The Confusion Matrix: A New Model

J.L. Danhauer and L.E. Lucks

Abstract

For years researchers and clinicians have faced the problem of consolidating massive amounts of raw data representing individuals' responses to various speech stimuli into a form that is manageable for either descriptive or statistical analyses. The traditional "confusion matrix" has been extremely useful in organizing such data for various analyses, particularly those involving distinctive feature theory. However, it is difficult to visualize more than one feature at a time on a single matrix. A model is presented here that expands upon the earlier confusion matrix and allows simultaneous visualization of multiple feature distinctions on a single matrix. This matrix is specifically for consonant sounds but could be modified easily to accommodate vowels. Implications for use of the matrix in clinical and research analyses and in teaching distinctive features are discussed.

Introduction

The evaluation of an individual's responses to given stimuli is a vital clinical and research skill in speech-language pathology and audiology. Accurate evaluation of a response to a specific target stimulus is critical for the speech-language pathologist evaluating a patient's production of certain phonemes and for the audiologist assessing a patient's ability to perceive sounds on a speech discrimination test. Evaluation of responses to given stimuli does not present major problems for small amounts of data. However, when the clinician/researcher conducts in-depth assessments of patients'/subjects' abilities to produce or perceive large sets of stimuli, the process can become unwieldy.The difficulty in performing these analyses increases as a function of the number of stimuli in the set. Despite this difficulty, the results of such stimulus-responses (S-R) analyses are critical for making clinical judgements about a patient's speech production or perception abilities and subsequently in determining the course of management. In the research setting, these results are used to determine how subjects produce or perceive certain stimuli, and ultimately they are important in forming the theoretical bases of speech production and perception.

The management of large amounts of data by hand becomes almost impossible as the number of observations increases. For example, speech-language pathologists frequently assess patients' abilities to produce sounds in various contexts such as consonants in pre-, inter- and post-vocalic positions. Multiple observations of the same S-R items are necessary to obtain a representative sample of how the patients produce each target in each context. If the clinician were to evaluate only 4 observations of a patient's ability to produce a set of 24 consonants in each of the 3 positions, the number of responses for analysis would be 288 (24 consonants × 3 positions × 4 observations). This problem is compounded for the researcher who frequently evaluates several responses from several subjects. The number of observations grows rapidly and becomes unwieldy for analysis if only 10 to 15 subjects are tested. The same problems are apparent for studies involving perceptual data (Danhauer, Singh, 1975).

The "confusion matrix" (Miller, Nicely, 1955) has been used by clinicians/researchers for the past three decades to condense and manage large amounts of raw data for analysis. The traditional confusion matrix (Figure 1) serves as a visual representation of the S-R paradigm in which the stimuli are listed down the side of the matrix and the responses are represented across the top in the same order as the stimuli. This may be called a "symmetric" matrix. Thus, by starting at the appropriate intersection or "cell" for the responses, one can determine how the subject responded to the stimulus. The stimuli and the responses are organized symmetrically so that by moving horizontally across the rows and vertically up and down the columns, one can assess what the subject's response is (e.g., when /p/ is the stimulus and /b/ is the response or when /b/ is the stimulus and /p/ is the response). The symmetric confusion matrix contains only responses that are within the stimulus set and is considered to be "square". In some cases, however, the confusion matrix is expanded to the right, making it "rectangular", to show omissions and substitution responses made by the subject that were not part of the original stimulus set. Confusion matrices can be so constructed for both consonants and vowels. In the confusion matrix, the diagonal cells represent correct responses, while confusions (errors) to the target stimuli are shown in the off-diagonal cells. Empty cells indicate that no errors were made for those pairs. The confusion matrix has also been modified to accommodate similarity or dissimilarity judgement data for paired comparison stimuli (Danhauer, Singh, 1975). In this case, all the cells of the matrix are filled, because the subject rates the similarity or dissimilarity of each stimulus paired with all others. These judgements result from the use of procedures such as equal-appearing-interval scaling and magnitude estimation tasks wherein subjects rate the similarity or dissimilarity of paired stimuli on fixed (e.g., 1 to 7) or open-ended scales. Yet another data collection method involves the use of ABX (triadic judgement)

