Acoustic confusion of digits in memory and recognition*

B. J. T. MORGAN†, S. M. CHAMBERS††, and J. MORTON

M.R.C. Applied Psychology Unit, 15, Chaucer Road, Cambridge, CB2 2EF, England

It has been shown that, contrary to current opinion, digits do not confuse arbitrarily and that similar patterns of confusion are observed in memory and auditory recognition tasks. Two alternative measures of stimulus similarity are compared.

Existing work on the confusion of digits has yielded no *structure* of confusion, that is, the type of structure that we observe in the acoustic confusion of letters (see, for example, Conrad, 1964; Miller & Nicely, 1955; Morgan, 1973a). For example, Moser and Dreher (1954) make the following comment on the results of their recognition study: "The most serious digit confusions involve 2, 3, and 6." Similarly, we may quote memory studies—for example, "The analysis of digit confusions indicates that there is no feature of the immediate memory function which could lead to systematic confusions of one digit with another for whatever reason. It appears that if confusions occur, they must be ascribed to weakness elsewhere in the communication system [Conrad, 1958]." "Digit 3 was too frequently substituted for 2, but the reverse substitution was not different from chance expectation. It is difficult to think of any unique principle which might underly this substitution, so that the question whether substitutions are due to perceptual or recall error has little significance [Conrad, 1959]." "The present experiment confirmed . . . that there are no systematic confusions between digits in short-term recall [Wickelgren, 1965]."

While carrying out a number of experiments with serial recall of digits, we noticed that Ss often commented (in contradiction to the results mentioned above) that they tend to confuse the digits 1, 5, and 9. Frequently, Ss are poor guides to their own performance, but, indeed, a preliminary analysis of the results of these experiments seemed to show that the Ss were, in fact, persistently confusing these digits. It seemed worthwhile making a more thorough investigation of the data, and this paper describes the results obtained and also the results of a *recognition* task; we were interested to find out whether the same confusions occurred in the recognition task as seemed to occur in the memory task, a relationship which Conrad (1962) showed for letters.

METHOD

As this is a post hoc investigation, the method contains several factors not utilized in the following analysis. It was assumed that these factors do not interact with the mechanism of confusion. Ss were housewives, aged 20 to 60, drawn from the APU Ss panel.

(a) Memory of Digits

Data were analyzed from 29 experiments where Ss were presented with lists of sometimes 8, sometimes 7, digits that were recalled serially. The stimulus digits were presented by two female voices, one an American (J.F.), the other an Australian (S.C.). The majority of the experiments included conditions with a stimulus suffix which differed from the digits in location, pitch, semantic content, or "speech likeness," depending on the particular aspect being investigated (see Morton, Crowder, & Prussin, 1971). With the J.F. stimulus lists, 230 Ss were presented with 23,652 digit lists in 12 experiments. A further 328 Ss listened to 24,750 digit lists in 17 experiments in which the digits were recorded in S.C.'s voice. The digits, presented at the rate of 2/sec, came from the set 1-9 and were ordered so that each digit occurred equally often at each serial position. The following variations were included in a small number of the experiments: Ss were required to prefix their recall with a tick or cross (three experiments); Ss were required to recall the final four digits before the first four (one experiment); Ss were required to recall the digits in reverse order (one experiment); the digits were presented to alternate ears (one experiment); the digits were presented at the rate of 1/sec (three experiments). The resulting confusion matrices did not seem, from cursory inspection, to show interaction with type of stimulus, and thus results were pooled to give a single confusion matrix per voice.

The total possible lists that Ss could have recalled correctly was 48,402. A list was included in the analysis only if the S had a single error on that list. This restriction was necessary as we were interested only in errors brought about by confusions between digits. Multiple errors in a list could have a different etiology, and including them would make the data noisy (cf. Conrad, 1962). There was, therefore, a possibility of 5,378 errors for each digit. The final overall proportion of errors was 6.5%.

(b) Recognition of Digits
[Different Ss From Those In (a)]

Four blocks of digits were recorded. two with a female voice (S.C.) and two with an English male voice (J.M.). There were 15 presentations of each digit from the set 1-9, making a total of 135 items per block; balancing with respect to serial position occurred as in (a). The digits were recorded on one channel of a Vortexion tape recorder. Bursts of white noise were recorded on the second channel to mask the digit completely, with about a 4-sec gap before the next item (an item was presented every

Table 1
Acoustic Confusion of Digits, Recognition Task, Male Voice (JM)

Response	Stimulus									Total
	1	2	3	4	5	6	7	8	9	
1	188	23	33	55	21	20	31	21	56	448
2	9	117	60	16	2	15	23	30	6	278
3	12	161	143	20	5	29	31	44	5	450
4	17	33	37	300	7	21	30	36	15	496
5	150	16	20	58	445	11	41	32	113	886
6	8	83	37	10	0	346	38	56	2	580
7	21	51	71	29	5	52	274	59	7	569
8	16	32	60	18	6	27	30	219	11	419
9	119	21	62	29	51	14	29	34	324	683
Total	540	537	523	535	542	535	527	531	539	4809

Table 2
Acoustic Confusion of Digits, Recognition Task, Female Voice (SC)

Response	Stimulus									Total
	1	2	3	4	5	6	7	8	9	
1	770	96	121	98	128	77	122	114	218	1744
2	41	385	224	48	35	94	79	91	60	1057
3	56	438	480	54	35	100	85	141	49	1438
4	63	89	125	1023	67	63	100	177	61	1768
5	254	66	81	119	937	64	83	59	324	1987
6	30	110	98	38	16	805	110	242	22	1491
7	41	189	168	64	26	191	807	154	51	1691
8	64	87	139	58	28	111	109	513	48	1157
9	242	96	119	60	292	44	65	55	720	1693
Total	1561	1556	1555	1562	1564	1549	1560	1546	1553	14006

