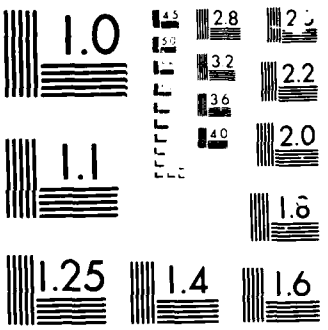
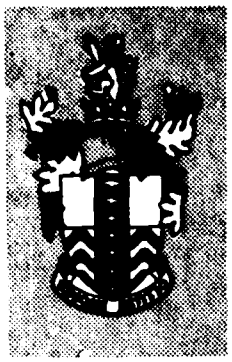
AD-A169 546 RANK-ORDERING OF SUBJECTS INVOLVED IN THE EVALUATION OF 1/1
AUTOMATIC SPEECH..(U) ROYAL SIGNALS AND RADAR
ESTABLISHMENT MALVERN (ENGLAND) D C SMITH ET AL.
UNCLASSIFIED JAN 86 RSRE-MEMO-3926 DRIC-BR-99243 F/G 17/2 NL


1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
2.2
2.0
1.8

UNLIMITED



RSRE MEMORANDUM No. 3926

ROYAL SIGNALS & RADAR ESTABLISHMENT

RANK-ORDERING OF SUBJECTS INVOLVED IN THE EVALUATION OF AUTOMATIC SPEECH RECOGNISERS

Authors: D C Smith, M J Russell
M J Tomlinson

PROCUREMENT EXECUTIVE,
MINISTRY OF DEFENCE,
RSRE MALVERN,
WORCS.

RSRE MEMORANDUM No. 3926

Memorandum 3926

TITLE: RANK-ORDERING OF SUBJECTS INVOLVED IN THE EVALUATION OF AUTOMATIC SPEECH RECOGNISERS

AUTHORS: D C Smith, M J Russell and M J Tomlinson

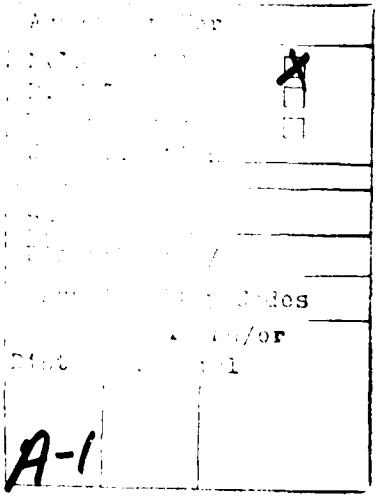
DATE: January 1986

SUMMARY

It is well known that the performance of current automatic speech recognisers varies significantly between talkers. Hence it is essential for anyone involved in the evaluation of speech recognition systems, or the assessment of the acceptability of such systems in a particular application, to be able to 'calibrate' potential talkers.

The purpose of this memorandum is to provide measurements of the expected performance of a typical speech recogniser on a group of forty subjects who have provided recordings for speech recogniser evaluation. The group includes all but one of the speakers from the RSRE speech database, all of the speakers who contributed towards the UK part of the NATO RSG10 spoken digit database, and a group of pilots involved in the assessment of automatic speech recognition in avionics applications at the Royal Aircraft Establishment at Bedford and Farnborough.

A method of ranking the subjects based on simpler measurements was also considered. This method potentially requires much less computation and was found to correlate well (rank correlation 0.687) with the ranking by expected error-rate.



INTRODUCTION

It is well known that the performance of current automatic speech recognition equipment varies significantly between speakers [1], [2]. Some speakers, who have come to be refered to as 'sheep' [2], achieve acceptable performance on relatively unsophisticated systems, while others, the 'goats' [2], are able to confound even more advanced machines. It is therefore essential for anyone involved in speech recognition research to be able to calibrate potential speakers. This is particularly important in recogniser or algorithm evaluation, speech database compilation, or in the assessment of the acceptability of Direct Voice Input (DVI) in particular applications.