Jeffrey L. Danhauer and Lisa E. Lucks
Department of Speech and Hearing Sciences
University of California Santa Barbara

RESPONSE

STIMULUS	p	t	k	f	θ	s	ʃ	b	d	g	v	ð	z	ʒ	m	n
p	200	10	5	1			14								2	
t	80	150		1		1										
k		5	180	2	3	1				41						
f	20	10	5	122	15		40				10	10				
θ	30	10	5	21	160	3						3				
s	15	16	2	20	10	165	8				4		12			
ʃ	2	2	3	20	2	30	170							3		
b	2							222	2	2	1				3	
d		3						7	210	6		4			2	
g			15							187	10	9	10			1
v				30				12	23	7	110	16	22	1	1	
ð					3			8	27	12	15	152	15			
z						33		2		2	2		193			
ʒ								4	10	18	17	7	63	103		
m								10							205	17
n									1						12	219

Figure 1: **Traditonal confusion matrix using hypothetical data**

schemes wherein subjects receive two stimuli and rate the similarity or dissimilarity of a third stimulus to the first two.

The construction of the initial matrix determines the way the clinician/researcher can use the resulting matrix to perform description analyses of specific *a priori* (predetermined) distinctive feature properties among the stimuli. The confusion (or similarity/ dissimilarity) matrix has also been used to organize data for input to nonparametric analysis procedures, such as the various multidimensional scaling or hierarchical clustering schemes, that can result in *a priori* or *a posteriori* (an undetermined set) features (Danhauer, Singh, 1975; Doyle, Danhauer, Edgerton, 1986; Pruzansky, 1975; Singh, 1976; Wang, Bilger, 1973; Wilson, 1963; Wish, Carroll, 1973). Thus, the traditional confusion matrix has been helpful in condensing large amounts of data for analyses.

A problem exists with the traditional confusion matrix, however, when the clinician/researcher wants to evaluate large numbers of stimuli, condition or subject variables that result in the need for multiple confusion matrices. While tools like the various multidimensional scaling programs (Carroll, Chang, 1970; Dixon, Brown, 1979; Pruzansky, 1975) have helped to address these issues, they may not be appropriate for all types of data. Often the researcher/clinician is interested in a more descriptive type of analysis that can be done by hand or by eye "on-line" rather than by computer. In this case it is difficult to visualize more than one feature on a given matrix. That is, if one wishes to look at stop versus continuancy, it is relatively easy to arrange the stimuli so that these manner of articulation features will be categorized in mutually exclusive groupings. Further, it is also relatively easy to represent various other manner features such as frication, sonorancy, nasality or liquid glide on the same matrix. However, if one is also interested in looking at voicing or place of articulation features, it is difficult and cumbersome to observe these distinctions from the same matrix. In that case, the same data must be rearranged and replotted on other matrices for visualization of these specific features. In cases in which many subjects or conditions are used and the investigator wants to evaluate several *a priori* distinctive features or feature systems (Chomsky, Halle, 1968; Jakobson, Fant, Halle, 1963; Klatt,1968; Miller, Nicely, 1955; Singh, Black, 1966), the number of matrices needed to visualize the possible distinctions is increased substantially, making the task even more unmanageable.

A New Model

Clinicians/researchers have been frustrated with the limitations inherent in the traditional confusion matrix. Having dealt, usually unsuccessfully, with this problem for the past several years, we are proposing a model that expands upon the traditional confusion matrix to account for most of the issues raised here.