5 sec). The signal and noise were mixed using external mixing gear and played to the groups of Ss over a loudspeaker. The S/N ratio was approximately −10 dB, this being determined by a preliminary test on several Ss to discover that ratio at which Ss would be making the target of 50% errors. The final overall proportion of errors was 53%. For the first block, the signal was begun at 5 dB higher than this level and turned down slowly to the stated level after 60 items. This procedure incurred an approximately constant error probability in spite of the Ss' becoming more practiced. Ss were asked to write down the digit as they heard it, on prepared response forms, guessing if unsure rather than leaving blanks. Ss listened to the digit channel without the noise for 7 items and were then given 14 items for practice before beginning each block. Thirty-nine Ss were presented with one block of digits spoken by the female voice. A second group of 28 Ss was presented with four blocks of digits, the first two being spoken in the female voice, the last two by the male voice. Again, the results for individual Ss were pooled to give two confusion matrices, one for the male voice and one for the female.

Analysis

Tables 1 and 2 present the confusion matrices for the two voices in the recognition task; Tables 3 and 4 present the confusion matrices for the two voices in the memory task.

In previous work, analysis of such matrices has been on the "micro" level, comparing observed confusions with those that would have occurred "by chance," i.e., those that would have occurred had each digit been equally confusable with each other digit. Such an analysis is inappropriate here as it is clear that there is a nonuniform distribution of confusion in all four tables.

For the skeptical, however, let us take as an example the confusions for 1 from Table 4. We have·

Response	2	3	4	5	6	7	8	9
Stimulus 1	28	28	34	40	10	42	31	32

Under the null hypothesis of equal confusion, the expected number in each category of response is 30.63, yielding a chi square of 21.87 on 7 df, significant at the 0.5% level. This result may be due to a genuine nonuniform confusability or just due to a bias in response; but perhaps it is not merely fortuitous that 7 is the digit most frequently confused for 1!

An *overall* test of independence of confusion that may be applied to the entire matrix is a test of "quasi-independence." This term is explained in papers by Goodman (1968) and Morgan[1] and is appropriate in that, in testing for it, we test for equal confusability, allowing for response bias. The test developed by Morgan tests the null hypothesis,

$$H_o: P_{ij} = \frac{P_j}{1 - P_i}, i \neq j,$$

where P_{ij} is the probability that Response j is made for Stimulus i, and the (P_j) form a probability distribution, so that for all i,

$$P_i > 0,$$

and

$$\sum_{i=1}^{9} P_i = 1.$$

If H_o is true, then the presence of bias would imply that the P_i were not all the same.

Table 3
Acoustic Confusion of Digits, Memory Task, Female Voice (JF)

Response	Stimulus									Total
	1	2	3	4	5	6	7	8	9	
1	2340	4	16	13	25	4	4	13	49	2468
2	23	2533	54	42	24	11	14	13	28	2742
3	34	28	2465	32	26	19	5	12	21	2642
4	35	12	13	2421	56	7	3	14	10	2571
5	48	6	8	42	2315	10	8	7	72	2516
6	11	12	19	11	17	2506	7	19	11	2613
7	33	21	24	34	58	45	2570	21	25	2831
8	41	6	22	21	34	22	9	2508	25	2688
9	63	6	7	12	73	4	8	21	2387	2581
Total	2628	2628	2628	2628	2628	2628	2628	2628	2628	23652

Table 4
Acoustic Confusion of Digits, Memory Task, Female Voice (SC)

Response	Stimulus									Total
	1	2	3	4	5	6	7	8	9	
1	2505	11	7	25	23	0	6	16	42	2635
2	28	2609	35	31	23	9	6	17	21	2779
3	28	31	2612	23	35	14	11	13	13	2780
4	34	16	16	2530	43	11	7	14	9	2680
5	40	11	13	54	2426	4	5	18	44	2615
6	10	24	24	23	22	2654	14	23	9	2803
7	42	19	19	24	51	37	2692	24	28	2936
8	31	15	14	13	37	17	4	2600	18	2749
9	32	14	10	27	90	4	5	25	2566	2773
Total	2750	2750	2750	2750	2750	2750	2750	2750	2750	24750

It can be shown that (see Altham, 1970) H_0 is equivalent to the following hypothesis (for more than three stimuli):

$$\frac{P_{ik}}{P_{il}} = \frac{P_{jk}}{P_{jl}}$$

for all possible sets of four (i, j, k, l) so that no two are the same. This is clearly what we mean when we talk in terms of equal confusability.

The resulting test statistic has a chi-square distribution (under H_0) on $(s^2 - 3s + 1)$ degrees of freedom, where s is the number of stimuli. Here s = 9 so that the degrees of freedom are 55. Applying this test to the data of Tables 1, 2, 3, and 4, we obtain the values of Table 5. As $\chi^2_{55}(0.1\%) \approx 90$, the absurdly high values above imply rejection of H_0 in all four cases. This was a case when the data were so obviously significant that a statistical test was really unnecessary!

Therefore, we have established nonuniformity of confusion; we must now examine this nonuniformity in more detail.

Shepard (1963) emphasized the utility of nonmetric multidimensional scaling (we shall refer to this technique as "MDSCAL"; for a description, see Kruskal, 1964a, b) as a means of illustrating and summarizing large matrices of data by a single "macroanalysis"; in his case, some of the matrices were of Size 36. Although the stimulus-response matrices of this study are only of Size 9, it was decided that macroanalyses such as MDSCAL would be the most appropriate here too. Thus, the matrices were all analyzed by the complementary techniques, MDSCAL and single-link cluster analysis (we shall refer to this technique as "SLINK"; for a description, see Johnson, 1967). This paper assumes a basic familiarity with these two techniques (see, e.g., Morgan, 1973a; Shepard, Romney, & Nerlove, 1972), but the following preliminary comments will be made.