In circumstances where it is only possible to gather data from a limited number of speakers the need for speaker calibration is heightened. For example in avionics applications it may be considered essential to gather data from a pilot during flight. If that pilot is an extreme "sheep" or "goat" then there is a danger that the assessment of the acceptability of DVI in that application will be biased. Choice of speaker is equally important when gathering data to test advanced recognition algorithms. Much time can be wasted collecting and processing data from a "sheep", since this is unlikely to produce the recognition errors which are needed to reveal weaknesses in the algorithm.

The primary purpose of this memorandum is to calibrate a group of speakers currently involved in speech recogniser evaluation. The group includes all but one of the speakers recorded in the RSRE Speech Database [3], [4], all of the speakers recorded in the UK part of the NATO RSG10 Spoken Digit Database [8] and two groups of pilots, based at the Royal Aircraft Establishment at Bedford and Farnborough, who are participating in the assessment of DVI in avionics applications.

The subjects are calibrated by measuring the performance of a speaker-dependent isolated digit recogniser on their speech. The recognition algorithm used is a 'textbook' whole-word pattern matching Dynamic Time-Warping (DTW) algorithm [5] [6]. This calibration criterion is clearly relevant, since such algorithms are the basis of the majority of current commercial recognisers. However the method is also of interest in a much wider sense. Classification in a whole word template matching speech recogniser is based solely on measurments of the dissimilarity between a pattern representing an unknown word and each of a set of known reference words. Such a system will only perform well if the distances between patterns representing the same word are small. Hence, provided acount can be taken of other variables such as vocabulary and environment, the expected performance of such a recogniser for a particular speaker can be regarded as a measure of the consistency of pronunciation of that speaker.

The method used to compute expected recogniser performance is taken from [7]. It was shown in [7] that for a typical whole-word template matching recogniser, recognition accuracy depends crucially on the choice of reference words. This effect is accomodated by computing the expected error rate, over a large number of randomly chosen sets of reference words, for each speaker in the study. The algorithm used is taken from [7] but is described briefly in the text.

A secondary goal of the study is to investigate alternative methods for speaker calibration. This arises from two considerations. First, calibration by expected error-rate is expensive in terms of the quantity of speech data needed from each subject and in terms of computing requirement. Furthermore, because of the relative ease of speaker dependent isolated

digit recognition the method is ineffective for discriminating between consistent speakers, who will all score a percentage error-rate of approximately zero.

Hence, two alternative methods for speaker calibration are considered. These are based on measurements of the distance between corresponding spectra in different utterances of a given word. The correspondence between spectra is determined using a standard dynamic time-warping algorithm [5]. Potentially, calibration using measurements of spectral distance requires less speech data than the expected error-rate method described above, and hence is computationally less expensive. Measurements of correlation between the rankings of the speakers obtained using the different approaches are presented.

METHOD

Subjects

A total of forty normal speakers, including eight female speakers, were recorded. The ages of the speakers ranged from nineteen to fifty-three years. Seven of the speakers are pilots from the Royal Aircraft Establishment at Bedford and Farnborough involved in assessing the acceptability of DVI in avionics applications, two are involved in Air Traffic Control and the remainder are members of staff at RSRE. The latter include all but one of the speakers recorded in the RSRE Speech Database [1], [2] and all of the speakers who contributed to the UK part of the NATO RSG10 Spoken Digit database [8].

All of the speakers were aware of the purpose of the recording session, and each speaker completed a brief questionaire giving details of their linguistic background.

Vocabulary

The experiments performed call for many repetitions of a relatively small set of words. The digits zero to nine were chosen as a suitable vocabulary. All of the subjects spoke four lists of one hundred isolated digits in English. Each list contains ten examples of every digit, ordered according to NATO RSG10 lists SB, 1A, 1B and 1C [8]. The first list, list SB, was used as training data and the remaining three as test data.

Recording Procedure

The procedure and equipment used to record the subjects was identical to that used to record the original RSRE speech database, and the reader is referred to [1] and [2] for details.