Figure 2 depicts a matrix that can accommodate 24 of the English consonants. This matrix looks very similar to the traditional confusion matrix; however, some modifications have been made that help distinguish several features displayed on the same matrix. Note that the percentage of correct scores can still be obtained by tallying the entries in the diagonal cells, and error scores can be computed using the off-diagonal elements. In addition, the matrix has been extended to the right, making it rectangular and capable of accounting for omis-

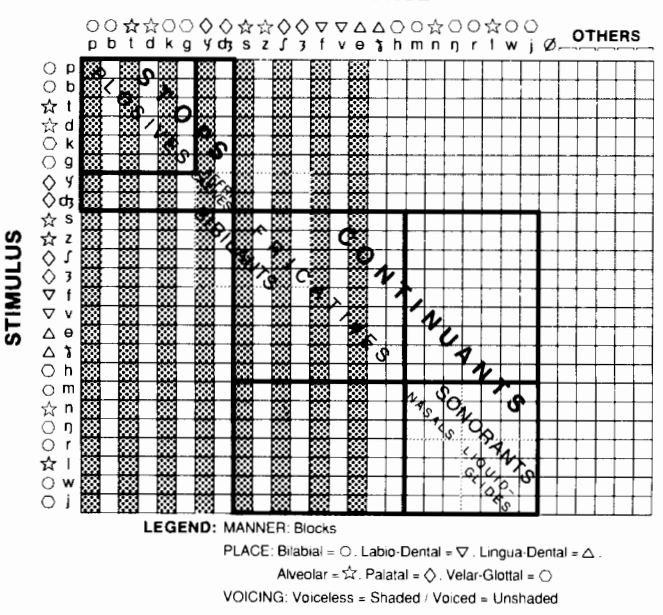

Figure 2: **New model, distinguishing manner, place and voicing on same matrix**

sions and non-stimulus-set substitution responses if needed. Also, certain S-R blocks are highlighted to distinguish manner features, specific cells are shaded to indicate presence or absence of voicing, and a variety of symbols is used to designate place features. Thus, this model encompasses all three features on one matrix without having to replot the data for further feature analyses.

Manner of Production

Boldlined blocks are highlighted on the matrix to indicate the major manner of production feature classes. Moving down the diagonal entries, two major boldlined S-R blocks are highlighted; the first encloses the stop phonemes /p, b, t, d, k, g, tʃ, dʒ/ and the second encloses all the remaining phonemes, which are continuants. Any entry in a cell outside the block enclosing the stops indicates that the response violated the stop distinction. Likewise, responses outside the block for continuants indicate that the continuant feature was violated.

Figure 2 reveals that the stop category is subdivided into an S-R block for plosives /p, b, t, d, k, g/ and another for affricatives/tʃ, dʒ/. Likewise, the continuant block is subdivided into fricatives /s, z, ʃ, ʒ, f, v, θ, ð, h/ and sonorants / m, n, ŋ, r, l, w, j/. Note that a dotted block is used to separate the sibilants /tʃ, dʒ, s, z, ʃ, ʒ/ from the other sounds; this category borrows the affricatives /tʃ, dʒ/ from the stops and /s, z, ʃ, ʒ/ from the fricative continuants. The sonorant block is also subdivided into nasal / m, n, ŋ/ and liquid glide /r, l, w, j/ categories.

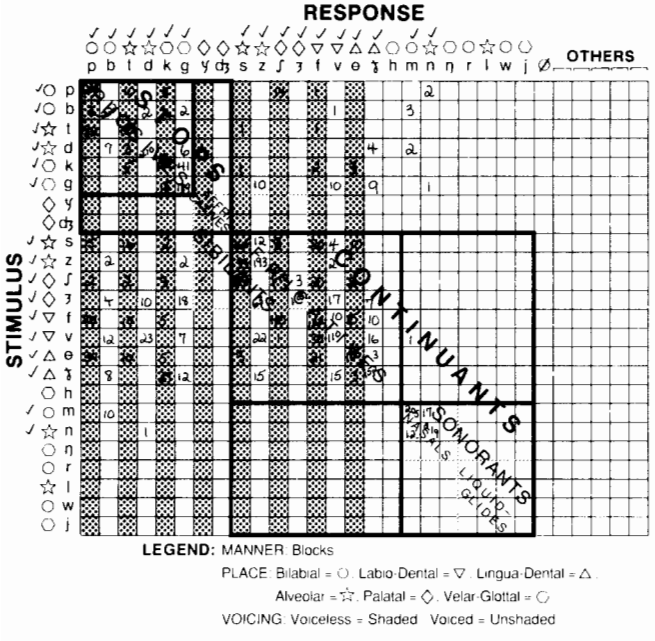

Figure 3: **Application of new model using hypothetical data from Figure 1**

So far, this matrix offers little that could not be gained from using the traditional matrix, assuming the phonemes were listed in the same order. However, visualizing the voicing and place of articulation features on the same traditional matrix is cumbersome and requires replotting of the data on additional matrices with the phonemes listed in different order. This problem has been alleviated somewhat in the new model. Here, shaded cells and symbols are used to help visualize the voicing and place of articulation features as well as manner features.