Preliminaries

The MDSCAL representations sought were in two dimensions.

the resulting picture then being the most evocative (see, e.g., Kendall, 1969, 1971). Fortunately, the stress values obtained were never too high to force us to try a fit in a higher dimensional space. As an aid to interpreting the MDSCAL configurations, points with the "strongest" similarities are linked together. The MDSCAL results given below were the best (in terms of small stress) of several separate runs on each data table. Each run used global scaling and primary treatment of ties and utilized a different random initial configuration. All runs continued for 50 iterations, by which time the procedure had, in all cases, converged to the final solution.

SLINK is a continuous hierarchical cluster method, providing as output a dendrogram which can be characterized by an *ultrametric* function (see Johnson, 1967). In particular, this enables us to measure the distortion imposed on the original similarities by SLINK and also to compare two dendrograms (see Jardine & Sibson, 1971). The ultrametric quality of a SLINK dendrogram can be a drawback of the method, for if a large distortion has been imposed on the data by SLINK, then further SLINK analyses on similar data could result in large (but still continuous) changes in the dendrogram. Gower (1971), for example, considers that too much emphasis has been placed on the continuity of a cluster-analysis method. Nevertheless, we intend to compare dendrograms, and for such comparisons the continuity of the method that gives the dendrograms is of vital importance.

Before either of these techniques (MDSCAL and SLINK) may

Table 5
Test Statistics for Independence of Confusion in Tables 1, 2, 3, and 4

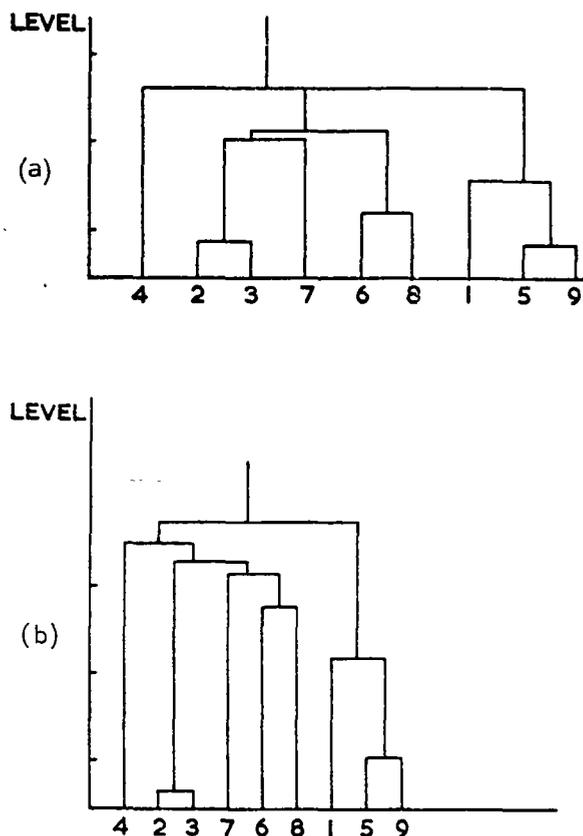| Table	Test Statistic
I	1121.62
II	2619.98
III	375.26
IV	274.76

Fig. 1. SLINK dendrograms for the two voices on the recognition task (Eq. 1 used for similarity). (a) Male voice (J.M.). We have taken as the measure of the distortion imposed by each dendrogram on the original dissimilarity matrix the $\hat{\Delta}_1$ measure suggested by Jardine and Sibson (1971, p. 104). Here, $\hat{\Delta}_1$ = 0.18. (b) Female voice (S.C.); $\hat{\Delta}_1$ = 0.11.

be applied, the matrix of stimulus-response scores must be converted to a symmetric similarity matrix:

$$\Sigma = (\sigma_{ij}, \, i,j = 1, 2, \cdots, 9),$$

where σ_{ij} denotes the similarity between elements i and j. The most obvious conversion to use is,

$$\sigma_{ji} = \sigma_{ij} = \frac{r_{ij}}{r_{i\cdot}} + \frac{r_{ji}}{r_{j\cdot}},$$ (1)

where r_{ij} is the number of times j was the response to stimulus i and

$$r_{i\cdot} = \sum_{j=1}^{9} r_{ij}, \, etc.$$

A common problem that arises, however, is that the data matrices tend to be asymmetric due to a biasing of the responses; the presence of bias has been a problem since the very first studies of this kind (see Shepard, 1963). Were bias an important factor, Eq. 1 might become doubtful as a measure of similarity. This problem has already been considered by Townsend (1971) and Van der Kamp and Pols (1971); in particular, one means of allowing for bias is to assume what is essentially Luce's choice model (1963), i.e., that one can write,

$$p_{ij} \propto b_j \, f(\sigma_{ij}),$$ (2)

where the constant of proportionality is adjusted to ensure $p_{i\cdot}$ = 1 for all i, f is a monotonic increasing function, and b_j represents the biasing of responses towards j. Equation 2 thus separates bias from similarity. If one assumes σ_{kk} to be a constant for all k, then it is not difficult to show that

$$f(\sigma_{ij}) = \frac{P_{ij} \, P_{ji}}{P_{ii} \, P_{jj}},$$

and as both MDSCAL and SLINK give the same representations under monotone transformations of the similarity scale (see Sibson, 1972), we may take,

$$\sigma_{ij} = \sigma_{ji} = \frac{P_{ij} \, P_{ji}}{P_{ii} \, P_{jj}},$$

estimated from the data by the cross ratio (see Edwards, 1963),

$$\sigma_{ij} = \frac{r_{ij} \, r_{ji}}{r_{ii} \, r_{jj}}.$$ (3)

Clearly, problems could arise in estimating σ_{ij} from Eq. 3 when the data available are rather scanty; in particular, if many of the r_{ij} are zero, then one might well obtain the type of distortion noted by Kendall (1970) in which a straight line becomes bent, in a two-dimensional MDSCAL representation, to give a "horseshoe." However, the above matrices have no zero entries, and so all analyses were made on both Eqs. 1 and 3. The results on Eq. 3 were much the same as those on Eq. 1, and so the Eq. 3 analyses are given (for comparative purposes) only in Fig. 4.

