks at three dimensions when correlating the VCQ configuration with the VCN data, but at four dimensions when correlating the VCN configuration with the VCQ data. The reason for this asymmetry is not clear, and it may be related to the generally poor cross-set correlations for VCs, compared to speech errors and CVs. Whatever the reason, if we are concerned with how many dimensions are reliably present in both of two independent sets of data, then we must choose three for VCs and well as for speech errors and CVs.

Interpretation

The three-dimensional solutions for all six sets of data were rotated orthogonally (using Comrey's (1973) program) to linguistically plausible configurations. In fact, the configuraitons could be rotated so that roughly the same three linguistically-related dimensions appeared in all six data sets. This further supports the desirability of the three-dimensional solutions. The three dimensions could be interpreted as follows: I - stop consonants vs. fricatives, II - voiced vs. voiceless; and III - place of articulation.

The positions of the 16 consonants in the plane formed by dimensions I and II are shown in Figures 5 through 7, for CV, VC and speech error data respectively. The two data sets of a given type are shown side by side in the same figure, so that reliability can be assessed graphically. Dimension I divides the consonants into two groups according to the categories [± continuant] and dimension II divides them into groups according to the categories [± voice]. Wherever such division into groups is possible without the groups overlapping, a dashed line is drawn in the space. Only two cases show and overlap -- VCQ and SOT1 (Fromkin data) configurations for dimension II. Examination of these cases shows the overlap to be very small. Thus, the categories [± continuant] and [± voice] are reliably distinguished in the various data sets analyzed.

Examining Figures 5 through 7 more closely it is clear that the two data sets of a given type are highly similar to one another in detail. Systematic differences, however, can be found between different data types. For example, the affricates /tʃ/ and /dʒ/ are at the extreme stop end of dimension I in both VC data sets, but are at the border between stops and fricatives in both sets of CV data. There is a reasonable explanation for affricates being "extreme" stops syllable-finally, but only "weak" stops syllable-initially, particularly in an auditory task. Unlike other stops, an affricate in final position must be released, and thus the silent interval corresponding to the stop closure will always be in the acoustic signal. Other stops may be unreleased, or released very weakly, so that there is less likely to be a silent gap in the acoustic signal. In utterance-initial position, no stops will have silent gaps in the signal, of course. However, the release

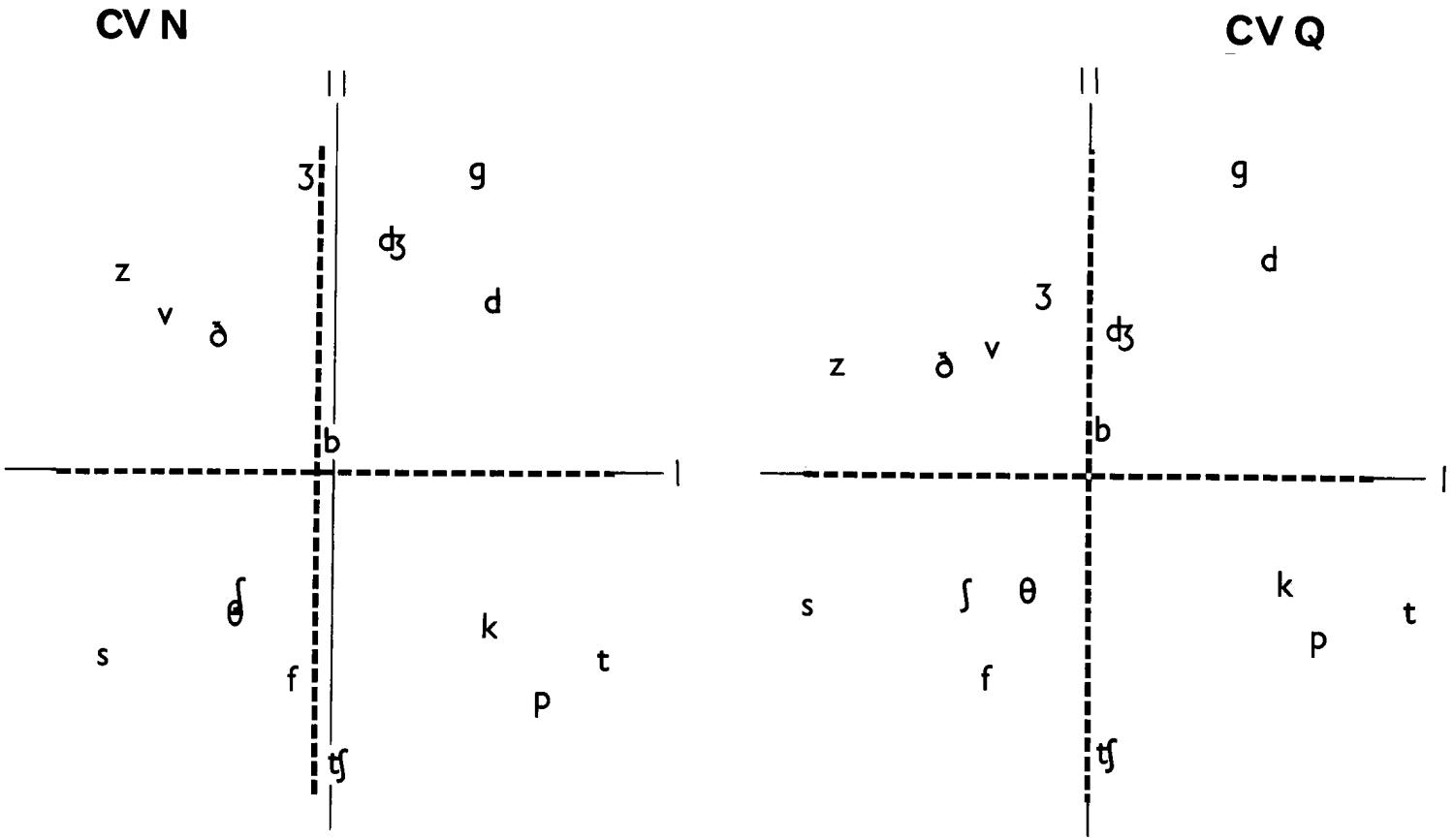

Figure 5. Plane formed by dimensions I and II in three-dimensional configuration for CVN and CVQ data. Dotted lines separate stops from fricatives (I), voiced from voiceless (II).

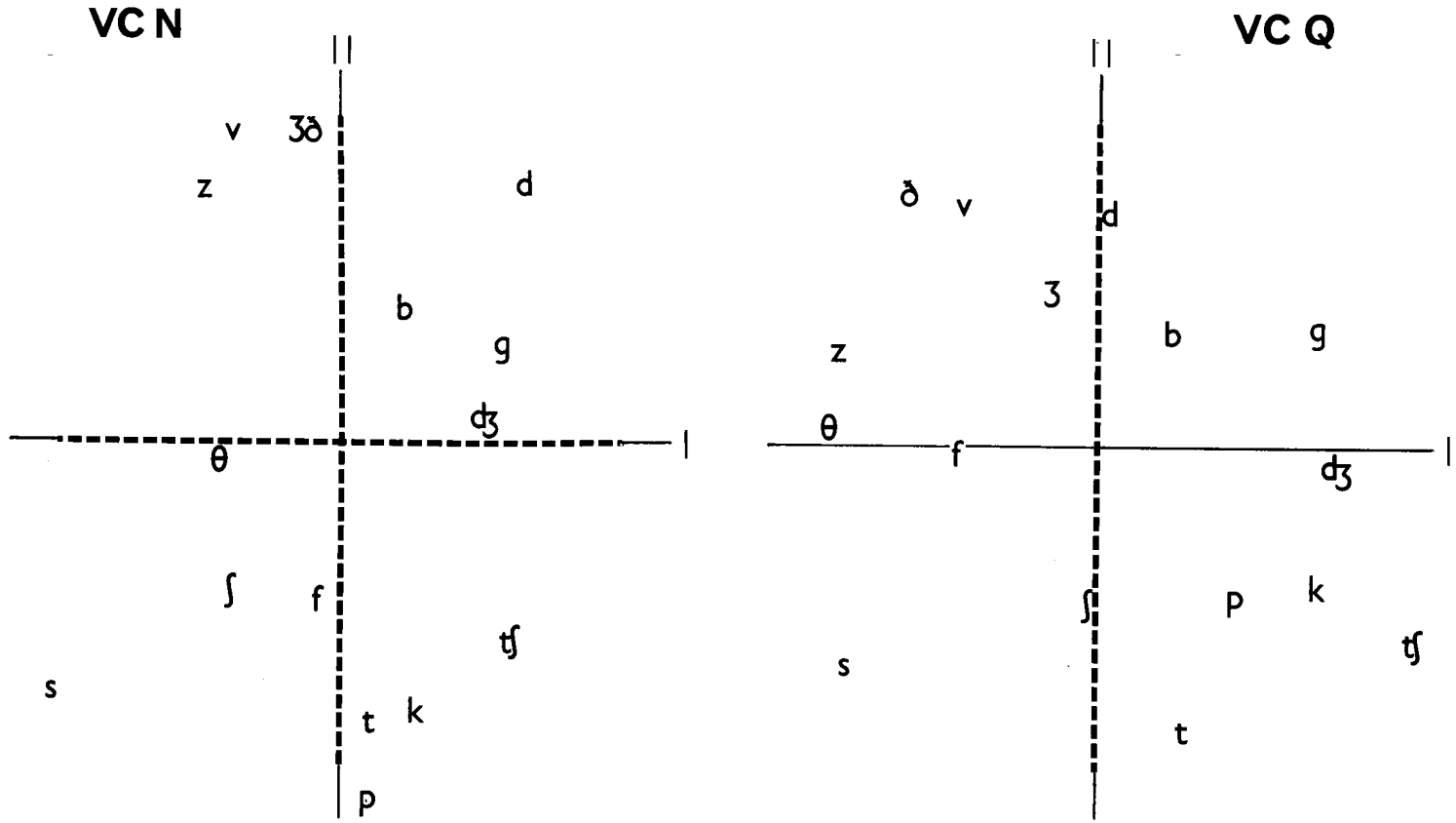

Figure 6. Plane formed by dimensions I and II in three-dimensional configuration for VCN and VCQ data. Dotted lines separate stops from fricatives (I), voiced from voiceless (II).

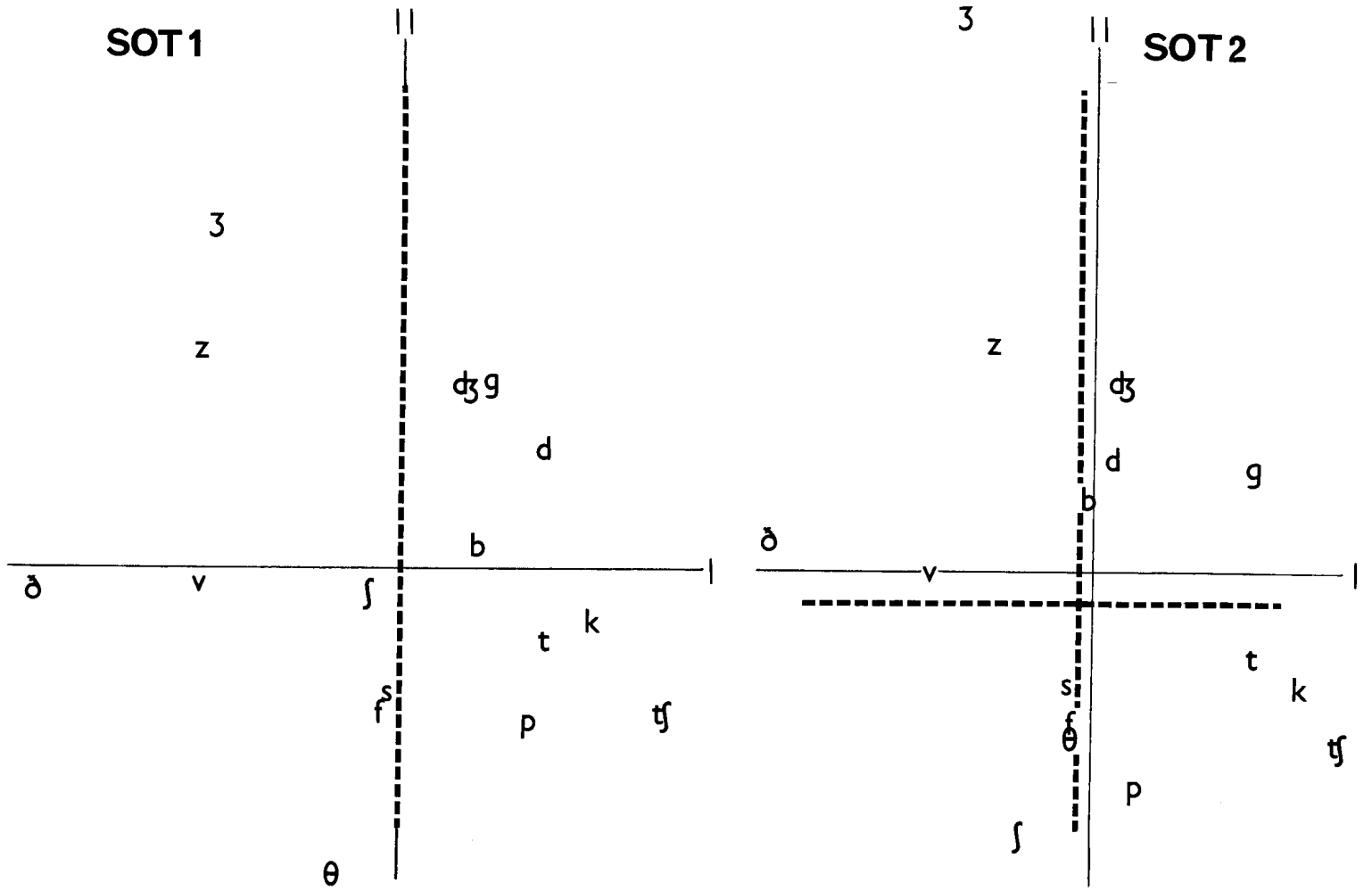

Figure 7. Plane formed by dimensions I and II in three-dimensional configuration for SOT1 and SOT2 data. Dotted lines separate stops from fricatives (I), voiced from voiceless (II).

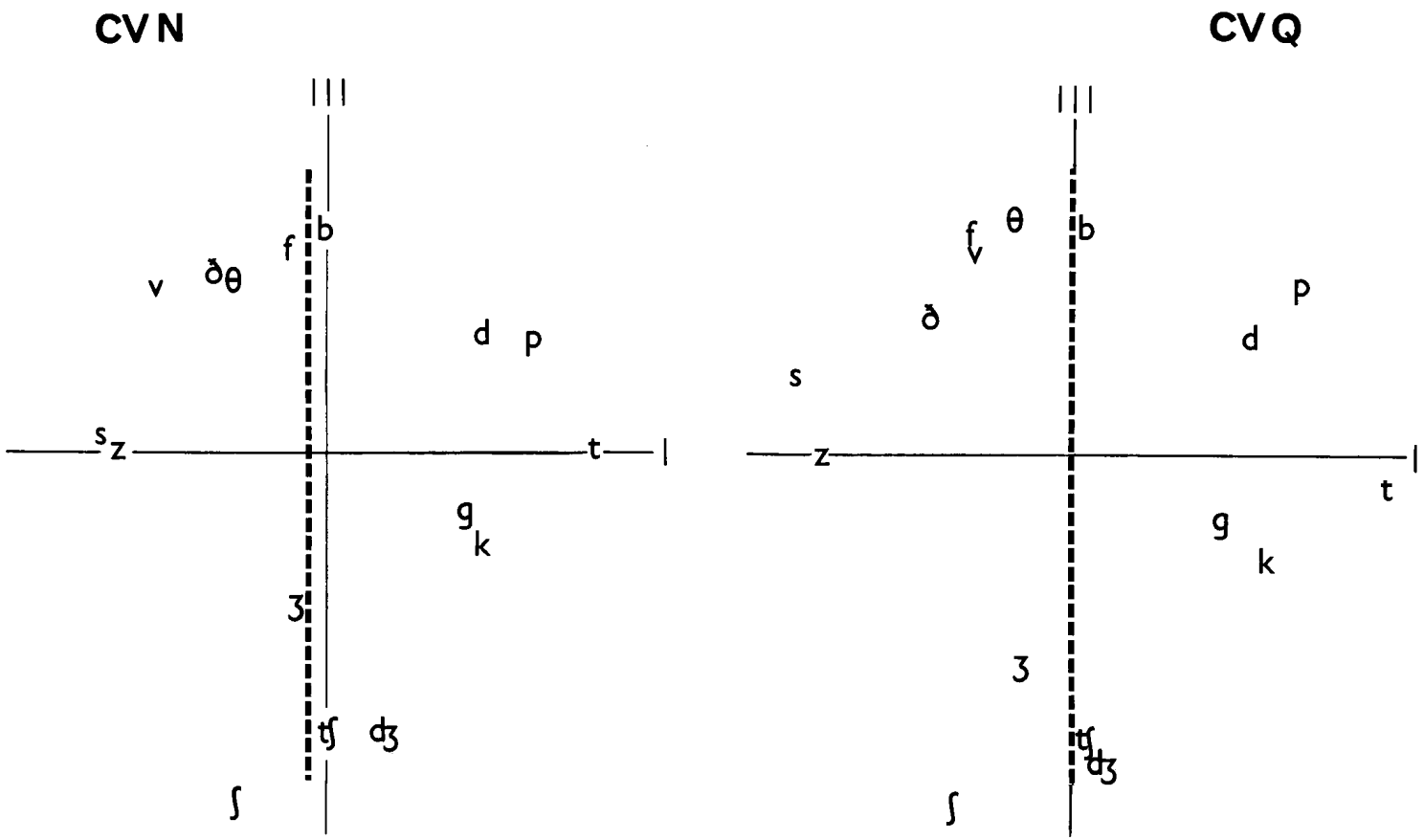

Figure 8. Plane formed by dimensions I and III in three-dimensional configuration for CVN and CVQ data. Dotted lines separate stops from fricatives (I).