The hypothetical data from Figure 1 have been replotted on the new model in Figure 3. This model clearly outlines the manner of production features so that correct responses for given manner attributes lie within the boldlined or dotted blocks, and errors are noted in the cells outside. Performance on particular features is easily identified by the labels within the blocks.

Voicing

Shading is used to indicate the presence or absence of voicing — that is, the symbols of voiceless target stimuli in the far left column and voiceless responses across the top are shaded as are the appropriate columns of the matrix; symbols and cells for the voiced stimuli and responses are unshaded. Thus, voiceless stimuli and voiceless responses intersect at shaded cells, indicating that the voicing feature is not violated, while voiceless targets (shaded) perceived or produced as voiced responses (unshaded) intersect at unshaded cells, indicating that voicing is in error. Likewise, voiced stimuli perceived/produced as voiced responses intersect at unshaded cells, showing that the voicing feature is correct, and voiced stimuli perceived/produced as voiceless responses intersect at shaded cells, showing that the voicing feature is incorrect.

Data entries in the cells corresponding to the shaded symbols and columns in the example provided in Figure 3 reveal that the voiceless feature was correctly produced/perceived. Entries in cells corresponding to unshaded symbols and unshaded columns reveal that the voiced feature was correct. Data entries in a shaded cell corresponding to an unshaded symbol or in an unshaded cell corresponding to a shaded symbol indicate that voicing was in error. In this fashion, the voicing feature, which required replotting of the data using the traditional matrix, can now be visualized clearly on the same matrix used for the manner features.

Place of Articulation

A variety of symbols is used to visualize place of articulation. Each symbol in Figure 2 represents a different place category. Although six places are shown, further modification could extend this place division as much as needed; for example, as many as seven place distinctions have been used in some feature systems (Singh, Black, 1966). The symbols indicating place of articulation are provided for each stimulus and response

phoneme listed along the outsides of the matrix. The bilabials /p, b, m, r, w/ are designated by O, the labiodentals /f, v/ by ▽, the lingualdentals /θ, ð/ by △, the alveolars /t, d, s, z, n, l/ by ★, the palatals /ʃ, ʒ, tʃ, dʒ/ by ◇, and the velars and glottals /k, g, ŋ, j, h/ by ○. Note that the phonemes /r, w/ are classified with the bilabials on the basis of their visual rather than their acoustic qualities, and /h/ is grouped with the velars, rather than using a separate class for glottal, for economic reasons.

Thus, by matching the symbols at the cell intersecting a stimulus and a response, one can determine if the place of articulation feature is in error. For example, in Figure 3 /b/, noted by the symbol O, was perceived as /p/, also noted by O; thus, place of articulation was correct because both phonemes share the bilabial place. Note that the ★ for the alveolar /d/ response did not correspond to the O denoting the stimulus /b/; because the S-R symbols failed to match, the subject erred on place. While the place distinctions are less evident, it is possible to see whether the stimulus and the response share manner, voicing or place characteristics all on the same matrix.

While this matrix accommodates 24 consonants, clinicians/researchers may use the same matrix for studies having smaller stimulus sets by placing a checkmark by the stimulus phonemes listed down the left side of the matrix appropriate for their stimulus sets, as noted in Figure 3. When clinicians/researchers use closed-set response modes (i.e., the S-R set is limited for the subject), all responses should fit within the square matrix. If an open-set response mode is used, the subjects' responses that were not included in the stimulus set can be accounted for across the top of the matrix. That is, any non-stimulus-set responses can be indicated to the right, producing a rectangular matrix. Further, separate columns have been reserved for omissions indicated by the symbol (Ø) and for nonstimulus responses indicated by "others", which also provide spaces for clinicians/researchers to add their own symbols or phonemes.