Speakers were recorded in a 'Burgess' acoustic booth using a Shure SM10A Professional Head-Worn Microphone adjusted to be three-quarters of an inch from the corner of the speakers mouth. Rate of speaking and 'list effects' were controlled to some extent by presenting the vocabulary one word at a time at 1.8 second intervals on a visual display unit controlled remotely by a BBC microcomputer. The microcomputer also monitored the recorded speech level, which was displayed to the speaker as a horizontal line under the presented word such that the length of the line was proportional to the speech level. The system was calibrated so that a mid-scale reading on this 'level meter' corresponded to optimum loading of the recording system, and speakers were asked to maintain this level.

All of the subjects were initially asked to say a small number of digits, at

a comfortable level, to enable the operator to set the microphone gain and to calibrate the system. These were not recorded.

The speech was recorded digitally on video cassette using a Sony PCM-F1 pulse code modulation system (set to 14 bit resolution) and a Sony SL-F1UB video cassette recorder (PAL).

Preprocessing

The speech was digitised using a 20ms frame rate, 19 channel vocoder. Hence, after processing, an utterance is represented as a sequence of vectors $u = u_1,\ldots,u_I$, where u_i is the 19-dimensional output of the vocoder at time i. The sequence u will be refered to as a word-pattern. The resulting file was segmented and edited using a semi-automated process which locates and labels each word in the file. The same process also removes any extraneous noises which occur between the digits and unwanted announcements from the beginnings and ends of the recording sessions.

Synthetic Control Data

In order to provide control data for the experiments, for example to identify any variations which are introduced during the preprocessing stage, an additional complete set of recordings was prepared using a commercial speech synthesiser. The system used was the standard internally-fitted Acorn speech synthesiser for the BBC microcomputer [9]. This is an LPC synthesiser based on the Texas Instruments TMS 5220 voice synthesis processor and TMS 6100 voice synthesis memory. The synthetic data is identified as speaker "KK" in the results.

Recognition Algorithm

The recognition algorithm used is a 'textbook' whole-word pattern matching, Dynamic Time-Warping (DTW) algorithm [5], [6]. The particular DTW algorithm used corresponds to case p=1 in [5]. A brief description is included for completeness.

Dynamic Time-Warping is a particular method for computing a measure of dissimilarity between two word-patterns. Its key feature is the use of timescale distortions to obtain an optimal match during the comparison process. This compensates for the natural temporal variability in speech.

The basic principle of a DTW based whole-word pattern matching speech recogniser is straightforward. First the dissimilarity D(r,u) is computed between an unknown utterance u and each pattern r in a set R of stored reference patterns. The pattern u is then assigned to the class of the reference r* satisfying:

$$D(r^*,u) = \min_{r \in R} D(r,u).$$

Equivalently, classification is performed according to a 1-nearest neighbour rule with respect to the DTW dissimilarity measure D.

In order to understand how the DTW dissimilaity D(r,u) is computed it is necessary to introduce the concept of a time-registration path. Let $r = r_1,\ldots,r_I$ and $u = u_1,\ldots,u_J$ be two word-patterns. A time-registration path p of length L (where $\max(I,J) < L < I+J-1$) between r and u is a mapping from the ordered set $\{1,\ldots,L\}$ into the (i,j)-plane satisfying:

(a) $p(1) = (1,1)$

(b) $p(L) = (I,J)$
and, (c) $p(1) = p(1\text{-}1) + (a,b)$, where $(a,b) = (1,0)$, $(1,1)$ or $(0,1)$.

Any such path defines an alignment between r and u. Points on the path indicate corresponding vectors in the two word-patterns.