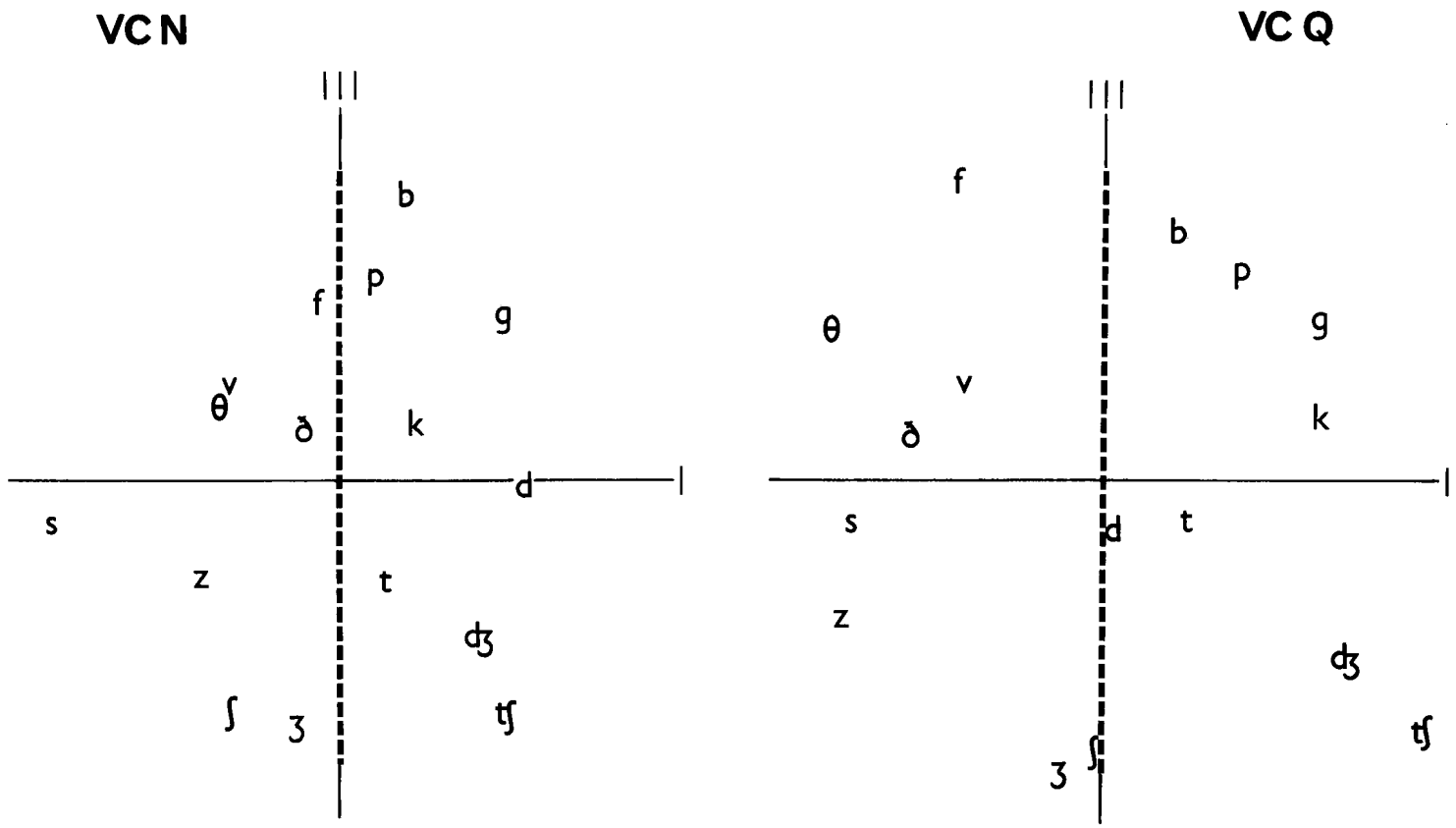

Figure 9. Plane formed by dimensions I and III in three-dimensional configuration for VCN and VCQ data. Dotted lines separate stops from fricatives (I).

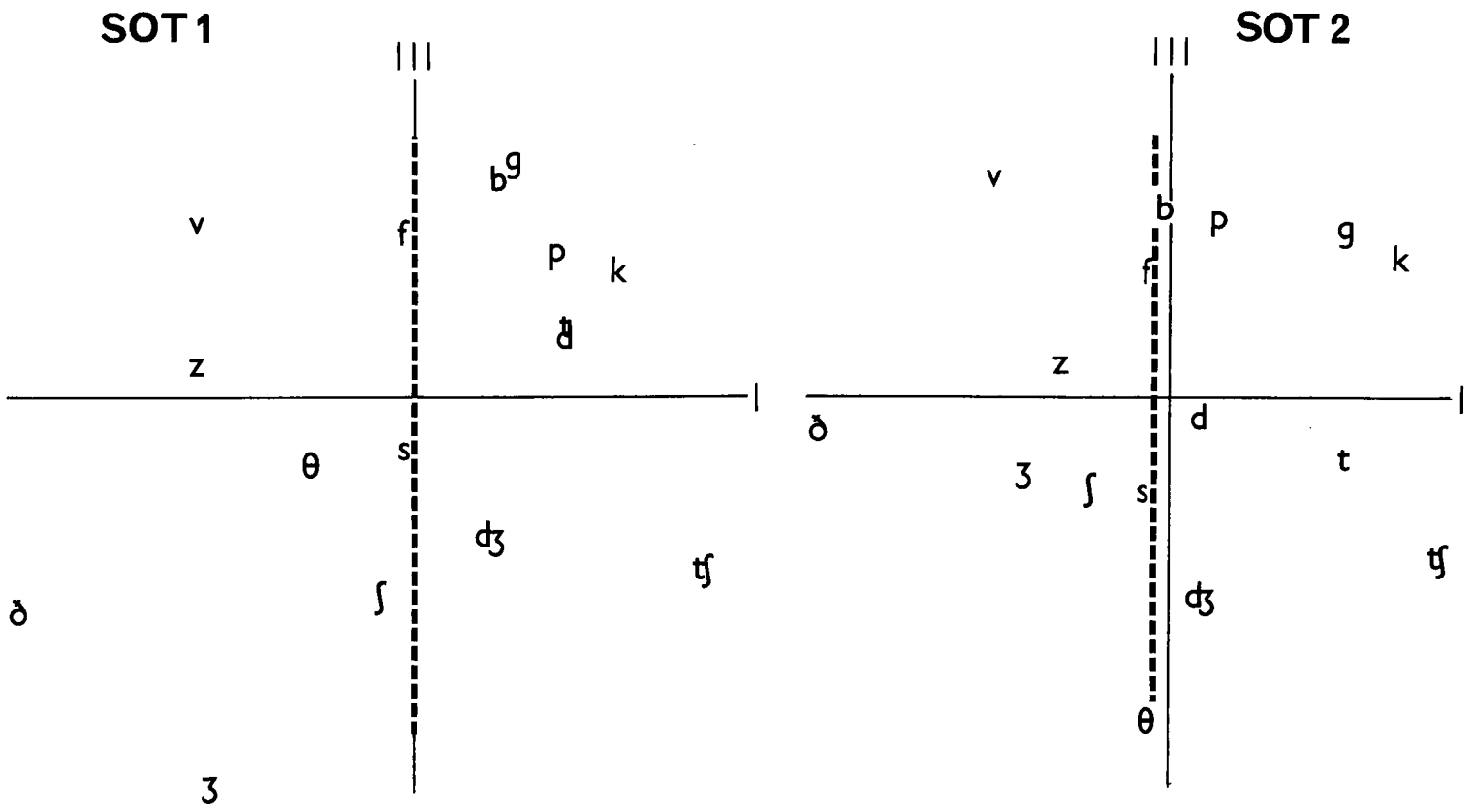

Figure 10. Plane formed by dimensions I and III in three-dimensional configuration for SOT1 and SOT2 data. Dotted lines separate stops from fricatives (I).

portions of /tʃ/ and /dʒ/ are very similar acoustically to /ʃ/ and /ʒ/. There is no comparable acoustic overlap between the other stops and any of the fricatives in the data set. This may account, then, for proximity of the affricates to the fricatives in syllable-initial position.

Figures 8 through 10 show the positions of the 16 consonants in the I-III plane, again for CV, VC and SOT types of data, respectively. For all data sets, dimension III divides the consonants into groups according to place of articulation. However, the particular grouping varies across the three types of data. Moreover, these differences are reliable, in the sense that the groupings within a data set type are the same. For CVs there are three groups, ordered from one end of dimension III to the other - labials, dentals and alveolars form one group, velars the second and palato-alveolars the third. VCs also show three groups, with different memberships. Labials, dentals and velars form one group, alveolars the second and palato-alveolars the third. For SOTs there seem to be only two reliable groups -- labials and velars against the rest. It is not clear how to account for these differences. The grouping of dentals, alveolars, palato-alveolars vs. labials and velars in production suggests the possibility of some articulatory parameter associated with raising the tip of the tongue. This same grouping, however, can also be made acoustically -- on the basis of the height of the second formant (high for dentals, alveolars and palato-alveolars, lower for labials and velars). The emergence of this grouping in SOTs might, then, be viewed as evidence of perceptual factors in speech errors. However, the failure of this grouping to appear for the CV perceptual data casts doubt on this latter hypothesis, particularly since the vast majority of speech errors (approximately 80%) involve syllable-initial consonants.

Reliability

In order to assess the reliability of the dimensions extracted in the analysis, (Pearson) correlation coefficients were calculated between data sets for the loadings of the consonants on a given dimension. Results for Dimension I are shown in the left-hand column of Table 3. The upper part of the table shows the correlation of the two replications within a given data type. The lower part shows correlations between data sets of different types. The significance levels associated with these correlations is shown in the second column from the left. As can be seen, the correlations are all highly significant[2], except for one marginal case -- SOT2 with VCN ($p<.011$). These correlations indicate that there is some reliable component of variance in dimension I, both within and across types of data. Similar calculations were performed for dimensions II and III and the results are shown in the left-hand columns of tables 4 and 5, respectively. The results show the same basic pattern as for dimension I-- highly significant correlations both within and across different types of data. The only exception to this pattern is in the correlation between the SOT and CV data sets on dimension III,

Table 3. Pearson correlation coefficients (r) and associated significance levels (p) for comparison of Dimension I values across pairs of data sets. Correlations are given both before (Zero-order) and after (Partial) controlling for categorical features by means of partial correlation.

Data Set Pair	Zero-order		Partial	
	r	p	r	p
	Within Data Type			
CVN-CVQ	.979	.001	.892	.001
VCN-VCQ	.831	.001	.508	.055
SOT1-SOT2	.933	.001	.622	.021
	Across Data Type			
CVN-VCN	.767	.001	.109	.375
CVN-VCQ	.729	.001	.117	.366
CVQ-VCN	.745	.001	-.062	.428
CVQ-VCQ	.709	.001	.157	.322
SOT1-CVN	.701	.001	-.101	.384
SOT1-CVQ	.678	.002	-.092	.394
SOT2-CVN	.653	.003	-.151	.329
SOT2-CVQ	.647	.003	-.053	.439
SOT1-VCN	.626	.005	.298	.186
SOT1-VCQ	.771	.001	-.284	.199
SOT2-VCN	.566	.011	.036	.459
SOT2-VCQ	.734	.001	-.440	.088

Table 4. Pearson correlation coefficients (r) and associated significance levels (p) for comparison of Dimension II values across pairs of data sets. Correlations are given both before (Zero-order) and after (Partial) controlling for categorical features by means of partial correlation.

Data Set Pair	Zero-order		Partial	
	r	p	r	p
	Within Data Type			
CVN-CVQ	.973	.001	.882	.001
VCN-VCQ	.934	.001	.715	.007
SOT1-SOT2	.885	.001	.658	.014
	Across Data Type			
CVN-VCN	.858	.001	-.477	.069
CVN-VCQ	.783	.001	-.364	.135
CVQ-VCN	.819	.001	-.443	.086
CVQ-VCQ	.786	.001	-.289	.194
SOT1-CVN	.859	.001	.631	.019
SOT1-CVQ	.799	.001	.502	.058
SOT2-CVN	.845	.001	.418	.100
SOT2-CVQ	.763	.001	.166	.313
SOT1-VCN	.626	.005	-.219	.259
SOT1-VCQ	.476	.031	-.346	.148
SOT2-VCN	.730	.001	.024	.472
SOT2-VCQ	.590	.008	-.107	.377

Table 5. Pearson correlation coefficients (r) and associated significance levels (p) for comparison of Dimension III values across pairs of data sets. Correlations are given both before (Zero-order) and after (Partial) controlling for categorical features by means of partial correlation.

Data Set Pair	Zero-order		Partial	
	r	p	r	p
	Within Data Type			
CVN-CVQ	.986	.001	.891	.001
VCN-VCQ	.973	.001	.905	.001
SOT1-SOT2	.724	.001	.061	.430
	Across Data Type			
CVN-VCN	.800	.001	.546	.041
CVN-VCQ	.807	.001	.537	.044
CVQ-VCN	.808	.001	.655	.014
CVQ-VCQ	.824	.001	.599	.026
SOT1-CVN	.536	.016	-.384	.121
SOT1-CVQ	.563	.012	-.184	.294
SOT2-CVN	.410	.057	-.466	.074
SOT2-CVQ	.386	.070	-.653	.015
SOT1-VCN	.794	.001	-.012	.486
SOT1-VCQ	.780	.001	.084	.403
SOT2-VCN	.662	.003	-.544	.042
SOT2-VCQ	.573	.010	-.688	.010

place of articulation. As noted above, the groupings on this dimension differ across data set type, and this is borne out by the lowered correlations for those comparisons.

It was hypothesized in the introduction that there would be reliable noncategorical variance along a particular dimension within perception or production, but not when comparing a perceptual dimension with a production dimension. This could be tested using the current data by examining the correlations within and across data types, after taking account of the variance contributed by categorical features. (This analysis also examines the common variance between syllable-initial and syllable-final perceptual dimensions, of course). The variance due to categorical features was taken into account by means of partial correlation. The technique can be described by means of an example: the correlation between CVN and CVQ on dimension I is .979. Suppose we want to know how much of this correlation is due to other things beside the fact that both data sets divide the consonants into the same groups -- stops and fricatives. One way to do this is to correlate each of these dimensions with a categorical variable that has the value 1 for all stops and 0 for all fricatives. It is then possible to remove from the CVN and CVQ loadings the variance that correlates with this categorical variable. We then have two sets of residuals that have zero correlation with the categorical variable. We can then calculate the correlation coefficient for this pair of residuals -- this determines whether they share any variance that is *not* predictable on the basis of the categorical variable.

The correlations in Tables 3 through 5 were recalculated, partialling out the effects of categorical features in the manner described above. For each dimension, the whole set of categorical features relevant to this subset of consonants was partialled out simultaneously. This was done rather than partialling only the single categorical features most obviously related to the dimension in question, for example [± continuant] for dimension I, and [± voice] for dimension II. As noted in the introduction, it is possible that within group variation on a particular dimension might be due to categorical features other than the major one correlated with a given dimension. It is for this reason that all categorical features were partialled out simultaneously. Five features relevant to the differentiation of this subset of consonants were chosen. These a are primarily the Chomsky and Halle (1968) features relevant to distinguishing these consonants -- [± anterior], [± coronal], [± voice], [± continuant]. As a fifth feature, [± sibilant] was chosen, rather than Chomsky and Halle's [± strident]. (The difference is that /f/ and /v/ are [+ strident] but [- sibilant].) The values of the sixteen consonants on these five features are shown in Table 6.

The results of the partial correlations are shown in the right-hand columns of Tables 3, 4 and 5. Examination of the within data type correlations indicates that there are still large corre-

Table 6. Representation of consonants in terms of features used for partial correlation analysis.

	continuant	voice	sibilant	anterior	coronal
p	0	0	0	1	0
t	0	0	0	1	1
k	0	0	0	0	0
b	0	1	0	1	0
d	0	1	0	1	1
g	0	1	0	0	0
f	1	0	0	1	0
θ	1	0	0	1	1
s	1	0	1	1	1
ʃ	1	0	1	0	1
v	1	1	0	1	0
ð	1	1	0	1	1
z	1	1	1	1	1
ʒ	1	1	1	0	1
tʃ	0	0	1	0	1
dʒ	0	1	1	0	1

lations for each of the three data types. Thus, we conclude that there is reliable variance on the dimensions that cannot be accounted for on the basis of the five categorical features partialled out. (The major exception is that the two speech error sets do not correlate on dimension III, after partialling out the five categorical features). Looking at the cross-data type partial correlations reveals a different pattern of results. For dimensions I and II almost all but two of the 24 cross-data type correlations become very small or negative. Thus for those two dimensions, the common variance across data set type can be adequately accounted for solely on the basis of the five categorical features used. For dimension III, however, there are significant residual correlations between the two types of perceptual data. Thus, for this dimension, there are some perceptual processes common to syllable-initial and final positions that cannot be accounted for solely on the basis the five categorical features employed. At least for dimensions I and II we can make the following conclusions:

(1) There is reliable variance in the dimensions of perception and the dimensions of production that cannot be attributed to categorical features.