The model (Eq. 2) is, when the size of the stimulus-response matrix is (3) (it is 9 above, of course), a statement of quasi-independence, and thus the fit of the model could be tested statistically (Morgan, 1972). When the size of the matrix is greater than 3, maximum likelihood estimates of b_j and σ_{ij} could be obtained (though the method would probably require numerical iteration), and the fit of the model tested by means of a chi-square goodness-of-fit test (see Grey & Morgan, 1972, for a description of this procedure).

RESULTS

In both the memory and recognition tasks, we had two voices; all four of the resulting confusion matrices were analyzed by SLINK and MDSCAL, once using Eq. 1 and once using Eq. 3. A statistical test of the equality of the confusion probabilities underlying, say, Tables 1 and 2 (comparing the two voices on recognition) or Tables 4 and 2 (comparing the two tasks on the same voice) probably could be devised on the lines of Goodman (1968) and Morgan (1972). However, it is felt that the present form of analysis is adequate for the purposes of this paper.

First of all, on both memory and recognition tasks, there was little difference between voices on either SLINK or MDSCAL, whichever equation, 1 or 3, was used. For example, taking just Eq. 1, for the recognition task, see Fig. 1, while for the memory task, see Fig. 2.

We may compare the dendrograms of Figs. 1 and 2 by using the measure of dissimilarity, Δ_1^*, suggested by Jardine and Sibson (1971, p. 106); we then obtain the results of Table 6.

We see, as expected, that within task and between voices, dissimilarities are much smaller than between task dissimilarities. Also, as Δ_1^* can take on values in the range $(0, 1)$, we see that the between-task dissimilarities are by no means very large. These, of course, would be our conclusions on just looking at Figs. 1 and 2, but Table 6 quantifies these conclusions. It would be of interest to conduct a Monte Carlo study of the significance of these Δ_1^* values (cf. Klahr, 1969; Stenson & Knoll, 1969; Wagenaar & Padmos, 1971).[3]

The similarity between voices was also present in the MDSCAL analyses; we thus felt justified in basing further analysis on the data obtained by pooling over voices to give a single confusion matrix for the recognition task and a single confusion matrix for the memory task.

A secondary aspect of this analysis, taken apart from the actual confusions, is the comparison of Eqs. 1 and 3 (see Figs. 3 and 4, respectively, for comparison). For the individual voices, the Eq. 3 dendrograms (not given here) were more "condensed" than those of Eq. 1. MDSCAL plots in two dimensions had an improved fit (smaller stress) when Eq. 3, rather than Eq. 1, was used. However, when the two measures are compared on the data matrices pooled over voices, the slight changes noted above become quite distinct: the stress values reduce from 6.44% to 0.44% (recognition) and from 14.60% to 10.70% (memory). The Eq. 1 dissimilarities

Table 6
Measures of Dissimilarity Between the
Dendrograms of Figs. 1 and 2

| | Recognition | | Memory | |
	(JM) Fig. 1a	(SC) Fig. 1b	(JF) Fig. 2a	(SC) Fig. 2b
Fig. 1a		0.07	0.22	0.24
Fig. 1b			0.18	0.19
Fig. 2a				0.07
Fig. 2b				

are little distorted by SLINK, whereas (and probably consequently) the Eq. 3 dissimilarities suffer a much larger distortion [$\hat{\Delta}_1$ = 0.74 rather than 0.11 (recognition) and 0.58 rather than 0.18 (memory)]. This is of interest, as previous studies comparing Eqs. 1 and 3 found little difference between the two measures (Townsend, 1971; Van der Kamp & Pols, 1971).

The transformation of the data represented by Eq. 3, as it incorporates the product $r_{ij}r_{ji}$, tends to increase the ratio of large:small similarities obtained under Eq. 1, and thus fairly distinct clusters under Eq. 1 can be expected to become more condensed and separated under Eq. 3. Compare Figs. 3a and 4a: the clusters of Fig. 4a are the L* clusters of Van Rijsbergen (1970), in which every within-cluster dissimilarity is less than every dissimilarity between any object in the cluster and any object outside it. The previous studies comparing Eqs. 1 and 3 had no fairly distinct clustering under Eq. 1, and so did not produce this condensing effect when changing from Eq. 1 to Eq. 3.

When we have a set of stimuli which separate into L* clusters, then MDSCAL could reveal no more information than just what these clusters are: it would just have to put all the stimuli in an L* cluster at a single point and then perform an MDSCAL analysis on the resulting points. Alternatively, MDSCAL might find a local minimum such as that of Fig. 4a, which is adequate for our purpose. This possibility of MDSCAL collapsing has been noted by Shepard (1962, p. 240).

Finally we come to the confusions themselves: the picture is the same whichever measure, Eq. 1 or Eq. 3, is used. The recognition confusions are more clearly structured than the memory confusions (cf. Table 5). In recognition the strongest confusors are 2 3 and 1 5 9; Figs. 3 and 4 indicate two major confusiong classes: 1 5 9 and 2 3 6 7 8. In memory, the MDSCAL fits in two dimensions are not as good as in recognition; Fig. 5 indicates the major confusing classes, 1 4 5 9 and 2 3, with 6, 7, and 8 all remaining fairly isolated. The strong confusions of the recognition task (with the exception of 4) are seen to be the strong confusions of the memory task (cf. Conrad, 1962).