Given a dissimilarity measure d, defined on the set of all possible vectors u_i and r_j, the accumulated distance $D_p(r,u)$ between r and u along the path p is defined by:

$$D_p(r,u) = \sum_{x=1}^{L} d(r_{i_x}, u_{j_x}),$$

where $p(x) = (i_x, j_x)$ is the xth point on the path. $D_p(r,u)$ is a measure of how well the time-registration path p aligns the two patterns: a small value of $D_p(r,u)$ corresponds to small values of $d(r_{i_x}, u_{j_x})$ along the length of p and indicates a good alignment, whereas a large value of $D_p(r,u)$ indicates a poor alignment. The time-registration path which gives the best alignment between the two patterns is the path p* which satisfies:

$$D_{p^*}(r,u) = \min D_p(r,u)$$

where the minimum is taken over all possible time-registration paths between r and u. The path p* is called an optimal time-registration path. D(r,u) is defined to be the accumulated distance $D_{p^*}(r,u)$ along the optimal path p* and is sometimes called the cumulative distance between r and u.

In the present experiments the dissimilarity measure d is the 19-dimensional euclidean metric.

Computation of Error-Rate Distributions and Expected Error Rate

Given a set U of word-patterns corresponding to utterances spoken by a given speaker, it was shown in [7] that the the number of classification errors E[D,R] which occur using the above system depends to a large extent on the choice of the reference set R. Since the purpose of the present experiments is to investigate the dependency of E[D,R] on the speaker, this effect is undesirable and must be removed. In [7] this is achieved by computing the distribution of E[D,R] over a large number of reference sets. From this distribution the expected error-rate E[D] is easily obtained. In the present experiments the reference sets were chosen randomly from the training list SB. To avoid unnecessary repeated calculations of the same dissimilarities, the following procedure from [7] was adopted:

First the dissimilarity D(r,u) is calculated between each pattern r in the training set (table SB) and each pattern u in the test set (tables 1A, 1B and 1C). Then, for each test pattern u, the training patterns are ordered according to increasing dissimilarity, so that the first training pattern in the ordering corresponding to u is the pattern r for which D(r,u) is smallest. Given a particular reference set R, classification of a test pattern u is achieved by simply scanning the ordering corresponding to u until a training pattern r belonging to R is found: u is then assigned to the class of r. In this way E[D,R] can be evaluated for several thousand reference sets in the time which would be required to process just ten or twenty reference sets directly.

Calibration using Mean Inter-Spectra Distances

Although it is efficient, the above procedure is still time consuming, both

in terms of the quantity of speech data required from each subject and in terms of computing resources. It was therefore decided that alternative methods for calibrating the subjects should be investigated, and the resulting ranking compared with that obtained using the expected error-rate criterion. Two alternative calibration criteria were considered. Both are based on the distances between pairs of vectors in different word-patterns which correspond according to an optimal time-registration path. Because these vectors represent short-term spectra, this distance is referred to as 'spectral distance'.

More precisely, let p* be an optimal time-registration path between r and u, then,

$$\overline{d(r,u)} = D(r,u)/L$$

is the mean value of the dissimilarity measure d along the path p* (here L is the length of p*). Hence $\overline{d(r,u)}$ is a measure of the average distance between corresponding vectors in the patterns r and u.

The Mean Inclass Spectral Distance (MID) (respectively Mean Outclass Spectral Distance (MOD)) is defined to be the average value of d(r,u) over all pairs of word-patterns r and u in the training set SB which belong to the same class (respectively different classes). Hence MID is a measure of the expected difference between corresponding regions in word-patterns representing examples of the same word, and can be regarded as a measure of variability, while MOD is a measure of the same distance between patterns which represent examples of different words and is therefore concerned with class separation.

Two criteria for ranking the subjects using MID and MOD were considered. Under the first criterion subjects were calibrated according to the MID parameter alone, while under the second criterion calibration was based on the ratio MID/MOD.

Computation of MID and MOD is easily accomplished during the computation of the expected error rate. In the present experiments the values of MID and MOD were those obtained by aligning all patterns in the training set with all utterances in the test set.

The advantage of this type of calibration is that potentially MID and MOD could be computed using much smaller amounts of data. In fact either parameter could be estimated from a single DTW alignment. However, experiments to determine the amount of data required to accurately estimate MID or MOD were not included in the present study.