(2) The overlap between dimensions of perception and production, and between dimensions of syllable-initial and syllable-final perception can be exhaustively described by categorical features.

DISCUSSION

The results presented suggest the following conception of the role of features in speech perception and production. For any particular kind of perceptual or production behavior, there is some small number of dimensions along which English consonants reliably vary. These dimensions have the effect of dividing consonants into groups that coincide with the categorical features used in phonological analysis. The nature of the dimensions themselves varies across the different kinds of behaviors, at least for the data that were examined here -- perception of CV syllables, perception of VC syllables and spontaneous speech errors. For two of three dimensions we extracted (i.e. voicing, stop/fricative) what is common to the dimensions across different kinds of data is limited to the division of the set of consonants into groups corresponding to categorical features. This, then, provides rather strong evidence for the importance of categorical features in the internal representation of speech -- they recur in three different kinds of behaviors, even though the continuous dimensions that correlate with them, differ from one another.

Some consideration needs to be given to the *way* in which the dimensions differ across the three types of behavior examined above. Note that there are significant residual correlations

between the two data sets of a given type, even after partialling out the effect of categorical features. Thus, there is some reliable, non-categorical information within a particular data type. However, there is no correlation between this reliable non-categorical information across data types, even though there is a good correlation between data set types with respect to categorical information. How can the different reliable properties of these data set types be accounted for? The potential sources of the differences between different data types can usefully be divided into two categories discussed below: (1) Differences due to inherent differences in the processes involved in the behaviors. (2) Differences due to the particular properties of the experimental or observational situations.

(1) Inherent differences. Differences between CVs and VCs might be due to the fact that the important acoustic cues for various features differ from initial to final position. For example, vowel length is an important cue for voicing in final position, but syllable-initially VOT is more important. As another example, the difference in acoustic cues was used to explain the difference in how affricates are represented in CVs and VCs. Presence of a silent interval may be used as a cue for stops in final position, but not in (utterance-) initial position. Similarly in comparing production and perception confusions, we are comparing acoustic and auditory properties of the consonants with properties relevant to articulation (in the broad sense). There is no reason to expect there to be a simple one-to-one relationship between those articulatory and acoustic properties.

In general, there is a welter of articulatory and acoustic variables that could be appealed to in order to account for differences in the data. A study specific to this point would be required to determine which set of variables exhausts the reliable variation within each type of data.

(2) Situational Effects. Van den Broecke (1976), in a review of the literature on perceptual features, has shown that there are substantial differences in the results of experiments using different tasks and number of stimuli. Differences in results include differences in the number of important features and in their relative weights. The dimensions used by a listener in any task will invariably be influenced by what (s)he is called upon to do and the strategies employed to do it. While the differences between CVs and VCs analyzed here ought to be minimal from this point of view, differences between this experimental situation and spontaneous speech errors would be expected to be substantial. Thus, some of the differences between types of data may be, themselves, uninteresting. The fact of these differences makes the good correlation with categorical features even more impressive, however.

As noted in the results, there are significant correlations between CVs and VCs on the place of articulation dimensions, even

after removing effects due to categorical features. At least a partial explanation for these correlations can be provided. Examination of the plots of CVN, CVQ, VCN and VCQ on dimension III (Figs. 8 and 9) shows that all four data sets have labials at one extreme end and palato-alveolars at the other end. The CV and VC sets differ, however, in the distribution of consonants between the extremes. Velars are represented as closer to the labials in VCs and closer to the palato-alveolars in CVs. With the particular categorical features used in the partial correlation analysis, there was no way to extract the commonality of the two data types with respect to labials and palato-alveolars. This is because the same features ([$\pm$ ant], [$\pm$ cor]) must be used to distinguish between velars and alveolars (for which the two data types differ) as well as to distinguish between labials and palato-alveolars. To take account of this, another partial correlation analysis was performed, using a different feature set. The new feature set used four separate features for distinguishing place of articulation: [$\pm$ labial], [$\pm$ lingual], [$\pm$ dorsal], [$\pm$palatal]. Using this feature set, the cross data type correlations no longer reached significance, after partialling out the effect of categorical features. This confirms the explanation suggested above. However, these correlations were still higher than most obtained for the cross data type comparisons. Thus, there may be some common variance between CVs and VCs that is not explained by categorical features, regardless of what categorical features are used. As an example of this, /v/ had a consistently smaller loading on dimension III than the other labials, for all perceptual data sets (but not for speech errors). Why this is true is unclear; it may reflect some artifact of the particular experimental situation.

Finally, the above-outlined role of features in speech perception and production has implications for active theories of speech perception. Such theories (Neisser, 1967; Stevens and Halle, 1967) postulate that speech perception involves the matching of some stored representation of a stretch of the auditory signal with a corresponding articulatory representation that is generated internally. A perplexing problem that faces such models is to determine the level at which this matching takes place. For example, Stevens and Halle (1967) have suggested that the internal articulatory signal is transformed into an auditory one, and that matching takes place at an auditory-memory level. The results of the present analysis suggests that matching could only take place at the level of categorical features, since only at this level are the internal representations (of consonants) the same for perception and production. However, given that perceptual analysis has proceeded to the point that a categorical representation is possible, the necessity for matching at all is undermined. Models for the process by which auditory representations are directly transformed into sets of categorical features seem to be required. Detailed analysis of the non-categorical components of perceptual dimensions may lead to some insight into this process.

NOTES

[1] I would like to thank Vicki Fromkin, Jean-Marie Hombert, Wendy Linker, Ian Maddieson, Lloyd Rice and Renee Wellin for various forms of assistance and encouragement. Awards for special heroism to Richard Harshman for teaching me to appreciate reliability; to Eric Holman for always having answers; to Peter Ladefoged for listening to all the details; to Marcel van den Broecke for providing challenges; and to Catherine Browman for reminding me at critical moments that even n-dimensional vector spaces have forests. Thanks to Stefanie Shattuck-Hufnagel for the use of her data. This research was supported by NIH.

[2] These significance levels should be regarded with some caution, however. It is not clear how appropriate the required independence and normality assumptions are for these comparisons.

REFERENCES

Carroll, J.D. and Chang, J.J. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young'decomposition. Psychometrika 35. 383-419.

Carroll, J.D. and Wish, M. 1974. Multidimensional perceptual models and measurement methods. In E.C. Carterette and M.P. Friedman (eds). Handbook of Perception 2. New York: Academic Press. 391-447.

Chomsky, N. and Halle, M. 1968. The sound pattern of English. New York: Harper and Row.

Comrey, A.L. 1973. A first course in factor analysis. New York: Academic Press.

Cooper, W.E. 1975. Selective adaptation to speech. In F. Restle, R.M. Shiffrin, N.J. Castellan, H. Lindman, and D.B. Pisoni (eds). Cognitive theory, Vol. I. Hillsdale, N.J.: Lawrence Erlbaum Associates. 23-54.

Fant, G. 1967. The nature of distinctive features. In To honor Roman Jakobson: Essays on the occasion of his seventieth birthday. The Hague: Mouton. 634-642.

Fromkin, V.A. (ed). 1973. Speech errors as linguistic evidence. The Hague: Mouton.

Gandour, J.T. and Harshman, R.H. 1977. Cross-language differences in tone perception: a multidimensional scaling investigation. Manuscript submitted to Language and Speech.

Harshman, R.H. 1970. Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis. UCLA working papers in phonetics 16.

Ingram, J.C. 1975. Perceptual dimensions of phonemic recognition. PhD dissertation. University of Alberta. Edmonton, Alberta.

Klatt, D. 1968. Structure of confusions in short-term memory between English consonants. J. Acoust. Soc. Amer. 44. 401-407.

Kruskal, J.B. 1964. Multidimensional scaling by optimizing goodness of fit to a monmetric hypothesis. Psychometrika 29. 1-27.

Kruskal, J.B., Young, F.W. and Seery, J.B. How to use KYST, a very flexible program to do multidimensional scaling and unfolding. Unpublished manuscript. Bell Laboratories. Murray Hill, New Jersey.

Ladefoged, P.N. 1975. A course in phonetics. New York: Harcourt, Brace and Jovanovich.

Lisker, L. and Abramson, A. 1964. A cross-language study of voicing in initial stops: acoustical measurements. Word 20. 384-422.

Luce, D. 1959. Individual choice behavior. New York: Wiley.

MacKay, D. 1970. Spoonerisms: The structure of errors in the serial order of speech. Neuropsychologia 8. 323-350.

Miller, G. and Nicely, P. 1955. An analysis of perceptual confusions among English consonants. J. Acoust. Soc. Amer. 27. 338-352.

Neisser, U. 1967. Cognitive psychology. New York: Appelton.

Peters, R.W. 1963. Dimensions of perception for consonants. J. Acoust. Soc. Amer. 35. 1985-1989.

Shattuck, S.R. 1975. Speech errors and sentence production. PhD dissertation. Massachusetts Institute of Technology. Cambridge, Massachusetts.

Shepard, R.N. 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. Psychometrika 27. 125-140, 219-246.

Shepard, R.N. 1972. Psychological representation of speech sounds. In E.E. David and P.B. Denes (eds). Human Communication: A unified View. New York: McGraw-Hill. 67-113.

Shepard, R.N. 1974. Representation of structure in similarity data: problems and prospects. Psychometrika 39. 373-421.

Singh, S. 1966. Cross-language study of perceptual confusion of plosive consonants in two conditions of distortion. J. Acoust. Soc. Amer. 40. 635-656.

Singh, S. 1975. Distinctive feature: A measurement of consonant perception. In S. Singh (ed). Measurement procedures in speech, hearing and language. Baltimore: University Park Press. 93-155.

Singh, S. and Black, J.W. 1966. Study of twenty-six intervocalic consonants as spoken and recognized by four language groups. J. Acoust. Soc. Amer. 39. 372-387.

Singh, S., Woods, D.R. and Becker, G.M. 1972. Perceptual structure of 22 prevocalic English consonants. J. Acoust. Soc. Amer. 52. 1698-1713.

Stevens, K.N. and Halle, M. 1967. Remarks on analysis by synthesis and distinctive features. In W. Walthen-Dunn (ed). Models for the perception of speech and visual form. Cambridge, Massachusetts: MIT Press. 88-102.

Studdert-Kennedy, M. 1976. Speech perception. In N.J. Lass (ed). Contemporary issures in experimental phonetics. New York: Academic Press. 243-293.

Torgerson, W.S. 1958. Theory and methods of scaling. New York: Wiley.

Van den Broecke, M.P.R. 1976. Hierarchies and rank orders in distinctive features. Van gorcum.

Wang, M.D. and Bilger, R.C. 1973. Consonant confusions in noise: A study of perceptual features. J. Acoust. Soc. Amer. 54. 1248-1266.

Wish, M. and Carroll, J.D. 1974. Applications of individual differences scaling. In E.C. Carterette and M.P. Friedman (eds). Handbook of Perception 2. New York: Academic Press. 449-491.

Chapter 2:

Perceptual salience of stressed syllables

Perceptual Salience of Stressed Syllables

Louis Goldstein

INTRODUCTION

Much of the research in speech perception has been devoted to understanding the processes involved in the perception of isolated CV and VC syllables, in particular to the question of how listeners transform acoustic representations of these syllables into phonetic representations specified in terms of segments and features. However, there has been very little attempt to specify how the acoustic representation of the speech input is used by listeners in recognition of higher-level linguistic units, ie. words, phrases and sentences. Most of the investigations that have been done in this latter area have looked primarily at how prosodic acoustic information is used perceptually, but not at how prosodic and segmental acoustic information interact in the process of word and phrase recognition. For example, prosodic information has been shown to provide cues as to syntactic structure of sentences, even when the segmental acoustic information is severely distorted by spectral rotation (Blesser, 1969), or is replaced altogether by humming (Svensson, 1974). Experiments employing synthesized versions of hummed speech have shown that intonation contour (Collier and 't Hart, 1975) and durational pattern (de Rooij, 1976) can both be used, independently of the other, as cues to the location of breaks between major syntactic constituents.

In the present paper, we are interested in a different set of prosodic effects; in particular, how prosodic acoustic information (stress) interacts with segmental acoustic information in the recognition of words and phrases. We assume the following general framework for describing the process of speech recognition. At least two components seem to be involved. One component is a mechanism for sampling the acoustic signal over some time window and making a gross and incomplete phonetic analysis. The second component is a higher-level decision mechanism that determines what word or phrase has been perceived. The input to the decision-making mechanism includes not only the incomplete phonetic analysis of the first component, but also phonological, syntactic and semantic context and expectations. Explicit recognition models, involving such components have been proposed by Pisoni and Sawusch (1975), for example, and have been in use in automatic speech understanding systems (eg., Lesser et. al., 1974). Given this framework it is reasonable to ask whether different types of segments or different types of syllables can vary systematically in the degree to which they constrain the choices of the decision-making component. Such differences will be referred to as differences in the

perceptual salience of the units in question. Over and above differences between types of segments, a likely hypothesis is that differences in perceptual salience can be found to be associated with prosody. In particular, it would be predicted that stressed syllables are more perceptually salience than unstressed ones.[1]

There are at least two possible factors that could contribute to differences in perceptual salience. The first factor will be referred to as inherent phonetic ambiguity. The information gained in the acoustic sampling and partial phonetic analysis could be inherently more detailed for some kinds of phonetic units than for others. Such low-level differences can be measured independently of the word and phrase recogniton situation by examining differences in confusability between the relevant units in a nonsense CV or VC recognition task. Since context and lexical structure does not enter into such tasks (although syllabic response bias might, see Goldstein, 1977), reliable differences in confusability (after removing effects of response bias) would reflect inherent phonetic ambiguity. Relative amount of energy in the signal is one determinant of phonetic ambiguity, since syllables presented to listeners in lower S/N ratios are generally more confusable. Thus, one would expect unstressed syllables to be more phonetically ambiguous than stressed ones, as they are generally shorter and of lower amplitude. On the basis of energy considerations, we might also expect consonants to be more phonetically ambiguous than vowels. However, as is well-known, cues for consonant recognition are contained in both vowel and consonant portions of the acoustic signal. The relevant experiments, directly comparing consonant and vowel confusions in the same set of nonsense stimuli have not been done. Moreover, effective S/N ratio is probably only one determinant of confusability. In particular, there may be specialized feature or property detectors for consonants (Cooper, 1975; Stevens and Blumstein, 1975) that may be quite sensitive to low amplitude signals of the proper form.

The second factor that contributes to perceptual salience is the relative attention that is directed toward successive portions of the input stream. This factor would also seem to support the relative perceptual salience of stressed syllables. Experimental evidence indicates that attention is focussed on stressed syllables in the perception of sentence-sized units. Cutler (1976) employed the phoneme-monitor technique to provide evidence for this point. In this task, first used by Foss (1969), subjects are required to listen to sentences for comprehension, and at the same time, to attend for the occurrence of a particular word-initial target phoneme (usually a consonant). They are instructed to press a button just as soon as they hear the target phoneme, and reaction time is measured. Cutler found that RT for phoneme targets in highly-stressed syllables was faster than for targets in low-stressed syllables. Moreover, using an ingenious control condition, she was able to demonstrate that the difference in RT was not due solely to acoustic differences in the target syllables themselves. This control condition was created as follows: A clause containing the target word was recorded in three different sentential contexts. The contexts produced different stress patterns in the

common clause. In one of these versions, the target word was contrastively stressed, in one it was normally stressed, and in the other it was unstressed (or had low stress). The target word was spliced out of the normally stressed version and was inserted into the other two versions. RT to the target was faster in the stressed version than in the unstressed version, even though the target word itself was acoustically identical in the two cases. Cutler concludes from this faster RT that stressed syllables are processed earlier than unstressed ones, as a result of listeners' focussing attention on the portions of the sentence where stressed syllables are expected.

Evidence in favor of stressed syllables as attentional foci was also presented by Shayne and Gass (1976). They used a version of the click paradigm (Ladefoged and Broadbent, 1960; Fodor and Bever, 1965), in which the subject is required to listen to a sentence that has an extraneous noise superimposed somewhere on it, and to report where in the sentence they heard the noise. (Shayne and Gass used a tone 'beep' rather than a click). They varied the position of the beep with respect to the position of the main stress in a number of sentences. They found that subjects were most accurate in localizing the beep when it coincided with the main stress of a sentence; and that subjects tended to report the beep on the stressed syllables as often, or more often, than on the correct syllable when they did not coincide. This result alone does not necessarily demonstrate that attention is actively focussed on the stressed syllable during perception, since the subjects could have associated the beep with the stressed syllable during some post-perceptual organization of the response. However, the result clearly supports the role of the main-stressed syllable as a kind of temporal anchor point. Taken together with the Cutler result, it seems likely that this anchoring plays an active role in sentence perception.