An alternative way of comparing the data of Tables 1-4 is by simply computing correlation coefficients for the between-table confusions. In Table 7, the rank-order correlations between the columns of Tables 1-4 (omitting correct responses) are given, together with the
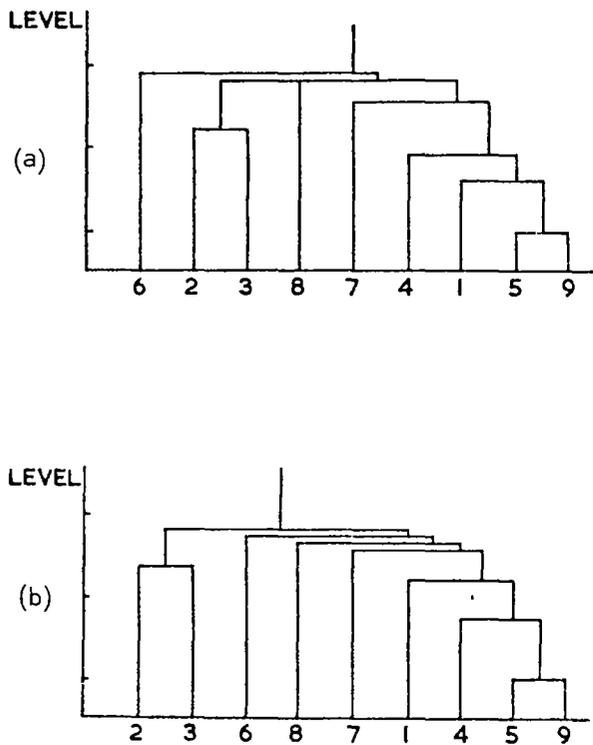
LEVEL

(a)

6 2 3 8 7 4 1 5 9

LEVEL

(b)

2 3 6 8 7 1 4 5 9

Fig. 2. SLINK dendrograms for the two voices on the memory task; Eq. 1. (a) Female voice (J.F.): $\hat{\Delta}_1$ = 0.20. (b) Female voice (S.C.): $\hat{\Delta}_1$ = 0.16.
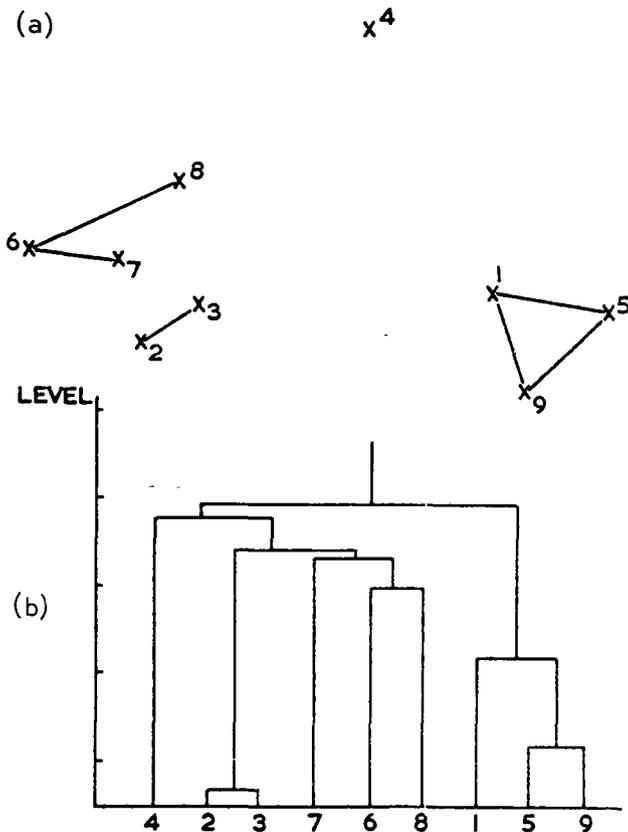
(a) x^4 $(\tilde{r})$ $_x4$

Fig. 3. Recognition task pooled over voices; Eq. 1. (a) MDSCAL: stress = 6.44%. (b) SLINK dendrogram; $\hat{\Delta}_1$ = 0.11.

Fig. 4. Recognition task pooled over voices; Eq. 3. (a) MDSCAL: stress = 0.44%. (b) SLINK dendrogram; $\hat{\Delta}_1$ = 0.74.

average Spearman correlations (Lyerly, 1952) for each pair of tables. A similar analysis could be performed for the rows, but there are difficulties in interpreting such a row analysis. Thus, we see that when we confuse the digit 1 (voice, S.C.), the confusions in the memory task have a Spearman correlation of 0.874 with those in the recognition task, etc.

For all six comparisons, the average Spearman correlation coefficients are highly significant (in all cases, $p < .0005$, see Lyerly, 1952). The within-task average correlations are greater than the between-task correlations (cf. Table 6).

W. S. Mueller, of Monash University, has analyzed the data of Tables 1-4 by the technique of INDSCAL (see Carroll & Chang, 1970), and his results, not yet published, form an interesting complement to the results obtained here. The "subject space" of the INDSCAL analysis contains just four points corresponding to the four tables, 1-4.