Use of the Matrix

The matrix presented here is very similar to the traditional confusion matrix, but it provides some modifications that permit visualization of multiple features on the same matrix. The matrix is not necessarily "exhaustive", in that every phoneme may not be totally distinguished from all the others on the matrix. Further, this matrix does not account for all possible consonants (e.g., the /ʍ/or /hw/ common to the midwestern dialect is missing). Also, this matrix may not cover all possible parameters of the stimuli; other features could be used. In particular, the features included here are more phonologically than acoustically based.

At first this matrix may appear more cumbersome than the traditional matrix, but it should help distinguish features. Application of this matrix should simplify descriptive analyses of data as well as the preparation of raw data for submission to the various multivariate analysis programs available.

This matrix may also be useful in teaching distinctive feature theory and its application to students and future clinicians. The teacher can use this matrix to demonstrate the features common to any specific S-R paradigm. By inspecting the cells of the matrix, the student can determine the category or block in which a response falls, whether or not shading is present, indicating the voicing characteristic and what symbol is used to show place of articulation. If the response to a given stimulus is outside the manner block, the subject has erred on manner of production; if the shading is not consistent, the subject has erred in voicing; and if the symbols do not line up, the subject has erred on place of articulation. It should thus be easier to determine whether a response is correct or by how many features it differs from the target.

We hope that clinicians/researchers can use this model to analyse data in a simpler fashion than was possible with the traditional confusion matrix. Also, similar matrices can be constructed easily for vowels. While the new model may not be in its final form, may need further modification and may not be useful in all situations, we hope it will prompt clinicians/researchers to use it and investigate alternative ways of handling large amounts of data.

References

Carroll, J., Chang, J. (1970). Analysis of individual differences in a multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition. *Psychometrika, 35,* 283-319

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

Danhauer, J., Singh, S. (1975). *Multidimensional Speech Perception by the Hearing Impaired: Treatise on Distinctive Features.* Baltimore: University Park Press

Dixon, W., Brown, M. (1979). *Biomedical Computer Programs (BMDP), P-Series (Computer Program).* Los Angeles: University of California Press

Doyle, K., Danhauer, J., Edgerton, B. (1986). Vowel perception: experiments with a single-electrode cochlear implant. *Journal of Speech and Hearing Research, 29,* 179-192

Jakobson, R., Fant, G., Halle, M. (1963). *Preliminaries to Speech Analysis: the Distinctive Features and Their Correlates.* Cambridge, Massachusetts: MIT Press

Klatt, D. (1968). Structure of confusions in short-term memory between English consonants. *Journal of the Acoustical Society of America, 44,* 401-407

Miller, A., Nicely, P. (1955). An analysis of perceptual confusions among English consonants. *Journal of the Acoustical Society of America, 27,* 338-352

Pruzansky, S. (1975). *How to Use SINDSCAL, A Computer Program for Individual Differences in Multidimensional Scaling.* Murray Hill, New Jersey: Bell Telephone Laboratories

Singh, S. (1976). *Distinctive Features: Theory and Validation.* Baltimore: University Park Press

Singh, S., Black, J. (1966). Study of twenty-six intervocalic consonants as spoken and recognized by four language groups. *Journal of the Acoustical Society of America, 39,* 372-387

Wang, M., Bilger, R. (1973). Consonant confusions in noise: a study of perceptual features. *Journal of the Acoustical Society of America, 54,* 1248-1266

Wilson, K. (1963). Multidimensional analysis of confusions of English consonants. *American Journal of Psychology, 76,* 89-95

Wish, M., Carroll J. (1973). Applications of "INDSCAL" to studies of human perception and judgment. In Carterette, E.C., Friedman, M.P. (eds.) *Handbook of Perception.* 4th ed. New York: Academic Press, Inc.

Acknowledgements

Portions of this work were supported by a University of California Santa Barbara (UCSB) General Research Grant (#8-587529-19900-7) and a Biomedical Research NIH Grant (#RR07099-18-8-448790-24328) administered by the UCSB Social Processes Research Institute.