If stressed syllables are, indeed, more perceptually salient than unstressed ones, ie., if stressed syllables serve to constrain the choices made by the higher-level component more than unstressed ones, then this makes predictions about error patterns when words and phrases are misperceived. In particular, we would expect that the misperceptions should involve unstressed syllables being in error more often than stressed ones. The experiment to be reported below attempts to test this hypothesis directly. In addition, the experiment was designed so that possible differences in perceptual salience between consonants and vowels could be tested.

A technique that can be used for generating mispercpetions was first described by Pickett and Pollack (1963) and Lieberman (1963). It involved splicing words and short phrases out of normally spoken sentneces. Pickett and Pollack found that such samples are very hard to identify, and that recognition is a function of the duration of the sample. Near perfect recognition was not achieved until samples

were 800 msecs long. At least part of the reason for the poor intelligibility of these segments is the absence of the particular context in which they were originally read. This has been demonstrated by Lieberman (1963), who found that intelligibility of excised segments was inversely related to the predictability (or redundancy) of the segments in the original sentences. This technique was also employed recently by LaRiviere and Winitz (1977). They spliced single-syllable CVC(C) words out of context. In other conditions, they included the vowel immediately preceding and following this word in the excised sample, to see if coarticulation cues improved intelligibility. No effect was found.

This technique was considered more appropriate for the current investigation than presentation of words and phrases in noise. The latter technique has often been used in word recognition experiments (eg. Savin, 1963; Fredriksen, 1971), but such experiments have generally tested word-level variables (such as word frequency). Since we are currently concerned with examining patterns of segmental errors in mispercpetions, it was considered inappropriate to use added noise, which might have the effect of selectively impairing the perception of particular segments. The excising technique involves no artificial noise; it generates confusions by withholding (from the higher-level component) the contextual information necessary to correctly identify these words.

METHOD

Stimuli

The stimuli chosen included 70 English words and short phrases. The stimuli were selected so as to be as phonetically balanced as possible, while at the same time controlling for word frequency. Frequency was controlled because we want the differences in intelligibility between stimuli to reflect, as much as possible, the phonetic properties of the words, rather than differences in expectation. All of the single word stimuli had frequencies of 1 or 2 per million in the Kucera and Francis (1967) word count. For phrase stimuli, at least one word in the phrase had a frequency of one or two per million (and thus the frequency of the phrase had to be somewhat lower). Low frequency words were chosen so as (hopefully) to maximize the number of errors. Generally, more errors are found in word recognition with low frequency words because listeners seem to choose the most common word that is consistent with their gross phonetic analysis of the signal(see Savin, 1963). Frequency was also found to effect error rate with the excised word technique (LaRiviere and Winitz, 1977).

For similar reasons (maximization of errors), stimuli included both single words and phrases, rather than just single words. With both words and phrases possible as responses, the number of response alternatives is greater than if the response had to be a single word. A greater number of alternatives would be expected to produce

a greater liklihood of error. Moreover, uncertainty as to the word-structure of a particular acoustic sequence more closely approximates the normal speech perception situation than does a situation in which the listeners know that they can expect a single word.

The stimuli were either one two or three syllables long. The 44 polysyllabic stimuli were either single words, or two- or three- word phrases. The 26 monosyllabic stimuli were, of course, all single words. The choice of stimuli was constrained by three principles. First, since we wanted to compare the number of errors in stressed and unstressed syllables, position of stress in the stimuli had to be varied systematically. Secondly, stressed and unstressed syllables were required to have comparable distributions of consonants and vowels, so that stressed-unstressed comparisons would not be confounded by differences among consonants or vowels. Finally, since a comparison of consonant and vowel errors was planned, the entire set of English consonants and vowels had to be represented in the stimuli. Consistent with these three principles, the details of the choice of the 70 stimuli were as follows:

The 26 monosyllabic stimuli were all stressed and each of the standard American English vowels (except [ɔɪ]) appeared twice. The initial consonant portions of these words were all single consonants. Each of the 22 English syllable-initial consonants (except [ð]) appeared once. The other five words began with a vowel. For polysyllabic stimuli, the single stressed syllable of each of the stimuli fell approximately an equal number of times in each syllable position. The initial consonant portion of all the syllables consisted of a single consonant, with all English (syllable-initial) consonants occurring at least once in stressed and once in unstressed syllables. All American English vowels (again except [ɔɪ]) occurred at least twice in stressed and twice in unstressed syllables. In addition, twenty of the unstressed syllables included a reduced vowel ([ə] or [ɪ]). All of the consonants except [p,θ,l,h] occurred once in initial position of these reduced syllables. Two began with vowels. Syllable-final consonants were left uncontrolled in the entire stimulus set, and include clusters as well as single consonants. Therefore, all comparisons between stressed and unstressed syllables will be limited to syllable-initial consonants and vowels. The 70 stimuli are presented in the Appendix, organized by number of syllables, word structure and stress pattern.

Sentences were constructed in which the 70 stimuli were embedded, with the constraint that the last phoneme *before* the stimulus, and the first phoneme *after* the stimulus were voiceless stops. This facilitated the editing process. The sentences were constructed in such a way as to effect the desired stress patterns on the stimuli. Contrastive stress, for example, was occasionally required.

Procedure

The sentences were recorded by a speaker of general American English, whose pronunciation is similar to that used by network television news announcers. He was naive to the purpose of the experiment. For cases in which the constructed sentence did not effect the desired stress pattern, new sentences were constructed, and the speaker returned for another session. This was continued until all the stimuli had the desired stress patterns.

Each sentence on the high-quality recording was low-pass filtered at 4.5 KHz and then digitally sampled at 12 KHz by a PDP-12 computer program for waveform sampling. The stimulus segments were isolated and stored using a program for waveform viewing and editing (written by Lloyd Rice), and then the 70 stimulus segments were rerecorded onto analog tape in random order with eight seconds between each stimulus. The tape was carefully transcribed by a phonetician, whose transcription agreed with that of the author.

The stimulus tape was played, over a loudspeaker at a comfortable listening level, to an introductory linguistics class. Students were told that the stimuli were words and phrases spliced out of context, and were instructed to write down some English word or phrase for each stimulus; they were told to guess if they were unsure.

Listeners' written responses were assigned a phonemic transcription by the author, using Southern Californian English as a reference dialect. In the few cases in which a listener's response was not an English word or phrase, conventions of English orthography were used to assign a transcription. In cases of orthographic ambiguity, the transcription chosen was the one that was closest to the stimulus. Responses of 12 subjects were analyzed.

Scoring

In order to determine which phonemes are reported correctly and which are in error, the individual phonemes of a stimulus have to be matched to the phonemes of the response. This can be a complex process, given that the stimulus and response can differ in number of phonemes and even number of syllables. The matching was therefore accomplished using an interactive computer program developed for this purpose by Catherine P. Browman and the author. The program attempts to do as much of the matching as possible automatically. The only user-intervention is at the very earliest stage in which each of the syllables of the stimulus have to be matched to syllables in the response. (This procedure is very difficult to automate). For the analysis of the present data, syllables were matched as follows: if the stimulus and response had the same number of syllables, the syllables were matched in simple left-to-right order, first syllable of the stimulus with first syllable of the response, etc. If the response involved deletion or addition of a syllable (which was rare in this experiment), syllables were matched so as to maximize

the number of phonemes in common to the syllables being matched, with the constraint that the linear order of the stimulus syllables was identical to the linear order of the syllables that were matched to them. Thus, syllable metatheses were not considered a possible error type, and none of the data suggested that this type of error had occurred.

Once the syllables were matched, a set of algorithms proceeded to automatically match the phonemes of the stimulus with those of the response. The details of the algorithms are quite involved and they are described in detail in Browman (1977a) and Browman and Goldstein (forthcoming). Briefly, the processes can be summarized as follows: The vowels in matched syllables are matched to each other. The program then attempts to match syllable-initial consonants to initial consonants in the matching response syllable, first by looking for identical phoneme matches, then by attempting to optimize feature agreement in matched phonemes, and finally by resorting to linear order in the case of feature ties. The same procedure applies to syllable-final consonants. There are a number of additional passes through the data to attempt to find matches for consonants of inserted and deleted syllables, and to look for matches across syllable boundaries, and to look for consonant metatheses across vowels.

RESULTS

For each phoneme in a given stimulus that is compared to a given response, the output of the matching program indicates one of the following possibilities: Either the phoneme is matched to the identical phoneme in the response, or the phoneme is matched to some different phoneme in the response, or the phoneme has no match at all in the response. These three categories are referred to as correct reports, confusions, and deletions, respectively. Confusions and deletions will be referred to together as errors. Table 1 shows the percentages of stimulus phonemes that were confused and deleted, separately for consonants and vowels, and separately for stressed and unstressed syllables. These are the combined results from all subjects and all stimuli. The overall error percentage is higher for unstressed syllables than for stressed syllables, both for initial consonants and for vowels. These differences are both significant using a Wilcoxon matched-pairs, signed-ranks test, in which each subject contributes a pair of error percentages ($p<.005$ and $p<.025$, respectively, one-tailed). Examining the confusions and deletions separately, we can see that they both show the same trend as the overall error percentages.

The above differences confirm the prediction that there would be more errors in unstressed syllables than in stressed ones. There is, however, a possible objection to this analysis. Twenty-six of the stimuli were monosyllabic words, and as such they included only a stressed syllable. Pickett and Pollack (1963) have shown (as noted above) that recognition of excised segments is a function of

Table 1. Percentage of syllable-initial consonants and vowels that were deleted or confused in in subjects' responses. Deletions and confusions are combined as total errors. Results represent all stimuli and all subjects.

	Unstressed Syllables		Stressed Syllables	
	C	V	C	V
% Deletions	6.8	4.7	2.7	2.6
% Confusions	20.6	19.2	19.0	17.3
% Total Errors	27.4	23.9	21.7	19.9

Table 2. Percentage of syllable-initial consonants and vowels that were deleted or confused in subjects responses. Deletions and confusions are combined as total errors. Results include only polysyllabic stimuli for all subjects.

	Unstressed Syllables		Stressed Syllables	
	C	V	C	V
% Deletions	6.8	4.7	3.8	2.2
% Confusions	20.6	19.2	15.3	17.9
% Total Errors	27.4	23.9	19.1	20.2

the duration of the excised segment. Since the monosyllabic stimuli are all shorter than the polysyllabic ones, this variable will confound the difference between stressed and unstressed syllables. Therefore, error percentages were calculated excluding the monosyllabic stimuli. Results are presented in Table 2 (which are, of course, the same as in Table 1 for the unstressed syllables). The pattern of errors is apparently very similar regardless of whether the monosyllabic stimuli are included or not. Differences between stressed and unstressed syllables are again significant ($p<.005$ for consonants, $p<.01$ for vowels, one tailed). Again, confusion and deletion trends are both in the same direction as the overall error percentages.

Differences between consonants and vowels, in terms of percent error, are somewhat more problematic. Examining Table 1, there is a tendency for there to be more errors on initial consonants than on vowels. However, this is only marginally significant for unstressed syllables ($p<.05$, two-tailed) and is not significant for stressed syllables (again using a Wilcoxon test). The problem with including the monosyllabic stimuli in the analysis does not arise for consonant-vowel comparisons, because monosyllabic stimuli include both consonants and vowels. Thus, properties of monosyllabic stimuli should not confound the consonant-vowel comparison. Yet, as examination of Table 2 shows, removing the monosyllabic stimuli does affect the relationship between consonant and vowels errors. The confusion percentage (but not the deletion percentage) is lower for consonants than for vowels when examining just the polysyllables. This is the only case examined in which the confusion and deletion trends differ, and is not at all clear how to account for this difference. Thus, the relative perceptual salience of consonants versus vowels cannot be simply determined from this experiment.

A further analysis of the data was undertaken to determine what factors were responsible for the difference in perceptual salience observed in this experiment between stressed and unstressed syllables. It will be shown that inherent phonetic ambiguity does have an effect on error rate in the current experiment, independently of differences between stressed and unstressed syllables. Therefore, it may contribute to the perceptual salience of stressed syllables over unstressed ones, as well.

We can demonstrate the effect of relative phonetic ambiguity by showing that in the present experiment, subjects tend to make more errors on those segments that produce more errors in a nonsense recognition task. The perceptual confusion data of Miller and Nicely (1955) was used to estimate the relative phonetic ambiguity of different consonants. The stimuli for that experiment consisted of /Ca/ syllables under different conditions of noise and filtering. Consonants included /p,t,k,f,θ,s,ʃ,b,d,g,v,ð,z,ʒ,m,n/. For the current purposes, we examine the experimental condition that most closely approximated the conditions of the present experiment, +12 db S/N ratio, low-pass filtered at 5KHz. For each of the consonants, (except for /ʒ/, which did not occur in the present experiment) the percent of stimuli incorrectly reported

was calculated. This error rate, as a function of consonant, is referred to as m(x). The same calculation was performed for each subject in the present experiment, based on the data from polysyllabic stimuli. The error rate for those 15 consonants was calculated separately for stressed and unstressed syllables: $p_s(x)$ and $p_u(x)$ respectively. To the extent that a subject tends to make more errors in the current experiment on the consonants that have high error rates in the Miller-Nicely data, then the quantities As and Au in (1) below will tend to be large:

$$(1) \qquad As = \frac{\sum_{x=1}^{N} p_s(x)\ m(x)}{\sum_{x=1}^{N} p_s(x)} \qquad Au = \frac{\sum_{x=1}^{N} p_u(x)\ m(x)}{\sum_{x=1}^{N} p_u(x)}$$

N = number of different consonants

If there were no relationship between m(x) and $p_s(x)$, then the observed As would tend to be the same as its value if all of the error rates in $p_s(x)$ were equal to the mean error rate, $\overline{p_s}$. Thus, the expected value of As, assuming no relationship of m(x) to $p_s(x)$ is:

$$E(As) = \frac{\sum_{x=1}^{N} \overline{p_s}\ m(x)}{\sum_{x=1}^{N} \overline{p_s}}$$

$$= \frac{\overline{p_s} \sum_{x=1}^{N} m(x)}{N\ \overline{p_s}}$$

$$= \frac{\sum_{x=1}^{N} m(x)}{N}$$

Thus, for each subject we can calculate the observed As and compare it to the value that would be expected if there was no effect of m(x) on $p_s(x)$. The same can be done for unstressed syllables. The values for As and Au, and the value expected on the basis of no relationship are shown in Table 3. Both As and Au were significantly greater than the value that would be expected on the basis of there being no

Table 3. Values of As and Au for each of the twelve subjects. As and Au represent the degree to which errors on stressed and unstressed syllables tend to show the same pattern across consonants as in nonsense syllable recognition (see text). Also given is the expected value of As or Au, given no relationship between error patterns in this experiment and in nonsense syllable recognition.

Subject	As	Au
1	.172	.239
2	.227	.205
3	.250	.255
4	.150	.248
5	.229	.290
6	.261	.194
7	.216	.230
8	.265	.306
9	.177	.275
10	.163	.225
11	.213	.178
12	.347	.301
Expected	.168	.168

relationship between m(x) and the error rates in this experiment. ($p<.025$ for As and $p<.001$ for Au, sign test).

The above result indicates that the relative phonetic ambiguity does effect the error rate for consonants in both stressed and unstressed syllables. Further evidence for the role of phonetic ambiguity in the word and phrase recognition task can be found by examining word position effects on the error rates for vowels. Oller (1973) has shown that vowels are longer in word-final syllables than in word-initial syllables, for nonsense words. Klatt (1975) has shown that in running discourse, stressed vowels are slightly longer word-finally than word-initially, and substantially longer if they are also phrase-final. Assuming that an increase in the duration of a vowel makes it less phonetically ambiguous, we would predict that vowels in word-final stressed syllables should be less ambiguous than vowels in word-initial stressed syllables. The overall error rates are 13.% and 30%, respectively for polysyllabic words. In testing these for significance, account must be taken of the fact that the same vowels do not appear in both initial and final syllables, since it has already been shown that the relative ambiguity of segments can effect error rate (at least for consonants). The relative phonetic ambiguity for vowels was estimated on the basis of the vowel confusion data of Strange et al, 1976, for the vowels (i,ɪ,ɛ,æ,ɑ,ʌ,ɔ,u). In the particular experimental condition chosen, vowels were presented in a C-C context (in which the consonants were stops) with tokens from an individual speaker presented together in a block. On the basis of the relative phonetic ambiguity of these vowels the expected error rate for vowels in word-initial syllables was calculated as in (3):

$$(3)\quad E_I(e) = \frac{\sum_{x=1}^{8} f(x)\ s(x)}{n}$$

where:

s(x) = the rate of errors for vowel x in Strange et al.

f(x) = the number of tokens of vowel x in word-initial syllables

n = total number of vowel tokens in word-initial syllables.

A similar expression of course can be calculated for the expected error rate for vowels in word-final syllables, $E_F(e)$.

For each subject, the actual error rates in initial and final syllables was calculated and each was divided by the expected error rate appropriate to that condition. The difference between these corrected error rates was significant, using a Wilcoxon matched-pairs, signed ranks test ($p<.01$). Thus, position of a vowel within a word does seem to effect the number of errors. This can be explained on

the basis of the fact that vowels in word-final syllables tend to be longer, and therefore, less phonetically ambiguous than the vowels in word-initial syllables.