DISCUSSION

In order to compare the results of these experiments with those of previous studies, we must take into account several extraneous factors. For example, in their recognition study, Moser and Dreher (1954) have a
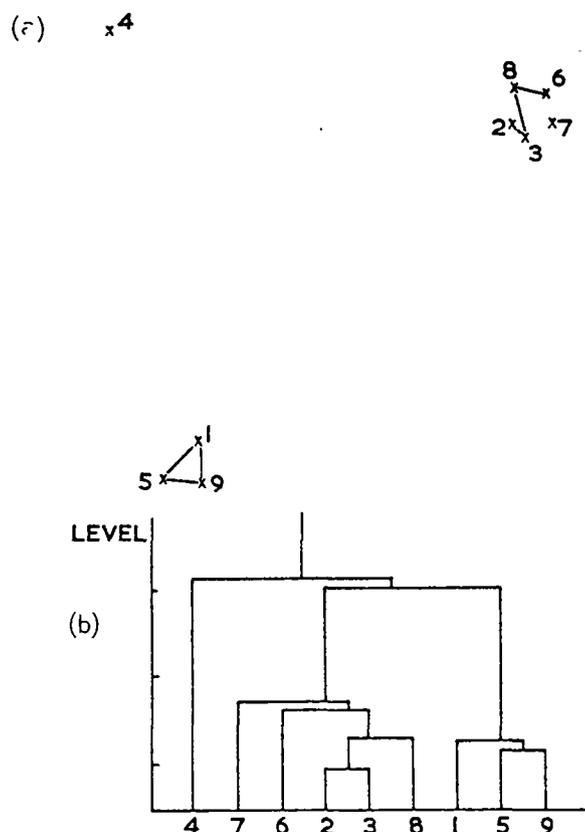
stimulus set containing letters and elements of a modified ICAO alphabet, as well as digits. Unless one subscribes to the somewhat dubious (see Cane, 1960; Morgan, 1973b) "constant ratio rule" (CRR) (Clarke, 1957), also known as "Luce's choice axiom" (Luce, 1959), then these factors will influence the predictions of the experiments and confound any comparisons one might wish to make between experiments. However, a MDSCAL analysis of the acoustic confusions in a letter-digit alphabet containing all the letters and digits (35 elements in all) has resulted in a fairly good fit in three dimensions (stress = 12.50%) and provided a very similar picture, as regards the digit confusions, to those of Figs. 3a and 4a, in that 1 5 9 and 2 3 stood out as the most strongly similar sets (see Morgan, 1973a).

Moser and Dreher also used "fife" and "niner" as alternative pronunciations of 5 and 9. The existence of these alternatives might have affected the error patterns, contra the CRR, and eliminated the 1 5 9 errors from their results. It is not clear why the other studies cited in the introduction failed to find syntactic errors in data from memory experiments since our own data is unequivocal.

We have shown, then, in both recognition of digits in noise, and in serial recall of digit strings, that there is strong evidence for structure in the error data. In both

situations, there are two fairly well-defined groups, 2 3 and 1 5 9. In recall, 4 is closely tied to the 1 5 9 group. In recognition, 2 3 are more highly confused than they are in recall, compared, for example, with 5 9.

The interpretation of these data is not clear. Previous work on letter confusions (Conrad, 1962, 1964; Wickelgren, 1965; Morgan, 1973a) would lead us to expect that items with similar vowel sounds would be likely to confuse both in memory and in recognition. However, with the exception of "five" and "nine," there are no exact vowel correspondences in the digit set. Thus, our account will have to use other dimensions of dissimilarity. Morgan (1973a) found the same problem in trying to disentangle the microstructure of letter name confusions. Indirect evidence that the confusions were complex lay in the relation between the stress measure and the number of dimensions used in MDSCAL analysis. If the number of dimensions of dissimilarity is small, one would find that the stress measure would fall rapidly and then level out to an asymptote. Morgan, however, found that the stress measure showed no signs of reaching an asymptote, even up to six dimensions. Klatt (1968), in an analysis of Wickelgren's (1966) data on consonant confusions in memory, concludes that at least five dimensions are exerting an influence. We can expect digit confusions to be more complex than this.

The simplest account we can come to is based on the following premises:

(P1) When recognizing items in noise, the first phoneme tends to be lost. Some asymmetries in response matrices could be accounted for in this way. Thus, Morgan (1973a) found that in the data he used, A was given as an incorrect response for K on 688 occasions. The reverse happened only 54 times. The same pattern held for JA confusions as well as for confusions of E with B, C, D, G, P, T, and V. These patterns would follow if the initial consonant was completely lost. In memory, such a phenomenon would not be expected.

(P2) The first formant of the vowel is more important than the second. In articulatory terms, this means that the tongue height is more important than its place. The reason for supposing this lies in the strong links between
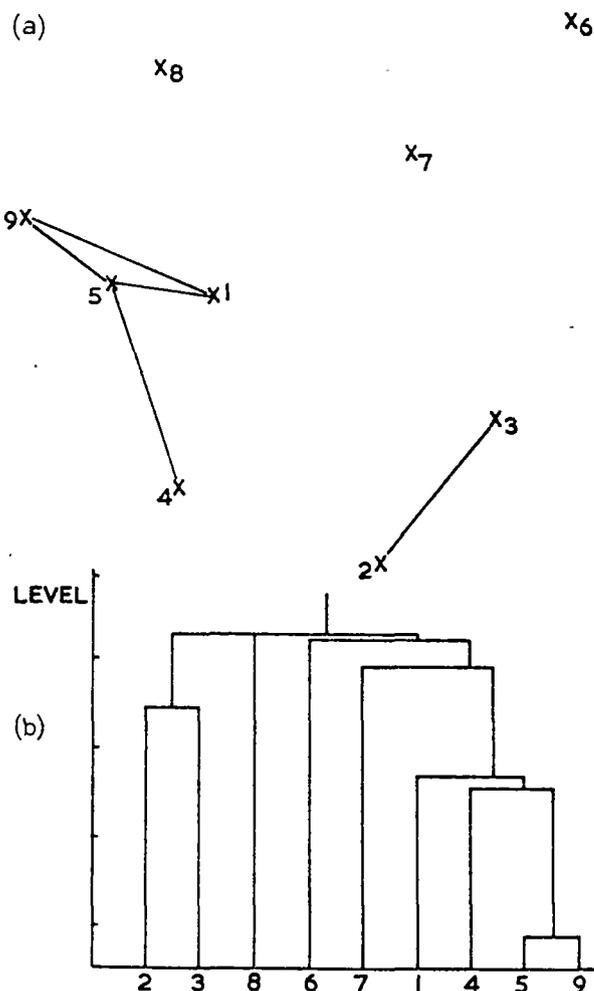

Fig. 5. Memory task pooled over voices; Eq. 1. (a) MDSCAL: stress = 14.6%. (b) SLINK dendrogram; $\hat{\Delta}_1$ = 0.18.

"two" and "three," especially in recognition. The vowels /u/ and /i/ have only the first formant in common. The 2-3 confusion was also found by Moser and Dreher. As the energy in speech falls off with frequency, we would expect the second formant to be more highly masked by white noise. This should not hold in memory.

Table 7
Spearman Rank-Order Correlations Within Columns, Between Tables 1, 2, 3, and 4

Stimulus (Column)	Tables					
	1, 2	1, 3	1, 4	2, 3	2, 4	3, 4
1	0.790	0.810	0.874	0.970	0.488	0.503
2	0.778	0.835	0.958	0.679	0.741	0.914
3	0.615	0.386	0.313	0.786	0.476	0.714
4	0.970	0.265	0.518	0.287	0.455	0.494
5	0.837	0.659	0.602	0.383	0.355	0.994
6	0.762	0.695	0.826	0.898	0.743	0.898
7	0.470	−0.376	0.274	−0.503	0.313	−0.570
8	0.643	0.494	0.238	0.241	−0.262	0 675
9	0.833	0.563	0.635	0.599	0.659	0.952
Average	0.744	0.481	0.582	0 482	0.441	0.619

Table 8
Classification of Digits by Phonemic Structure in
Terms of Consonant and Vowel Position

VC	CVC (C)	(C) CV	CVCVC
8-/eɪt/	5-/faɪv/	2-/tu/	(7-/sevθn/)
	9-/naɪn/	3-/θri/	
	1-/wʌn/	4-/fɔ/	
	6-/sɪks/		
	(7-/sevn/)		

(P3) Phonemic structure, in terms of consonant and vowel positioning, is important. These structures are shown in Table 8. Phonemic realizations of the digits differ from speaker to speaker. "Four" almost always contains a diphthong, and "seven" sometimes has a vocalic component between /v/ and /n/. In Fig. 6, we provide a transcription of the vowels in the digits of one of the speakers in which these points are illustrated.[2]

On the basis of phonemic structure (P3), we have four groups of digits. In recognition, P1 tends to operate, merging 8 and the 1 5 9 6 group. The former would still remain distinct from 1 5 9 on the basis of vowel quality. In addition, 6 and 8 contain a stop consonant and end with unvoiced sounds, whereas 1, 5, and 9 all end with voiced consonants. In Group C, 4 is distinguished from 2 and 3 by the first formant of the vowel.

In recall, we suppose the initial sound plays a greater role. One consequence of this is that the 4 5 confusion increases relative to other confusions. In addition, 4, 1, 5, and 9 are the only digits with low vowels. In recall, P2 should be weakened. However, although the 2 3 confusions are relatively fewer (as can be seen in the dendrogram in Fig. 5b compared with that of 3b), they are still strongly linked and separated from the other digits. Thus, the vowel height must still be playing a large role in the memory code.

This brief account is clearly incomplete but serves to indicate the order of complexity necessary to explain the patterns in the data. It is doubtful whether the problems can be resolved in detail by taking a wider range of words than the digit set, as it is likely that Ss
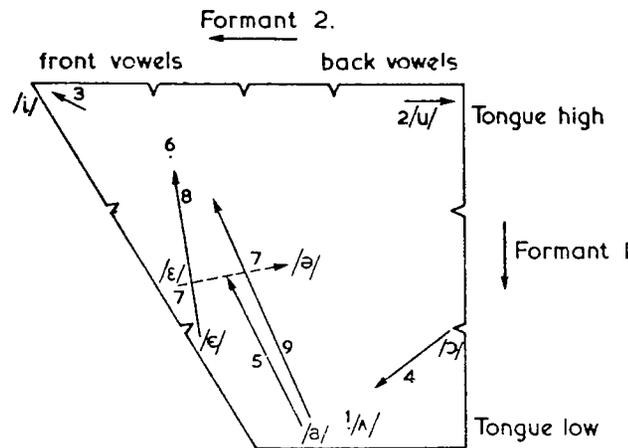

Fig. 6. A transcription of the vowels in the digits.

adopt different analytic strategies in recognition and different coding strategies in recall as a function of the vocabulary. Thus, prediction from the full alphabetic confusion matrix to small vocabularies using the CRR could overestimate the errors for low-confusion subsets such as C F J. The reasons for this would be that, with such subsets, the vowel alone gives all the information necessary. When these letters occur with a larger vocabulary, the consonants must be coded as well, presumably reducing the efficiency of the vowel coding. Thus, the /s/-/f/ relation, for example, could give rise to C-F confusions. There is insufficient data in the literature to permit a present test of such hypotheses.

Our primary conclusions are that the digits are not equally memorable or recognizable and that confusions are systematic. The pattern of confusions differs between the two tasks, especially with respect to the linking of 4 (although this does not bear on the discussion as to whether items to be recalled are represented in an acoustic or an articulatory code—see Sperling & Spellman, 1970, p. 183; Morton, 1970, pp. 239-240). The practical importance of these findings is that the use of different random digit strings in separate conditions in memory experiments is liable to lead to systematic experimental error. Thus, Morton and Chambers (unpublished data) have found that in experiments involving the stimulus suffix (see Morton, Crowder, & Prussin, 1971), some digits (e.g., "nine") are considerably more likely than others (e.g., "six") to give rise to errors on the final item (the crucial position for a "suffix effect"). Balancing the digits for occurrence at all serial positions in each condition is an elementary and essential precaution.

REFERENCES

Altham, P. M. E. The measurement of association of rows and columns for an r x s contingency table. Journal of the Royal Statistical Society B, 1970, 32, 63-73.