The effect of relative phonetic ambiguity on the error rate in word and phrase recognition has been demonstrated, independently of the difference between stressed and unstressed syllables. It seems reasonable, therefore, that the stress effect is, at least partly due to differences in inherent phonetic ambiguity. In addition, we want to test whether attentional and higher-level processes contribute to this difference.

To help clarify the rationale of the following analysis, consider the following obviously exaggerated model of the role of stressed and unstressed syllables in word and phrase recognition. Let us suppose that higher-level decision processes weight stressed syllables very highly and that the choices made by these processes are constrained to be *completely* consistent with the information from the gross phonetic analysis associated with stressed syllables. This information will reflect the inherent phonetic ambiguity of these syllables. Let us suppose further that unstressed syllables receive very little weight, and choices are not at all constrained to be consistent with whatever information there is in the gross phonetic analysis associated with unstressed syllables. These two suppositions are an extreme version of the preference for stressed syllables.

The relative phonetic ambiguity of different phonemes can be assessed by reference to their relative confusability in a nonsense syllable recognition task. The model described above would predict that the relative error rates for different phonemes in stressed syllables of words and phrases should correlate with the relative phonetic ambiguity of these phonemes as measured in nonsense syllable recognition tasks. However, for unstressed syllables, no such correlation would be predicted, since word recognition is not assumed to be consistent with the phonetic analysis of unstressed syllables.

This model is clearly too extreme particularly since we have already seen that there is an effect of phonetic ambiguity on error rate. for *both* stressed and unstressed syllables. However, one could imagine a greater weight being assigned to the phonetic analysis of stressed syllables than to that of unstressed syllables in making lexical decisions. One might then expect to find a better correlation between word recognition and nonsense results (in terms of relative error rates for different phonemes) for stressed syllables than for unstressed ones.

In order to examine the pattern of relative error rates for stressed, unstressed and nonsense syllables graphically, the fifteen consonants of the Miller-Nicely experiment were divided into five classes and the percent error in each of these classes was calculated. Classes were: voiceless stops, voiced stops, voiceless fricatives, voiced fricatives (but not including [ʒ], since it did not occur in

in the current experiment), and nasals. The same calculation was made separately for stressed and unstressed syllable-initial consonants in the data from the current experiment. Only data from polysyllabic stimuli was used.

The relative error rates for the different phoneme classes in all three sets of data are shown in Fig. 1. As is clear from this figure, the stressed syllable error rates agree quite well with those from the Miller-Nicely experiment. The relative ordering of the obstruent classes is the same for both. The only difference is that nasals have a lower error rate than voiceless stops in the Miller-Nicely data, but have an error rate between voiceless and voiced stops in the stressed syllables of the current experiment. The pattern of errors for unstressed syllables does not seem quite as similar to the pattern for nonsense syllables, however.

To test the hypothesis that the results for stressed syllables are more similar to the results for nonsense than is the case for unstressed syllables, we compared the values of As and Au, for each subject. According to the above hypothesis, As should be greater than Au since As represents the degree to which the distribution of error rates for stressed syllables is similar to that for the nonsense results, and Au represents this similarity for unstressed syllables. However, as is clear from Table 3, this is not the case. For eight of twelve subjects, Au is actually larger than As. It is not clear how to reconcile this result with graphs of Fig. 1. Apparently, differences among consonants within each class used in Fig. 1 and differences among subjects may be substantial enough to account for the discrepancy. There is also a theoretical problem with this analysis, however. The quantities Au and As will tend to be maximized to the extent that $p_s(x)$ or $p_u(x)$ are high for those values of m(x) that are very high. The values of $p_s(x)$ and $p_u(x)$ for those consonants with a low m(x) will have little effect on the calculation of As and Au. Examining Fig. 1, we can see that voiced fricatives (and particularly /ð/, in the individual consonant data) have the highest error rate in the Miller and Nicely data. While the data for unstressed consonants doesn't seem to fit the *general* pattern of M/N there is a very large percent error for voiced fricatives, the class with the largest values of m(x). It has a larger percent error for this class then the corresponding error rate for the stressed consonants. This difference will contribute to Au being larger than As. Thus, it is not quite clear that Au and As provide the best measure of overall similarity of the error rate patterns. In the limiting case, if all of the errors for unstressed syllables were on voiced fricatives, or on /ð/ in particular, then Au would be even larger than it presently is. However, it is not clear that we would want to consider such a pattern of error rates to be more similar to the m(x) than it presently is. In any case, we cannot, on the basis of this analysis, confirm the hypothesis that stressed syllables are given more weight in higher-level decision-making than unstressed ones.

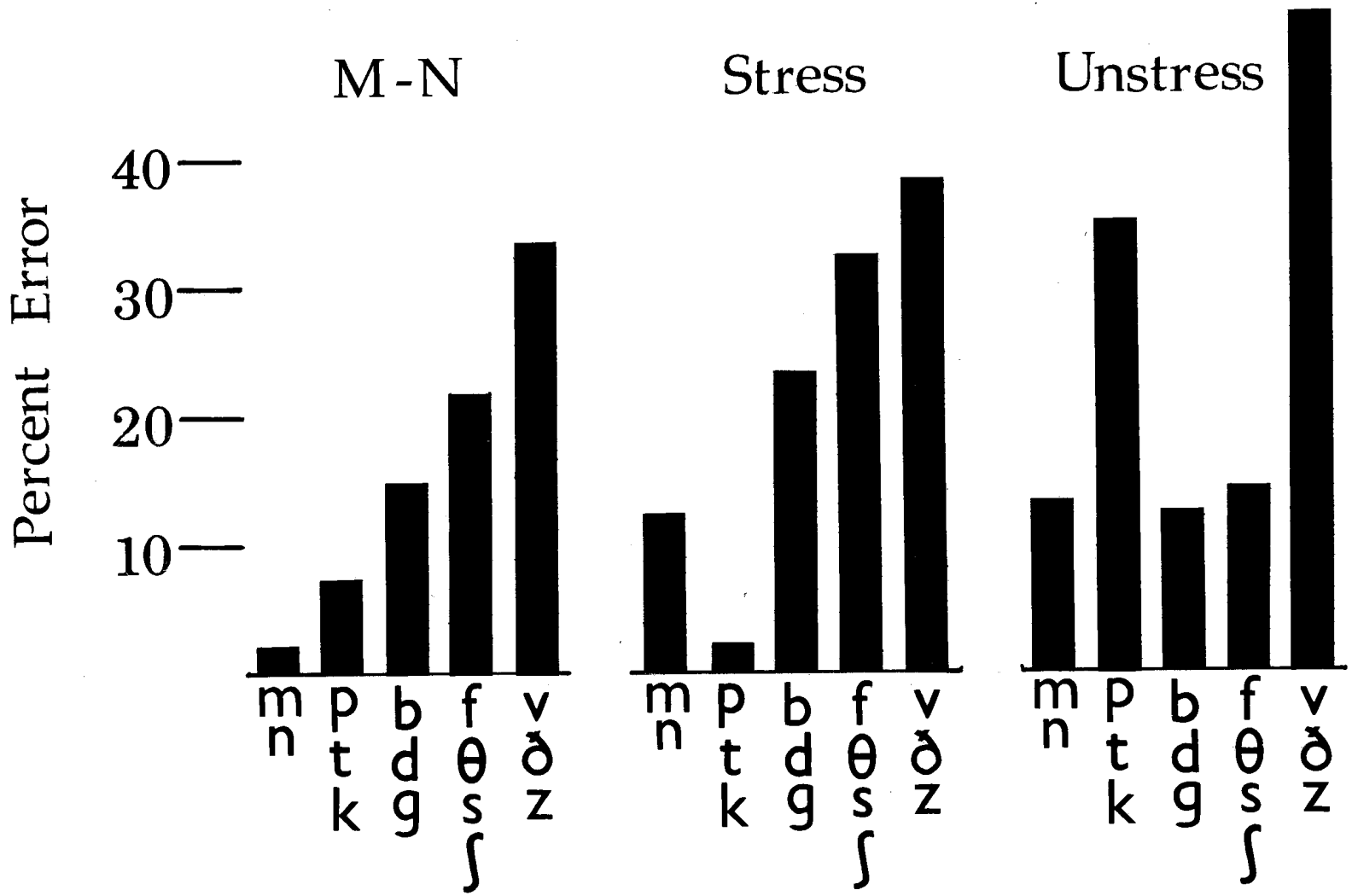

Fig. 1. Total percentage of stimulus consonants that are in error, shown separately for differerent classes of consonants. M-N represents Miller-Nicely data, Stress and Unstress data from this experiment.

DISCUSSION

We have demonstrated that stressed syllables are more perceptually salient than unstressed ones in the recognition of words and short phrases. We have suggested this effect is at least partly due to the greater phonetic ambiguity of unstressed syllables. If part of this effect were, in addition, due to a greater weight being given to the phonetic analysis of stressed syllables in the higher-level decision process, it is interesting to consider what mechanism might be responsible for such a differential weighting or attention. Cutler(1976) suggests, on the basis of her phoneme monitor RT results that stressed syllables, which can be identified on the basis of prosodic patterns, are processed earlier than unstressed ones (even if the unstressed ones precede). This is an interesting hypothesis, but phoneme monitor results can give only weak support for such a conclusion. It has been shown that phoneme-level decisions required in a phoneme-monitor task may be made *after* higher-level word decisions (see Rubin, Turvey and van Gelder, 1975). If the greater contribution of stressed syllables to the process of word recognition is to be explained on the grounds that they are processed earlier, we need to show this by some technique that measures processing that occurs *before* (or simultaneously with) word-level processing.

A possible technique for examining the hypothesis of earlier processing of stressed syllables as explanation for their perceptual salience might be as follows. Word recognition *latency* could be measured in a word and phrase task, similar to the one used in this investigation, in addition to error rate. If stressed syllables are processed earlier than unstressed ones, then the longer a listener takes to identify a word (or phrase), the greater the liklihood that some substantial processing of unstressed syllables occurs. This would predict that the difference in error rate between stressed and unstressed syllables would decrease as the latency for word (or phrase) recognition increased. Such an experimental approach is currently being explored.

Some comments need to be made about the results of comparing the relative perceptual salience of consonants versus vowels. While there was clearly a trend in favor of vowels, the results were somewhat ambiguous. Part of the difficulty may be that there are considerable differences among consonants in the degree to which crucial information is provided by consonantal or vocalic parts of the signal. In fact, examination of the error patterns in different consonant classes suggests that this might be the case. Error rates for the Miller-Nicely condition shown in Fig. 1 were re-calculated to include only within-class errors. Thus, for example, the error rate for /p/ was calculated as the percent of /t/ and /k/ responses, divided by the number of /p/, /t/ and /k/ responses. These within class error rates are shown in Fig. 2. As can be seen in Fig. 2, nasals have the lowest error rate. /m/ and /n/ are primarily distinguished from one another by the formant transitions from the nasal murmur into the vowel.

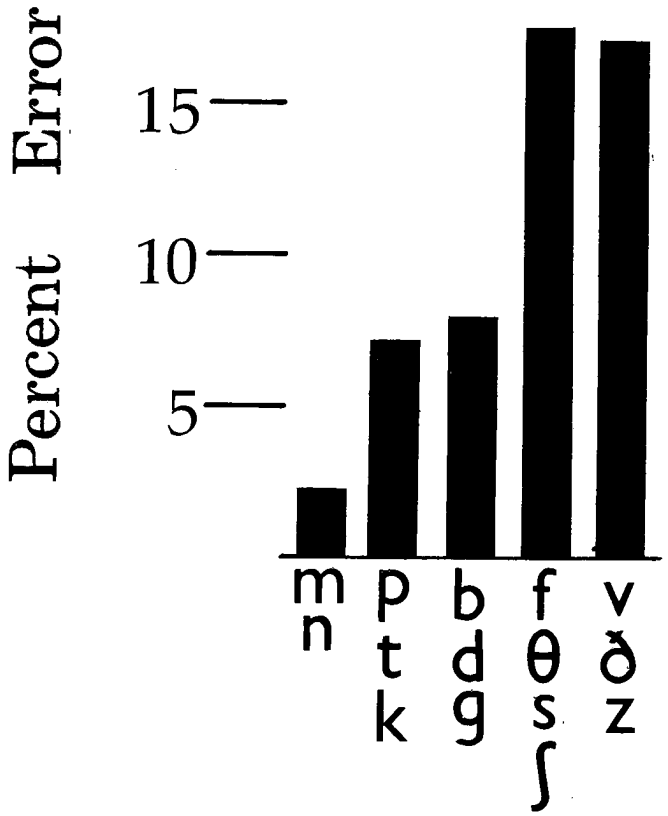

Fig. 2. For each class of consonants, percentage of within-class responses that are errors (see text). Data are from Miller and Nicely (1955).

Similarly, for voiced stops (and to a lesser extent, voiceless stops), sufficient perceptual information is probably present in the vowel for the discrimination of these stimuli. However, for fricatives, which show the highest error rates, it seems that the high-frequency consonant information present in the stimuli is necessary to the discrimination of different fricatives. Thus, it is at least possible that differences in error rate for consonants could be explained on the basis of how much perceptual information is carried in the following vowel that is relevant to the discrimination of a class of consonants. Resolution of this point will have to await further studies that attempt to precisely describe how different kinds of acoustic information interact in recognition.

Finally, it would be important to discover whether the results of the current experiment, using words and phrases in isolation, generalize to natural perceptual processing. Browman (1977a) has collected naturally-occurring misperceptions and has analyzed differences in segmental error rates, for segments in different word and syllable positions. She separates a component of error rate that is due to acoustic differences among the various positions from one that is due to differences in attention of the lexical decision mechanism. This is done by dividing the error rate in each position into the sum of two theoretical probabilities: the probability of making an *acoustic* error (ie., an error caused by the phonetic-analyzing and acoustic-sampling mechanism) and the probability of making a *lexical* error (ie., an error caused by the choice of the wrong lexical item, possibly consistent with the phonetic analysis of some other parts of the word). From the viewpoint of the present paper, a high *acoustic* error rate for some syllable corresponds to a high degree of phonetic ambiguity, whereas, a high *lexical* error rate corresponds to the tendency of the higher-level decision mechanism to ignore some syllable. Thus, the claim stressed syllables are *both* less phonetically ambiguous and are more highly weighted by higher-level mechanisms would predict that, in Browman's analysis, stressed syllables should show *both* lower acoustic and lower lexical error rates than unstressed syllables. Browman's naturally-occurring misperceptions were, therefore, analyzed to test this hypothesis and this prediction was confirmed (see Browman, 1977b) for details of this analysis . Thus, the perceptual salience of stressed syllables seems to hold in a very similar way in natural perceptual situations as well.

NOTES

[1] After preparation of this manuscript, the work of Cole and Jakimik (1977) became available to the author. They present a theoretical framework very similar to the one presented here, but more explicit with respect to modelling the word recognition process. In addition, they present data in support of the perceptual salience of stressed syllables using the listening for mispronunciations paradigm. Errors in stressed syllables were detected more rapidly than errors in unstressed syllables. In addition, evidence is presented for the perceptual salience of stops and of word-initial consonants.

ACKNOWLEDGMENTS

C.P. Browman, Eric Holman, and Peter Ladefoged all made substantial contributions to the content of this paper, and to its final form as well. Thanks to Vicki Fromkin and Ian Maddieson for valuable discussion. This research was supported by NIH.

REFERENCES

Blesser, B. (1969). Perception of Spectrally rotated speech. PhD dissertation, Massacusetts Institute of Technology, Cambridge, Masschusetts.

Browman, C.P. and Goldstein, L.M. (forthcoming). "A set of algorithms for matching strings of phonetic segments." To appear in UCLA Working Papers in Phonetics.

Browman, C.P. (1977a). "Perceptual processing: evidence from slips of the ear." Paper presented at 12th International Congress of Linguists, workshop on errors of speech and perception, Vienna.

Browman, C.P. (1977b). The word as a unit--evidence from slips of the ear and tips of the tongue. Phd dissertation, UCLA, Los Angeles, California.

Cole, R.A. and Jakimik, J. (1977). "Understanding speech: How words are heard," to appear in G. Underwood (ed). Strategies of Information Processing (Academic Press, New York).

Collier, R. and 't Hart, J. (1975). "The role of intonation in speech perception," in A. Cohen and S.G. Nooteboom (eds). Structure and Process in Speech Perception (Springer-Verlag, New York), 107-121.

Cooper, W.E. (1975). "Selective adaptation to speech," in F. Restle, R.M. Schiffrin, N.J. Castellan, H. Lindman and D.B. Pisoni (eds). Cognitive Theory: Volume I (Erlbaum Associates, Potomac, Maryland), 23-54.

Cutler, A. (1976). "Phoneme-monitoring reaction time as a function of preceding intonation contour," Percep. and Psychophys. 20, 55-60.

Fodor, J.A. and Bever, T.G. (1965). "The psychological reality of linguistics segments," J. Verbal Learning and Verbal Behavior 4, 414-420.

Foss, D.J. (1969). "Decision processes during sentence comprehension: effects of lexical item difficulty and position upon decision times," J. Verbal Learning and Verbal Behavior 8, 457-462.

Fredriksen, J.R. (1971). "Statistical decision model for auditory word recognition," Psychol. Rev. 78, 409-419.