Cane, V. R. Review of: "Individual choice behaviour: A theoretical analysis," by R. D. Luce (1959). Journal of the Royal Statistical Society A, 1960, 122, 486-488.

Carroll, J. D., & Chang, J. J. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckert-Young" decomposition. Psychometrika, 1970, 35, 283-319.

Clarke, F. R. Constant-ratio rule for confusion matrices in speech communication. Journal of the Acoustical Society of America, 1957, 29, 715-720.

Conrad, R. Accuracy of recall using keyset and telephone dial and the effect of a prefix digit. Journal of Applied Psychology, 1958, 42, 285-288.

Conrad, R. Errors of immediate memory. British Journal of Psychology, 1959, 50, 350-359.

Conrad, R. An association between memory errors and errors due to acoustic masking of speech. Nature, 1962, 193, 1314-1315.

Conrad, R. Acoustic confusions in immediate memory. British Journal of Psychology, 1964, 55, 75-84.

Cormack, R. M. A review of classification. Journal of the Royal Statistical Society A, 1971, 134, 321-367.

Edwards, A. W. F. The measure of association in a 2 x 2 table. Journal of the Royal Statistical Society A, 1963, 126, 109-114.

Goodman, L. A. The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries. Journal of the American Statistical Association, 1968, 63, 1091-1131.

Gower, J. C. Discussion on: "A review of classification" by R. M. Cormack. Journal of the Royal Statistical Society A, 134, 321-67.

Grey, D. R., & Morgan, B. J. T. Some aspects of ROC curve-fitting: Normal and logistic models. Journal of Mathematical Psychology, 1972. 9, 128-139.

Holloway, J. D., & Jardine, N. Two approaches to zoogeography: A study based on the distribution of butterflies, birds and bats in the Indo-Australian area. Proceedings on the Linnaean Society of London, 1968, 179, 153-188.

Jardine, N., & Sibson, R. The construction of hierarchic and non-hierarchic classifications. Computer Journal, 1968, 11, 177-183.

Jardine, N., & Sibson, R. Mathematical taxonomy. London: Wiley, 1971.

Johnson, S. C. Hierarchical clustering schemes. Psychometrika, 1967, 32, 241-254.

Kendall, D. G. Some problems and methods in statistical archaeology. World Archaeology, 1969, 1, 68-76.

Kendall, D. G. A mathematical approach to seriation. Philosophical Transactions of the Royal Society of London (A), 1970, 269, 125-134.

Kendall, D. G. Construction of maps from "Odd bits of information." Nature, 1971, 231, 158-159.

Klahr, D. A Monte-Carlo investigation of the statistical significance of Kruskal's nonmetric scaling procedure. Psychometrika, 1969, 34, 319-330.

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

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

Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. Psychometrika, 1964b, 29, 115-129.

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

Luce, R. D. A threshold theory for simple detection experiments. Psychological Review, 1963, 70, 61-79.

Lyerly, S. B. The average Spearman rank correlation coefficient. Psychometrika, 1952, 17, 421-428.

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

Morgan, B. J. T. Cluster analyses of two acoustic confusion matrices. Perception & Psychophysics, 1973a, 13, 13-24.

Morgan, B. J. T. A statistical test of Luce's choice axiom. Journal of Mathematical Psychology, 1973b. in press.

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

Morton, J., Crowder, R. G., & Prussin, H. A. Experiments with the stimulus suffix effect. Journal of Experimental Psychology, 1971, 91, 169-190.

Moser, H. M., & Dreher, J. J. Letter-digit interaction. Technical Report No. 8, RF Project 519, Ohio State University Research Foundation, 1954.

Shepard, R. N. Analysis of proximities: Multidimensional scaling with an unknown distance function. II. Psychometrika, 1962, 27, 219-246.

Shepard, R. N. Analysis of proximities as a technique for the study of information processing in man. Human Factors, 1963, 5, 33-48.

Shepard, R. N., Romney, A. K., & Nerlove, S. (Eds.) Multidimensional scaling. Vols. I and II. London: Seminar Press, 1972.

Sibson, R. Order invariant methods for data analysis. Journal of the Royal Statistical Society B, 1972, 34, 311-349.

Sperling, G., & Spellman, R. G. Acoustic similarity and auditory short-term memory experiments and a model. In D. A. Norman (Ed.), Models of human memory. New York: Academic Press, 1970.

Stenson, H. H., & Knoll, R. L. Goodness of fit for random rankings in Kruskal's nonmetric scaling procedure. Psychological Bulletin, 1969, 71, 122-126.

Townsend, J. T. Theoretical analysis of an alphabetic confusion matrix. Perception & Psychophysics, 1971, 9, 40-50.

Van der Kamp, L. J. Th., & Pols, L. C. W. Perceptual analysis from confusions between vowels. Acta Psychologica, 1971, 35, 64-77.

Van Rijsbergen, C. J. A clustering algorithm. Computer Journal, 1970, 13, 113-115.

Wagenaar, W. A., & Padmos, P. Quantitative interpretation of stress in Kruskal's multidimensional scaling technique. British Journal of Mathematical & Statistical Psychology, 1971, 24, 101-110.

Wickelgren, W. A. Acoustic similarity and intrusion errors in short-term memory. Journal of Experimental Psychology, 1965, 70, 102-108.

Wickelgren, W. A. Distinctive features and errors in short-term memory for English consonants. Journal of the Acoustical Society of America, 1966, 39, 388-398.

NOTES

1. Morgan, B. J. T. Testing finite first-order Markov chains for 'quasi-independence.' Unpublished work, 1972.

2. We are grateful to John Trim for providing this transcription.

3. G. Jenkinson of Kent University has been comparing randomly associated dendrograms, and his results imply that all the Δ_i^* values of Table 6 are significant at the 5% level.