Goldstein, L.M. (1977). Bias and asymmetry in speech perception." To appear in UCLA Working Papers in Phonetics.

Klatt, D.H. (1975). "Vowel lenthening is syntactically determined in a connected discourse," J. Phonetics 3, 129-140.

Kucera, H. and Francis, W.N. (1967). Computational Analysis of Present-day American English (Brown University Press, Provience, Rhode Island).

Ladefoged, P.N. and Broadbent, D.E. (1960). "Perception of sequence in auditory events," Quart. J. Exp. Psychol. 12, 162-170.

Lesser, V.R., Fennell, R.D., Erman, L.D. and Reddy, D.R. (1974). "Organization of the HEARSAY II speech understanding system," Carnagie Mellon University Working Papers in Speech Recognition 3, 11-21.

LaRivere, C. and Winitz, H. (1977). "Factors contributing to the recovery of monosyllabic words excerpted from natural speech," Kansas City Working Papers in Speech Science and Linguistics 6, 23-43.

Lieberman, P. (1963). "Some effects of semantic and grammatical context on the production and perception of speech," Language and Speech 6, 172-187.

Miller, G.A. and Nicely, P. (1955). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. 27, 338-352.

Oller, D.K. (1973). "The effect of position in utterance on speech segment duration in English," J. Acoust. Soc. Am. 54, 1235-1247.

Pickett, J.M. and Pollack, I. (1963). "Intelligibility of excerpts from fluent speech: effects of rate of utterance and duration of excerpt," Language and Speech 6, 151-164.

Pisoni, D.B. and Sawusch, J.R. (1975). "Some stages of processing in speech perception," in A. Cohen and S.G. Nooteboom (eds). Structure and Process in Speech Perception (Springer-Verlag, New York), 16-34.

deRooij, J.J. (1976). "Perception of prosodic boundaries," IPO Progress Report 11, 20-24.

Rubin, P., Turvey, M.T. and van Gelder, P. (1976). "Initial phonemes are detected faster in spoken words than in spoken nonwords," Percep. and Psychophys. 19, 394-398.

Savin, H. (1963). "Word-frequency effect and errors in the perception of speech," J. Acoust. Soc. Am. 35, 200-206.

Shayne, J. and Gass, S. (1976). "An investigation of the role of stress as a factor in speech perception," UCLA Working Papers in Phonetics 31, 78-85.

Stevens, K.N. and Blumstein, S.E. (1975). "Quantal aspects of consonant production and perception: a study of retroflex stop consonants," J. Phonetics 3, 215-233.

Strange, W., Verbrugge, R.R., Shankweiler, D.P. and Edman, T.R. (1976). "Consonant environment specifies vowel identity," J. Acoust. Soc. Am. 60, 213-224.

Svensson, S.G. (1974). "Prosody and grammar in speech perception," MILUS 2, Institute of Linguistics, University of Stockholm.

APPENDIX

Words and phrases for recognition experiment

I. Monosyllabic stimuli (N=26)

deem	fain	yawl	eel
rogue	zooms	tithes	ale
vow	bib	wed	oafs
cam	gull	nooks	owl
moth	thong	lute	elks
chive	hick	pap	
shun	soot	jot	

II. Bisyllabic stimuli (N=24) [S-U, U-S are stress patterns]

Single Words (N=12)		Two-word Phrases (N=12)	
S-U (N=7)	U-S (N=5)	S-U (N=6)	S-U (n=6)
yokel	gazelle	loused you	will zip
sulfide	rotund	footfall	it jibes
dilate	typhoon	bait thief	caused soot
shoddy	motet	knee-type	call thugs
catchy	beguile	this chow	build mounds
vacuum		home-bound	put shims
supine			

III. Trisyllabic stimuli (N=20)

Single Word (N=6)	Two word (N=8)	Three word (N=6)
U-S-U (N=2)	U-S-U (N=4)	U-S-U (N=2)
pulsation	his visage	could log-jam
demented	those bovines	that pay-off
	we chatted	
	looked jazzy	
S-U-U (N=2)	S-U-U (N=2)	S-U-U (N=2)
numinous	faucet nut	walk-out day
ravages	ketchup lump	woodwind show
U-U-S (N=2)	U-U-S (N=2)	U-U-S (N=2)
referee	gets shampoo	ate the gauze
guarantee	you adduce	not his chump

Chapter 3:

Bias and asymmetry in speech perception

Bias and Asymmetry in Speech Perception

Louis Goldstein

INTRODUCTION

Asymmetries in speech perception

Patterns of errors made by listeners when identifying auditorily presented speech sounds have frequently been studied to help in understanding the speech perception process. Such research has ranged from the work of Miller and Nicely (1955) who examined errors made in the perception of nonsense CV syllables under various conditions of noise and filtering, to recent investigations of perceptual errors that occur in normal conversation (Garnes and Bond, 1977; Browman, 1977). A common thread running through such research is the assumption that the more confusable a pair of speech sounds is, the greater their similarity with respect to the perceptual system. These similarities have been analyzed, by various techniques, into dimensions or features of perceptual similarity (e.g. for consonants, Miller and Nicely, 1955; Singh and Black, 1966, Shepard, 1972; Wang and Bilger, 1973; Wish and Carroll, 1974; Goldstein 1977a).

The concept of similarity employed by the above studies is a symmetric one -- the similarity between segment A and segment B is the same regardless of which is the stimulus, and which is the response. However, a casual perusal of the confusion matrices in published experiments reveals that this assumption is not always supported by the number of confusions in the data. For example, in the Miller-Nicely confusion matrices, /θ/ is reported as /f/ more frequently that /f/ is reported as /θ/. Such asymmetries have generally been attributed to a bias in favor of reporting some segments more than others. As such, this response bias has been considered irrelevant to the underlying similarity between the segments in question, and various techniques have been employed to remove such effects from confusion matrices (as will be discussed below) in order to analyze the pattern of symmetric similarities.

Other kinds of investigations have required the removal of a response bias component from a confusion matrix before conducting some major analysis of the pattern of confusions. Verbrugge et. al. (1976) wanted to examine various hypotheses about the effect different presentation conditions on error rate in vowel recognition. Looking only at the change in error rate from condition to condition, it is impossible to know whether the change in error rate is due to a change in the inherent distinctiveness of a vowel, or a change in the tendency to give certain vowels as a response. Verbrugge et. al. employed one of the models to be discussed below (that of Luce, 1959) to separate a change in response bias from a change in degree of ambiguity. Another example of this problem can be seen in Goldstein

(1977b), who has shown that the relative error rates for a set of consonants in the Miller-Nicely data is similar to the relative error rates for these consonants in a word and phrase recognition task. It could not be concluded, however, whether this similarity was to be explained in terms of the relative ambiguity of the consonants, or of their relative response biases (or both). Finally, Janson (1977) has shown that asymmetries in the confusion of Dutch vowels (in the data of Klein et. al., 1970) can be explained by a bias that is plausibly related to the fact that the range of F2 covered by the vowels in the experiment may differ substantially from the range normally encountered in Dutch.

Clearly, there is a need for an appropriate model to separate out bias and symmetric components in a confusion matrix. Part of the purpose of this paper is to compare the results of two different models for finding the bias in a confusion matrix -- the linear model proposed by Luce (1959) and the nonmetric model proposed by Holman (personal communication). These models will be discussed in detail below.

While the experiments noted above (and others) have attempted to remove bias from confusion matrices without interpreting it, it is possible that this bias is itself of some interest. Let us assume that bias can be shown to be reliable across different experiments, using the same stimuli, under different listening conditions. If this were the case, then this bias ought not to be discarded in an investigation of the perceptual system, as it would be potentially valuable information as to the working of the system itself. For example, reliable bias in perceptual confusions would differentiate speech perception and production systems with respect to errors. ShattuckHufnagel and Klatt (1977) have shown that confusions among segments in speech production (i.e., speech errors involving single segments), tend to be extremely symmetric (with only one or two isolated asymmetries). They show that there is essentially no bias in their (speech error) data.

Asymmetries in perception are also important to examine from the point of view of helping to explain some of the asymmetries that can be observed in phonological processes (both historical and synchronic) found commonly in languages. For example, many languages have a rule (or underwent a historical process) whereby a /k/ is palatalized to /tʃ/ before an /i/. However, examples of the converse (/tʃ/ becoming /k/ before /i/), are quite rare. Similarly, many languages undergo a process whereby syllable-final (or word-final) obstruents are devoiced, but a rule voicing syllable-final obstruents is, again, quite rare. It is possible that reliable asymmetries in speech perception can be shown to be related to such asymmetric processes.

Finally, on analogy to word recognition experiments, we might expect that frequency of occurrence of a segment in actual use would be responsible for a bias in perception. The recognition threshhold for a given word can be shown to be a function of its frequency of occurrence (see Howes, 1957). A common explanation for this lowered threshhold is that there is a response bias to emit common words as

responses, regardless of the stimulus. While the details of the explanation are quite varied (see e.g. Goldiamond and Hawkins, 1958; Savin, 1963) some notion of response bias is implicated in the explanation. We might, then, expect that subjects in a phoneme recognition task would also show a tendency to choose as responses those phonemes that occur most frequently in the language. Any reliable bias in perception should be compared, therefore, with data on frequency of occurrence.

Bias models of asymmetry

We will distinguish in the following discussion between two types of bias models -- metric and nonmetric. Metric bias models attempt to relate the observed data in a confusion matrix to a set of underlying symmetric and bias parameters by means of a linear equation. The nonmetric model assumes only a monotonic relationship between the underlying parameters and the observed data. The best-known metric model is that proposed by Luce (1959). To explicate this model, let us consider a confusion matrix of n objects, in which p(xy) represents the number of times the stimulus x is reported as response y. Let each row of the matrix represent a particular stimulus and each column represent a particular response. Thus, confusions of x with y are represented by p(xy) and confusions of y with x by p(yx). Correct responses for the set of stimuli are represented by p(1,1)... p(x,x)... p(n,n). If we divide the entries in each row of such a matrix by the total number of times each stimulus was presented, then for each entry in a given row, p(xy) represents the *proportion* of times stimulus x is reported as response y. The essence of the Luce model, is that the probability p(xy) can be represented as the product of a symmetric function of x and y and a bias function on y. This is shown in (1) below, (a scale factor in the denominator of (1) has been left out of the equation for the sake of expository simplicity):

(]) $p(xy) = b(y) \cdot s(xy)$
where $s(xy) = (yx)$

Thus, this model assumes that any observed asymmetry between p(xy) and p(yx) can be accounted for by a difference in the relative response bias of x and y, b(x) and b(y), respectively. We refer to this as the metric response bias model.

It would seem plausible that asymmetric properties in a confusion matrix could be due to differences among stimuli in terms of their tendency to be confused, as well as differences among responses in terms of their tendencies to be produced. Thus we might consider modelling this situation by adding another bias function (b') to (1), representing the confusability of each stimulus, as in (2):

(2) $p(xy) = b'(x) \cdot b(y) \cdot s(xy)$

However, Holman (personal communication) has shown that any set of

data that can be modeled by (2), with both response and stimulus confusability biases, can be modelled equally well by (1), with only a response bias. This can be shown as follows:

$$(2) \quad p(xy) = b'(x) \cdot b(y) \cdot s(xy)$$

$$(2a) \quad = \frac{b'(y) \cdot b'(x) \cdot b(y) \cdot s(xy)}{b'(y)}$$

by multiplying by $\frac{b'(y)}{b'(y)}$

$$(2b) \quad = \frac{b(y) \cdot [b'(x) \cdot b'(y) \cdot s(xy)]}{b'(y)}$$

by rearranging factors

$\frac{b(y)}{b'(y)}$ in (2b) defines a new function that depends on the response only. It differs from the original functions b and b', but it represents all the asymmetrical information in a single function on y. This is true because a $b'(x) \cdot b'(y)$ in the equation (2b) is a symmetric quantity, that is, it has the same value for p(xy) as for p(yx). In a parallel way, it can be shown that all the asymmetric information in b' and b can be represented in a new bias function defined only on stimuli. This function will be $\frac{b'(x)}{b(x)}$, or the reciprocal of the response function. Thus, mathematically there is no unique solution, in the metric model, for representing the asymmetry of the matrix in terms of bias functions. A response bias can be modelled as a stimulus bias with the appropriate change in the symmetric component, and *vice versa*. The bias functions will simply be reciprocals of one another.

Although stimulus and response bias models are mathematically equivalent, in terms of fit to the data, it can be demonstrated that the response bias model is more useful than stimulus bias model for confusion matrices in which the rows all sum to 1. As noted above, stimulus and response bias models of the same data will differ in terms of their symmetric component. Crucial to this argument is the value of the model for the similarity component of the diagonal elements (the so-called self-similarities) s(xx)...s(yy)...s(nn). Either the stimulus or response model will predict that the value of a diagonal entry p(xx) will be equal to b(x).s(xx). We can interpret b(x), as before, as a response bias, a tendency for x to occur as a response regardless of the stimulus. The self-similarity term can be thought of as representing the relative distinctiveness of a given

item. An item with a large s(xx) can be considered to be very distinct; it tends not to be involved with other items in confusions, either as stimulus or response. A low value of s(xx) can be interpreted as a relatively ambiguous item -- it enters into confusions readily - either as stimulus or response.

Let us consider two diagonal entries in a confusion matrix $p(x_1x_1)$ and $p(x_2x_2)$, such that $p(x_1x_1) > p(x_2x_2)$, and see how this situation could be represented in stimulus and response bias models. $p(x_1x_1) > p(x_2x_2)$ means, of course, that there are more correct responses for stimulus x_1 than for stimulus x_2. In the response bias model, this inequality could be predicted in one of two ways. If, in general, there are more x_1 responses (not including the diagonal) than x_2 responses, then $b(x_1)$ will be greater than $b(x_2)$. Thus, the differences in the bias components are in the same direction as differences in the diagonal values, and the self-similarity parameters will vary so as to predict just the right magnitude of difference between $p(x_1x_1)$ and $p(x_2x_2)$. If, however, there are generally more x_2 responses than x_1, $b(x_2)$ will be greater than $b(x_1)$ and $s(x_1x_1)$ will have to be greater than $s(x_2x_2)$ in order to account for the diagonal entries. Thus, in this model, the relationship between the diagonal entries and the response totals in the off-diagonals conspire to determine the best s(xx) values in an intuitively plausible way -- a stimulus which has a lot of correct responses, but is not reported frequently as a response to other stimuli will be considered to be a distinctive, non-ambiguous stimulus, with a low response bias.

The situation is rather different for the stimulus bias model, however. For this case, again, let us examine two diagonals $p(x_1x_1) > p(x_2x_2)$. If x_1 has more correct responses than x_2, than x_1 also has fewer confusions than x_2 since row sums must add to 1. Thus, assigning a low stimulus bias $b(x_1)$ to x_1 on the basis of the fewer confusions would make the *wrong* prediction about the number of correct responses, since the same $b(x_1)$ will appear both in the diagonals and the off-diagonals of a given row. $s(x_1x_1)$ will thus have to be inversely correlated with $b(x_1)$ to produce the right number of correct responses of $p(x_1x_1)$. For data with equal row sums, s(xx) will always turn out to be negatively correlated with b(x), since b(x) cannot simultaneously describe the tendency of a stimulus to be both correctly perceived and confused. For this reason, a response bias model is more convenient, because it allows us to calculate, for a set of data, a bias function and a set of self-similarities that are, at least in principle, independent of one another.

The nonmetric bias model also represents a confusion matrix in terms of two sets of underlying parameters -- a bias function, b, symmetric function s(xy). Unlike the metric model, however, the nonmetric model does not make an assumption about the particular form of the function that relates the parameters b and s to the data (p(xy)), it only assumes that p(xy) is monotonic on b and s, in the sense that (3) is assumed to be true:

(3) If $s(xy) \geq s(wz)$ *and* $b(x) \leq b(w)$ *and* $b(y) \geq b(z)$
this implies:
$p(xy) \geq p(wz)$

Thus, this model makes the following prediction about the relationship between the cells xy and wz of the confusion matrix. If the symmetric component of xy is greater than that for wz and the bias for y is greater than the bias for z, and the bias for x is smaller than the bias for w, then there should be more confusions p(xy) that p(wz).

The model attempts to order objects on b, and pairs of objects on s(xy) such that the number violations of (3) is minimized. It should be noted that this model only makes a prediction about the inequality between two of the cells in the matrix, just in case the conditions in (3) are met; if these are not met, *no* prediction is made. The bias function in this model is both a stimulus and response bias -- one can think of the objects (phonemes in this case) as being ordered on this function in such a way that those objects with low values on the bias function tend to be responded to as objects with high values on the bias function more often than vice versa. Put another way, objects with high bias values tend to intrude on objects with low bias values. The model uses only the confusions in the original matrix, it does not use the data on the diagonal at all. Thus, no estimates of self-similarities are produced by this model, and there is no problem in the relationship between the bias function and the self-similarities, as was encountered in the metric stimulus bias model.

The metric bias model is stronger than the nonmetric model, in that it assumes a particular, linear relationship between the underlying parameters and the observed data. This stronger model implies the weaker nonmetric model, in the sense that the b and s parameters derived in fitting the metric model should still fulfill (3), assuming that the model is appropriate to the data. This follows because the linear function assumed by the metric model is a monotonic function on b and s. Thus, it is possible to show the inappropriateness of the metric bias model for a particular set of data by showing that the b and s parameters from the metric model of a given data set lead to a substantially greater number of violations of (3) than does the nonmetric model that reduces such violations to a minimum.

In the analysis to be described below, metric and nonmetric bias models will be fitted to the same perceptual confusion data. The

models will be compared. At the same time, the confusion data are chosen so as to allow assessment of the reliability of the obtained bias functions. Thus, we will be able to compare metric and nonmetric models both with respect to violations of (3) and with respect to the reliability of the obtained bias functions. Moreover any reliable bias obtained will be interpreted with respect to general asymmetric linguistic processes, and will be compared with frequency of occurrence.

METHOD

Data: confusions

The perceptual confusion data published by Wang and Bilger (1973) was chosen for analysis. This data includes confusion matrices for four different sets of syllables referred to as CV1, VC1, CV2 and VC2. Each set of 16 syllables was presented under two different listening conditions. In one of the listening conditions, syllables were presented with background white noise of various S/N levels (this is referred to below as the N condition). In the other conditions, (referred to as Q below), syllables are presented without background noise at a variety of low signal levels. The particular consonants involved in these syllables are shown in Table 1. Note that the CV1 and VC1 sets include the identical set of consonants -- the English non-nasal stops, fricatives, and affricates. CV2 and VC2 sets include the English syllable-initial and syllable-final consonants that were not included in the CV1 and VC1 sets, and thus are not identical to one another. The vowels in all sets were /i/, /a/, and /u/. Each of the eight confusion matrices represents data summed over subjects, vowels and S/N level.

For each syllable type, the data in the noise and quiet conditions constitute potential tests for the reliability of extracted bias functions. It is easy enough to imagine that differences between these two presentation conditions are sufficient to introduce some differences in bias. However, there would not be much theoretical interest in a bias function that was not even reliable across conditions as similar as these.

Data: frequency

In order to test the hypothesis that the biases obtained in perceptual confusions would correlate with the frequency of occurrence of the consonants, data on consonant frequency was tabulated. Four different sets of frequency of occurrence data were obtained, one set from Carterette and Jones (1974), and three sets from Roberts (1965). Carterette and Jones recorded spontaneous, informal speech of children and adults, transcribed this speech and established phoneme frequency counts based on the transcriptions. For adults, the sample included 15,964 words of spoken speech, based on 24 speakers. From their totals for adults, the overall frequency of occurrence for consonants was obtained.

Table 1. Consonants used in the Wang and Bilger (1973) experiment, for each of the four conditions.

CV1 VC1	p	t	k	b	d	g	f	θ	s	ʃ	v	ð	z	ʒ	tʃ	dʒ
CV2	p	b	tʃ	dʒ	l	r	f	s	v	z	h	hʷ	w	j	m	n
VC2	p	b	g	m	n	ŋ	f	θ	s	ʃ	v	ð	z	ʒ	tʃ	dʒ

Table 2. Kendall's tau for correlation of metric and nonmetric bias functions. Associated significance levels in parentheses.

CV1N	.70	(.001)	CV1Q	.46	(.007)
VC1N	.70	(.001)	VC1Q	.80	(.001)
CV2N	.67	(.001)	CV2Q	.52	(.003)
VC2N	.72	(.001)	VC2Q	.70	(.001)

Roberts (1965) recorded a speaker reading sentences that included the 10,000 words in the Horn (1926) word count. These words were then transcribed and the frequency of occurrence of phonemes in the word list was determined. The frequencies of occurrence of the phonemes in the language were then computed on the basis of their frequencies in the words of this sample, and the frequency of occurrence of these words in the language, as reported by Horn (1926). From this phoneme count, the overall frequency of occurrence of consonants, and the frequency of occurrenence in word-initial position was obtained. In addition to the frequency of occurrence in the language of these consonants, the frequency in the Horn corpus was also noted. This can be considered a measure of *lexical* frequency, rather than frequency of occurrence, i.e., it is an estimate of the frequency of occurrence of the phonemes in a hypothetical dictionary of English in which each word is represented once.

Analysis

For each of the Wang and Bilger confusion matrices, the entries in each row were divided by the row totals, yielding estimates of row conditional probabilities. The resulting matrices were submitted to two computer programs. A program written by E. Holman found the best solution for the nonmetric model. The program converged within 15 iterations for all data sets analyzed. The second program, written by T. Wickens, used an iterative procedure to find the maximum likli-hood metric response bias model for a given set of data that included a diagonal. The program generally converged in less than 30 iterations, although for one of the data sets (CV2Q), the self-similarities were still changing slightly after 100 iterations.

RESULTS

Metric vs. nonmetric bias models

The rank order of the consonants in the bias functions was quite similar for metric and nonmetric models. Rank order correlations (Kendall's tau) between metric and nonmetric biases are shown in Table 2 for each of the eight confusion matrices. Each correlation is significant at better than the .01 level.

The following procedure was used to test the appropriateness of the metric bias model for the data. For each data set, we calculated the percent error for predictions made by the metric bias function under the assumption of monotonicity (made in equation 3). The nonmetric bias program was used to make this calculation. The percent error was compared to the comparable value for the nonmetric bias function. The results are shown in the left-hand columns of Table 3. It is clear that there are more violations of monotonicity for the metric than for the nonmetric bias, for all matrices. However, the differences are rather small for all data sets except CV1Q.

In some sense, the nonmetric biases should have an advantage in the above comparison, since they are actually calculated so as to

Table 3. Percent of predictions of inequalities in data that are incorrect. Results are shown for bias functions derived for metric and nonmetric models. Bias functions are used to predict data from which they have been derived (within-data) and data for contrasting noise condition (cross-data).

	within-data		cross-data	
	metric bias	nonmetric bias	metric bias	nonmetric bias
CV1N	.150	.110	.225	.214
CV1Q	.286	.173	.287	.271
VC1N	.079	.060	.129	.126
VC1Q	.108	.087	.143	.145
CV2N	.186	.150	.466	.508
CV2Q	.240	.197	.460	.467
VC2N	.111	.088	.184	.253
VC2Q	.124	.099	.231	.294

Table 4. Kendall's tau for reliability of metric-derived bias across conditions of noise and quiet. Associated significance levels in parentheses.

CV1N-CV1Q	.63	(.001)
VC1N-VC1Q	.64	(.001)
CV2N-CV2Q	.17	(.184)
VC2N-VC2Q	.53	(.002)

minimize the particular quantity being compared. A better analysis, therefore, would involve comparing violations of monotonicity for a bias function when used to predict, not the original data it was based upon, but the data of the other confusion set, having the same syllables in a different noise condition. Thus, bias functions generated for the quiet condition data sets were used to predict the noisy condition data and vice versa. The percentage violations of monotonicity for these cross data set comparisons are shown in the right-hand columns of Table 3. These results show a rather different pattern from the within data set comparisons. For the CV1 data sets, there are slightly more violations of monotonicty for the metric bias. For VC1 sets, differences are exceedingly small, one favoring the metric, the other the nonmetric. For CV2 sets, the cross data violations are so numerous (about half) that this data is largely irrelevant. (We will return below to this case). Finally, for VC2 sets there were substantially fewer violations for the metric model. Thus, the metric solution for the VC2 data seems to be producting bias functions that are more reliable across data sets than the bias functions produced by the nonmetric prodedure. Moreover, since the within data set violations are not much larger for the metric model than for the nonmetric, the advantage of the metric solution does not seem to come at the expense of making many more errors in predicting the original data. Thus, the stronger assumptions of the metric model seem appropriate to the confusion data at hand, in that its monotonic fit to the data cannot be considered worse than that for the nonmetric bias (if anything, it might be considered better for the metric bias). Since the metric model also has the advantage that the analysis provides an estimate of the self-similarity, or distinctiveness, of all the consonants, in addition to the bias function, it is the metric solutions that we will analyze in detail below.

Reliability of the metric bias

To test the hypothesis that the bias for a particular type of stimulus material would be reliable across presentation conditions, the metric bias functions from noise and quiet conditions were rank correlated (using Kendall's tau). Correlations for the four syllable types are shown in Table 4, along with associated significance levels. The sets clearly differ from one another in terms of reliability of bias. CV1, VC1, and VC2 sets all show highly significant correlations, while the CV2 set shows almost no correlation at all. It is odd that this set, with more nasals and approximants and fewer stops and fricatives than CV1, should have no reliable bias across noise and quiet conditions. However, as we shall see below, the CV2N and CV2Q sets behave differently from the other matrices in a number of ways.

Correlation of bias with frequency

The rank correlations of metric biases and consonant frequencies are shown in Table 5, for all eight sets of consonants. Correlations are given for each of the four frequency measures discussed above. Rank correlations are used, because the relationship between frequen-

Table 5. Kendall's tau for correlation between frequency and bias. Associated significance levels in parentheses.

	Carterette & Jones	Roberts general	Roberts word-initial	Roberts lexical
CV1N	.20 (.140)	.22 (.121)	.13 (.236)	.47 (.006)
CV1Q	.21 (.130)	.23 (.112)	.04 (.411)	.41 (.014)
VC1N	.45 (.008)	.43 (.01)	.45 (.008)	.58 (.001)
VC1Q	.24 (.096)	.29 (.057)	.34 (.032)	.51 (.003)
CV2N	.12 (.260)	.22 (.128)	.17 (.184)	.22 (.128)
CV2Q	.028 (.441)	-.10 (.293)	.03 (.429)	.16 (.200)
VC2N	.33 (.036)	.37 (.075)		.45 (.008)
VC2Q	.27 (.075)	.27 (.075)		.45 (.008)

Table 6. Percent of predictions of inequalities in data that are incorrect. Results are shown for frequency ranks used as bias functions.

	Carterette & Jones	Roberts general	Roberts word-initial	Roberts lexical
CV1N	.299	.345	.321	.244
CV1Q	.376	.421	.462	.312
VC1N	.179	.200	.129	.169
VC1Q	.256	.243	.231	.168
VC2N	.284	.320	.290	.252
VC2Q	.343	.322	.280	.238

cy and bias did not seem to be linear, either with the raw frequencies, or with log transforms of the frequencies. The hypothesis that reliable response bias would be a function of consonant frequency is partially supported by these results. For all eight sets of data, the highest correlations with bias are found for the lexical frequencies. Correlations of bias with lexical frequencies are significant for all four VC data sets and for CV1N. Correlations are marginally significant for CV1Q and, once again, virtually nonexistant for CV2N and CV2Q. VC syllables, in general, seem to show better correlations with frequency than CV syllables. VC1N has the best correlations whith frequency of all, showing significant correlations with frequency of occurrence measures, as well as the single largest correlation with lexical frequencies. VC1Q, VC2N and VC2Q show some marginal correlations with frequency of occurrence, and the CV sets show no correlation at all with frequency of occurrence, as opposed to lexical frequency.

The hypothesis that response bias is due primarily to frequency is strengthened by comparing the pattern of correlations to the patterns of goodness of fit of the model to the data, as shown by the percentage of monotonicity errors in Table 3. The percentage of violations of monotonicity is smaller for those data sets that show good correlations with frequency -- the VCs. The three data sets that have marginal or no correlation with frequency also have the greatest percent errors in Table 3 -- CV1Q, CV2N, and CV2Q. Thus, it seems that the greater the degree to which a bias funciton fits a set of data, the greater the correlation of the bias with frequency. This certainly supports the notion that phoneme frequency is an important determinant of reliable response bias.

Two analyses were undertaken to determine whether the reliable bias for a given syllable set could be considered to be exhausted by frequency, or whether there were, in addition, some other components of this bias. The first analysis was to use the phoneme frequencies themselves as the bias function, and to calculate, using the nonmetric bias program, the percent violations of monotonicity for this frequency bias function. These percent errors can then be compared to the values for the cross data set preditions of the metric bias functions, previously discussed, in Table 3. If frequencies are as good at predicting a given set of data as the biases from the paired data set, this would imply that there is nothing but frequency that is reliable in the bias. The percent errors of frequencies are shown in Table 6. Comparing these values to those in Table 3 for cross set predictions, it is clear that, in general, frequencies show more errors. The percent errors in predicting VC1N for the word-initial Roberts frequencies is the same as that for the VC1Q bias, but in every other case, frequencies are worse than metric biases in predicting the data. Thus, this seems to indicate that there is more than frequency that is reliable in the data.

The other procedure for deciding whether frequency exhausts the reliable bias is to partial frequency out of the bias functions of the noise and quiet sets of a given syllable type, and to see if they

still correlate significantly. Since we wanted to use frequency ranks rather than real values (given the linearity problem), we require some procedure for doing partial rank correlations. To approximate this, a Pearson correlation was performed, but using the ranks of frequency and bias as the values of the variables being correlated. This procedure essentially corresponds to a Spearman rank-order correlation. The correlation of these ranks, before and after partialling out frequency ranks, is shown in Table 7. The significance levels noted there should be regarded with some caution, given this partialling procedure. It is clear that the biases of noisy and quiet conditions for a given syllable type are still highly correlated, even after partialling out lexical frequency, the one that correlates best with the biases.

Other interpretations of bias

In order to determine what else is reliable in response bias, besides frequency, let us examine the actual bias functions themselves. In Table 8, the consonants are listed from left to right in order of decreasing bias, for each set of data. Let us first examine the bias for VC1N. The three voiceless stops /t,p,k/ show the highest bias and are followed closely by the three voiced stops and /s/: /g,s,b,d/. Thus, except for /s/, the high end of the bias function includes all and only the stop consonants, with the voiceless showing higher bias than the voiced ones. This reflects the rank of these consonants in terms of their likelihood to occur in the world's languages. Phonologies described by Hockett (1955) and the data summarizing the phonologies of 700 languages of the world in Ruhlen (1975) support the following claims: If a language has either stops or fricatives, but not both, it is much more likely to have stops; if it has either voiced or voiceless stops, but not both, it is much more likely to have voiceless stops. Moreover, if a language has only one fricative, it is more than likely to be /s/. Let us, oversimplifying for the present discussion, assume that the phonological naturalness of a given segment is directly related to its frequency among the languages in the world. Greenberg (1966) has shown that such distributional facts correlate with other criteria for "markedness" or naturalness, such as those discussed by Jakobson (1942), or Trubetzkoy (1958).
The bias for VC1N seems to coincide with this scale of phonological naturalness.

Unfortunately, what is phonologically natural also tends to occur frequency in English. Greenberg (1966) has shown that this is not true just for English -- in a variety of unrelated languages, markedness or naturalness of a segment was found to correlate with frequency of occurrence within the language. It is very difficult, therefore, to separate naturalness factors from frequency, in order to determine which, or both is responsible for bias in VC perception. For example, let us define a variable that can, in a limited sense, be considered to be an indication of phonological naturalness: [1] for all fricatives and [0] for all stops. (Affricates are considered fricatives in this analysis). This variable rank-correlates with lexical frequency

Table 7. Partial correlation (r) of bias ranks, after partialling out frequency ranks. Zero-order rank correlations (Spearman's) are also given. See text.

	Zero-order	Carter-ette & Jones	Roberts general	Roberts word-initial	Roberts lexical
CV1N-CV1Q	.82(.001)	.81(.001)	.81(.001)	.82(.001)	.75(.001)
VC1N-VC1Q	.82(.001)	.82(.001)	.80(.001)	.80(.001)	.65(.004)
CV2N-CV2Q	.31(.130)	.31(.142)	.41(.069)	.36(.102)	.27(.176)
VC2N-VC2Q	.73(.001)	.69(.002)	.69(.002)	.69(.002)	.59(.010)

Table 8. Bias from metric analysis. Consonants are ordered from left to right in order of decreasing response bias.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
CV1N	p	f	v	dʒ	t	g	s	z	d	b	k	ʃ	tʃ	θ	ð	ʒ
CV1Q	v	p	t	f	d	b	z	g	dʒ	tʃ	s	ʃ	ʒ	k	θ	ð
VC1N	t	p	k	g	s	b	d	v	ʃ	dʒ	f	ð	z	tʃ	θ	ʒ
VC1Q	p	k	t	s	g	d	f	ʃ	dʒ	v	tʃ	θ	b	ʒ	ð	z
CV2N	l	h	m	f	s	r	j	p	w	hʷ	dʒ	n	tʃ	z	v	b
CV2Q	z	s	m	f	p	l	b	r	v	dʒ	tʃ	hʷ	h	j	n	w
VC2N	s	p	n	ŋ	m	g	θ	b	dʒ	v	f	tʃ	ð	z	ʒ	ʃ
VC2Q	p	s	g	f	n	m	v	tʃ	dʒ	b	ŋ	θ	ʃ	ð	ʒ	z

(ρ = .53, $p < .02$). If one looks at the pattern of residuals after partialling frequency (lexical) out of the bias function for VClN, the consonants are no longer completely systematically ordered according to phonological naturalness. However, there is still a significant correlation of the feature stop with bias, even after partialling in this way ($r = 0.68$, $p < .002$, although, again the significance level should be regarded with some caution). Moreover, turning this around, there is still a substantial correlation of bias with lexical frequency, after partialling out the feature stop ($r = 0.63$, $p < .01$), the feature voice ($r = 0.74$, $p < .001$) or both features voice and stop ($r = 0.58$, $p < .02$). Thus, as far as can be determined at this point, lexical frequency and phonological naturalness are correlated but have separable effects on perceptual bias.

For the VClQ set, the bias is similar to that for the VClN set, except that /b/ has a very low bias, rather than being with the other stops at the top of the bias function. Moreover, the effects of frequency and phonological naturalness are more hopelessly intertwined in the VClQ case. Partialling out frequency makes the correlation of bias with the stop feature not significant, and partialling out the features stop and voice makes the correlation with frequency not significant.

More problematic results for the quiet conditions seem to be the rule in this data. For quiet conditions, in general, the bias functions fit the data worse than for the corresponding noise conditon (see Table 3). This should not be surprising, since there are many fewer errors in the quiet conditions, and therefore, the error distribution will be noisier, or less well determined. Thus, more problematic bias results in the quiet condition is somewhat less troubling than it would be in the noise condition.

There is a major objection to the phonological naturalness argument made above. As noted (in the Introduction), many languages devoice voiced stops in final position. Unfortunately for the present analysis, such devoicing is possible in English. The distinciton between voiced and voiceless final stops is often, in fact, cued by the length of the final vowel, rather than voicing in the consonant inself. (see Lisker, 1974, Javkin, 1976). Thus, it is possible that the stimuli in the Wang and Bilger experiment included devoiced final stops. They may have tended to be reported as voiceless stops more than vice versa (this is the implication of their relative positions on the bias hierarchy) because they actually were produced somewhat between fully voiceless and fully voiced final stops. Without the acoustic data from the experiment, it is not possible to fully resolve this point.

There is another point for which it would be useful to have the acoustic data for the syllables used in the Wang and Bilger experiment. It is possible that the final stops were released, and perhaps even followed by a very short vowel, if the reader was trying to articulate them very clearly. If this were the case, the somewhat unusual

acoustic marking might somehow be responsible for the bias in favor of the stops.

Returning to an examination of other interpretations of bias in the data, the bias in the VC2N condition also shows an effect of phonolgical naturalness. This stimulus set includes three stops -- /p,b,g/ -- and three nasals -- /m,n,ŋ/. These six consonants are included in the first eight positions of the bias function. Again, the voiceless stop has a higher bias than the voiced ones. Nasals are also common segments in the languages of the world, and in fact, there are languages (such as Peking Chinese) where the only syllable-final consonants are nasals. Thus, the fact that nasals have high bias once again supports the association of the bias function with phonological naturalness. The results for the corresponding quiet condition are, again less clear. Both /b/ and /ŋ/ have considerably lower values of bias in the VC2Q condition than in the VC2N condition.

Unlike the situation for VCs, it is very difficult to find a phonological naturalness interpretation of the bias for CV1N or CV1Q. In fact, it is very difficult to find any interpretation of the bias other than the somewhat weak correlation with frequency. The fricatives /θ,ð,ʃ,ʒ/ seem to be at the extreme weak end of the bias continuum for both CV1N and CV1Q. However, it is not clear how to interpret this. Similarly, there seems to be a preference for grave or non-coronal consonants at the high bias end of the continuum. Again, no explanation suggests itself. Similarly for CV2N and CV2Q, there is no obvious interpretation of the bias. The CV2 biases, as we see, are also not reliable, do not correlate with frequency, and have high proportions of errors in predicting the inequalities in the data. In addition, there is very little variability in the actual metric bias values assigned for each of these two conditions. Thus, it seems that there is only weak, uninterpretable, unreliable bias for these syllables. Possible reasons for this will be outlined in the discussion, below.

Self-similarities

One of the reasons for preferring the metric bias model, as outlined above, is that it is possible to obtain a measure of the relative distinctiveness (or ambiguity) of the consonants, in addition to a measure of bias. This measure of distinctiveness -- the self-similarities -- was also reliable in the data analyzed. The correlations between the noise and quiet conditions are shown in Table 9. Even the CV2 data shows a marginally significant reliability for self-similarities as opposed to bias. However, as it turns out, these self-similarities are highly correlated with the bias functions. This makes them very difficult to interpret, independently. The rank correlations of the biases and self-similarities for the eight data sets are shown in Table 10. It should be noted that these correlations are higher, in every case, for the quiet condition than for the corresponding noise condition. It is not clear why this is the case,

Table 9. Kendall's tau for correlation between self-similarities across noise and quiet conditions. Associated significance levels in parentheses.

CV1N-CV1Q	.85 (.001)
VC1N-VC1Q	.82 (.001)
CV2N-CV2Q	.45 (.04)
VC2N-VC2Q	.76 (.001)

Table 10. Kendall's tau for correlation of bias and self-similarities for each data set. Associated significance levels in parentheses.

CV1N	.53 (.002)	CV1Q	.59 (.001)
VC1N	.65 (.001)	VC1Q	.93 (.001)
CV2N	.58 (.001)	CV2N	.70 (.001)
VC2N	.70 (.001)	VC2Q	.85 (.001)

although it suggests the following: The computation of bias may be more heavily dependent on the relative sizes of the diagonals in conditions in which there are relatively few errors, than in conditions in which there are relatively more errors. Clearly, in conditions with few errors, differences among consonants, in terms of total number of responses, are going to depend very heavily on differences in the diagonal elements.

The consonants in each of the eight conditions are rank-ordered by self-similarity in Table 11. It is quite similar, of course to Table 8, since the biases and self-similarities are quite highly correlated. One way to interpret the self-similarities is to see how they differ from the bias functions. Comparing Tables 8 and 11, the most obvious difference is in the position of sibilants, particularly, /tʃ/ and /ʃ/. These consonants have very low biases, but in every condition but one, the rank of /tʃ/ and /ʃ/ is higher in self-similarities than in biases (in the one exception, VC1Q, /ʃ/ has the same rank in both). This suggests that while there is very little response bias in favor of /tʃ/ or /ʃ/, they are relatively unambiguous consonants. This is certainly a plausible result, and therefore suggests that the model may, in fact, be separating out response bias from relative ambiguity, as it should be. The high correlation between the two, therefore, may indicate that relatively distinct consonants, in fact, tend to attract responses.

DISCUSSION

We have seen that the metric response bias model is reasonably appropriate for consonant confusion data, that there are reliable biases in consonant perception, especially for VC syllables, and that such biases can be shown to correlate either with lexical frequency or phonological naturalness, or both. There are still a number of interesting issues raised by the biases found in CV and VC perception. Foremost among these is the question of why there is a difference between CVs and VCs in terms of bias. The bias function for VCs seems to fit the monotonic bias model better than that for CVs, the bias seems more generally reliable for VCs (CV2, it should be recalled, was very unreliable), and the bias seems more interpretable for VCs (both in terms of frequency and phonological naturalness). What can account for these differences?

There are two different explanations for the CV/VC distinction, depending on whether we think the bias is related to a perceptual preference for phonologically natural sequences or to a lexical frequency effect. On the theory that bias is related to phonological naturalness, the explanation for the CV/VC differences is quite straightforward. Languages in general seem to have a more restricted distribution of consonants syllable-finally than syllable-initially. There are many languages that have no syllable-final consonants at all, but a language with no syllable-initial consonants would be rare, indeed. Since languages are more restricted in their consonant

Table 11. Rank order of self-similarities. Consonants are arranged from left-to-right in order of decreasing self-similarity.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
CV1N	p	dʒ	g	s	t	d	ʃ	z	tʃ	f	v	k	ʒ	b	θ	ð
CV1Q	p	t	d	g	dʒ	z	tʃ	f	s	b	ʃ	v	k	ʒ	θ	ð
VC1N	s	t	p	k	g	b	d	ʃ	ʒ	v	tʃ	z	f	dʒ	θ	ð
VC1Q	k	p	t	s	g	d	f	ʃ	dʒ	tʃ	θ	v	ʒ	b	z	ð
CV2N	l	m	j	r	p	tʃ	s	n	dʒ	h	w	z	f	h^w	b	v
CV2Q	s	z	m	p	l	f	r	dʒ	tʃ	h	b	j	n	v	h^w	w
VC2N	s	p	ŋ	m	n	g	tʃ	b	dʒ	θ	ʃ	f	ʒ	v	ð	z
VC2Q	p	s	g	f	n	m	tʃ	dʒ	ŋ	ʃ	v	b	θ	z	ʒ	ð

inventories syllable-finally, syllable-final consonants are more likely to be among the phonologically natural ones. Thus, it would not be surprising that a bias in favor of natural segment types is found more strongly syllable-finally than syllable-initially. It should be noted, of course, that the phonological naturalness account of these facts does not explain (in the sense of providing proximal causes) why an individual's behavior in a perceptual experiment ought to reflect the distribution of consonants in the world's languages. The account suggests the possibility that this perceptual behavior may form part of the basis for the universal tendency, but no attempt is made, at this point, to explain why an individual's perceptual system should show this bias. (It is at least conceivable that some uninteresting principle like duration is partly responsible for this preference.)

Let us examine the frequency account for the consonant bias before attempting to see how this account can accomodate the CV/VC distinction. Certain consonants can be said to be more expected than others, on the basis of our language experience. When faced with an ambiguous stimulus, we choose the response that is more expected on the basis of this experience. The effect of this long-term experience has been modelled as a permanent criterion shift (in the signal detection sense) for units, depending on their frequencies (see Morton, 1964). While such models have been mainly proposed to account for word recognition, there is no reason not to extend them to segment recognition as well.

There is a major problem with the frequency account of consonant bias as outlined above. In the current analysis, bias correlates significantly with lexical frequency, but generally does not correlate well with frequency of occurrence. This is certainly not what would be predicted by the kind of model proposed to account, for example, for the word-frequency effect. Expectancy should be a function of the actual frequency of experience. This discrepancy, along with the CV/VC distinction suggests the following model to account for consonant bias. Let us suppose that, when presented with a relatively ambiguous nonsense stimulus, the listener's strategy is to sort through possible words that could plausibly include the ambiguous stimulus. The consonant decision is then made by determining which consonant is included in the greatest number of these plausible words. This model would predict, generally, a correlation of response bias with lexical frequency. Let us assume, in addition, that when listening to CV stimuli, the listeners sort through words whose *beginnings* are consistent with the stimulus while for VC stimuli, listeners sort through words whose *endings* are consistent with the stimulus. We could then predict the observed bias difference between CVs and VCs. English speakers are much better at listing words that end with a particular VC# than they are at listing words that begin with a particular #CV. (Baker, 1974). If part of the listener's strategy in the CV and VC recognition task is to match either the beginnings or endings of words (respectively) with the stimuli, then subjects' superior ability in the latter case could result in a much more stable frequency bias.

The model suggested above for accounting for the difference in bias between CVs and VCs is, admittedly, rather baroque. However, it could, in part, be tested. For example, one could examine individual differences in ability to produce words in a Baker-type task. To the extent to which degree of lexical frequency bias can be correlated with ability in the word-finding task, this would support the theory outlined above.

Finally, it is interesting to speculate on why there does not seem to be any bias in speech production errors (as reported by ShattuckHufnagel and Klatt, 1977) comparable to the bias in perceptual errors. Of course, a trivial explanation for this difference could claim that the difference is due to the perceptual results being based on isolated nonsense syllables, or being based on experimental, as opposed to naturally-occurring errors. There is no way, at present, to rule out these possibilites. There are two interesting explanations that are worth considering, however. The first of these would relate the difference to differences between perception and production. In perception, a listener is always faced with uncertainty. (S)he is attempting to map some internal categories onto an ambiguous external signal. Biases help listeners decide how to make their choices. They narrow down the list of alternatives. In fact, since words and phonemes do differ in frequency of occurrence, a comprehension strategy involving response bias in favor of frequent items would lead to the correct response more often than a strategy without such a bias. Passive models of speech perception (e.g., Morton, 1964, 1970), attribute the effect of context in speech perception to bias, effectively. Speech perception could not work at all without such contextual effects.

In short, speech perception can be seen as hypothesis-generation, and bias is one of the many ways that context and knowledge of the world guide this process. Errors in perception are simply hypotheses that happened to be incorrect. Speech errors, on the other hand, are not hypotheses about anything. Speakers generally know what it is they want to say. If we view bias in perception as part of the hypothesis-generating system, there is no reason to find it in production.

The above explanation for the difference between errors in speech perception and production is reasonable, as long as the bias we are discussing is, in fact, useful for the hypothesis-generating system. A frequency bias fulfills this requirement. However, a bias in favor of phonologically natural segments would not seem to be terribly useful to a perceptual system. Thus, there is no explanation for why there should be this kind of bias in perception, but not in production (if, in fact, such a bias could be separated out from frequency). The explanation may lie in the fact that most speech errors (about 80%) tend to involve syllable-initial consonants. While we do not understand why this is so, the failure of bias to show up in speech errors may be because they are mostly syllable-initial consonants, a position that shows only weak bias in perception as well. An

examination of exclusively syllable-final consonant errors might reveal bias in production, comparable to the phonological naturalness bias in perception.

ACKNOWLEDGMENTS

Thanks to Eric Holman for spending many hours discussing bias models and providing various programs. Tom Wickens supplied useful comments and programs. I had helpful discussions with C.P. Browman, Peter Ladefoged and Ian Maddieson. This research was supported by NIH.

REFERENCES

Baker, L.N. 1974. The Lexicon: Some Psycholinguistic Evidence. UCLA Working Papers in Phonetics 26.

Browman, C.P. 1977. Perceptual processing: evidence from slips of the ear. Paper presented at 12th International Congress of Linguists, workshop on slips of the tongue and ear, Vienna.

Carterette, E.C. and Jones, M.H. 1974. Informal Speech. Los Angeles: University of California Press.

Garnes, S. and Bond, Z. 1977. A slip of the ear: A snip of the ear?, a slip of the year? Paper presented at 12th International Congress of Linguists, workshop on slips of the tongue and ear, Vienna.

Goldiamond, I. and Hawkins, W.F. 1958. Vexierversuch: the logarithmic relationship between word-frequency and recognition obtained in the absence of stimulus words. J. exp. Psychol. 56. 457-463.

Goldstein, L.M. 1977a. Categorical features in speech perception and production. Paper presented at 12th International Congress of Linguists, workshop on slips of the tongue and ear, Vienna.

Goldstein, L.M. 1977b. Perceptual salience of stressed syllables. To appear in UCLA Working Papers in Phonetics 38.

Hockett, C.F. 1955. A Manual of Phonology. International J. of Amer. Linguistics, Memoir 11. Baltimore: Waverly Press.

Horn, E. 1926. A Basic Writing Vocabulary. University of Iowa Monographs in Education 4. Iowa City, Iowa.

Howes, D.H. 1957. On the relationship between intelligibility and frequency of occurence of English words. J. acoust. Soc. Am. 29. 296-305.

Greenberg, J.H. 1966. Language Universals. The Hague: Mouton.

Jakobson, R. 1942. Kindersprache, aphasie, und allgemeine Lautgesetze. Selected Writings I. The Hague: Mouton. 328-401.

Janson, T. 1977. Asymmetry in vowel confustion matrices. J. Phonetics 5. 91-96.

Javkin, H. 1976. The perceptual basis of vowel duration differences associated with the voiced/voiceless distinction. Report of the Phonology Laboratory, Berkeley, 1. 78-92.

Klein, W., Plomp, R. and Pols, L.C.W. 1970. Vowel spectra, vowel spaces, and vowel identification. J. acoust. Soc. Am. 48. 999-1009.

Lisker, L. 1974. On 'explaining' vowel duration. Glossa 8. 223-246.

Luce, D. 1959. Individual Choice Behavior. New York: Wiley.

Miller, G. and Nicely, P. 1955. An analysis of perceptual confusions among English consonants. J. acoust. Soc. Am. 27. 338-352.

Morton, J. 1964. A preliminary functional model for language behavior. International Audiology 3. 216-225.

Morton, J. 1970. A functional model for memory. In D.A. Norman (ed). Models of Human Memory. New York: Academic Press.

Roberts, A.H. 1965. A Statistical Linguistic Analysis of American English. The Hague: Mouton.

Ruhlen, M. 1975. A Guide to the Languages of the World. Palo Alto, California: Stanford.

Savin, H. 1963. Word-frequency effect and errors in the perception of speech. J. acoust. Soc. Am. 35. 200-206.

ShattuckHufnagel, S. and Klatt, D.H. 1977. Single phoneme error data rule out two models of error generation. Paper presented at 12th International Congress of Linguists, workshop on slips of the tongue and ear, Vienna.

Shepard, R.N. 1972. Psychological representation of speech sounds. In E.E. David and P.B. Denes (eds). Human Communication: a Unified View. New York: McGraw-Hill. 67-113.

Singh, S. and Black, J.W. 1966. Study of twenty-six intervocalic consonants as spoken and recognized by four language groups. J. acoust. Soc. Am. 39. 635-656.

Trubetzkoy, N. 1958. Grundzüge der Phonologie. Göttingen: Vandenhoeck und Ruprecht.

Verbrugge, R.R., Strange, W., Shankweiler, D.P., and Edman, T.R. 1976. What information enables a listener to map a talker's vowel space? J. acoust. Soc. Am. 60. 198-212.

Wang, M.D. and Bilger, R.C. 1973. Consonant confusions in noise: a study of perceptual features. J. acoust. Soc. Am. 54. 1248-1266

Wish, M. and Carroll, J.D. 1974. Applications of individual differences scaling. In E.C. Carterette and M.P. Friedman (eds). Handbook of Perception II. New York: Academic Press. 449-491.