EXPERIMENT RESULTS

Calibration by Expected Error-Rate

For each subject, the mean and variance of the error-rate were computed for a 300 utterance test set (lists 1A, 1B and 1C) using 5000 randomly selected reference sets from the 100 utterance training set (list SB). The results are shown in tables 1 and 2. Table 1 is ordered alphabetically according to the subjects surname, while table 2 is ordered by increasing mean error-rate.

The values of the expected error-rate vary between 0 and 11.25%, with a mean value of 1.29%. Between these extreme values the distribution of error rates is highly non-uniform: over half of the subjects have an expected

error rate of less than 0.1. The distribution of expected error rates, and the position of each subject in the distribution, is shown explicitely in figure 1. The clustering of values around 0 reflects the relative ease of speaker dependent isolated digit recognition, however it is clear from the extreme values of the distribution that for some subjects even this task causes the textbook recogniser severe problems. Notice that figure 1 can be used to estimate the ratio of 'sheep' to 'goats'.

The feature of the results which is apparent from table 2 is a correlation between the mean and variance of the error rate for each subject (see figure 2). This relationship, which was also observed in [7], can be explained as follows: The variance of the error-rate is a measure of the dependency of error rate on choice of reference set, therefore large variances can only occur if there are significant differences between reference sets. This in turn can only occur if there is variation between patterns in the training data representing the same classes. Any such variations will also be present in the test data and will result in classification errors.

A consequence of this result is that the dependency of error-rate on reference set selection is greater for 'goats' that for 'sheep'. Hence good reference selection is more important for 'goats'. Detailed information about this dependency is given in appendix 1, which includes histograms showing the distribution of error rate for each subject in the study. As in [7] the extent of the dependency of error-rate on the choice of reference set varies significantly between speakers.

Calibration by Spectral Distance

The values of MID and MID/MOD for all of the subjects in the study are shown in tables 3 and 4 respectively. The subjects are ordered according to MID in table 3 and MID/MOD in table 4.

Histograms showing the distribution of MID and MID/MOD are shown in figures 3 and 4. These should be compared with figure 1. As predicted, both MID and MID/MOD are better able to separate the consistent speakers than the expected error-rate measure.

Scatter plots of mean error-rate against MID and MID/MOD are shown in figures 5 and 6 respectively. The correlation between mean error-rate and MID is 0.31 and between mean error-rate and MID/MOD is 0.59. However, since the primary purpose of this study is to calibrate the speakers, a measure of the correlation between the rankings given by the different criteria is more relevant. The Spearman rank correlation coefficient between the rankings according to mean error-rate and MID is 0.39 (significant at the 1% level) and between the rankings for mean error-rate and MID/MOD is 0.687 (significant at the 0.1% level). Hence there is a significant agreement between the rankings according to the mean error-rate and MID/MOD criteria. Since the MID/MOD parameter contains information about both inclass and between-class distances, this result is not unexpected.

OBSERVATIONS

Noise

Despite efforts to minimise external noise during the recording sessions, it was possible to detect some extraneous noises on the recordings, the principle source being the speaker. The most common types were 'lip-smacking' between words and loud breathing. In some cases it was not possible to remove these unwanted noises during the editing stage because of

their close proximity to valid words.

However, despite repeated manual inspection of the speech files corresponding to 'goats', there did not prove to be any obvious correlation between subjectively 'noisy' speakers and high error-rate.

Intonation

For some speakers it was possible to detect audible changes in mood during the recording session, for example from subjectively sounding alert and enthusiastic to sounding uninterested and tired. Also, some speakers used more than one intonation pattern for the digits, for example both rises and falls.

In the experiments fundamental frequency information was not directly made available to the recognition algorithm. Moreover, because of the coarse resolution of the frequency analysis used, it is unlikely that changes in intonation pattern will have led to detectable differences in spectral shape. However, no experiments were conducted to explicitely measure the effect of varying intonation patterns on recogniser performance.

Level Variations

Although a real-time speech level meter was provided during the recording session, many speakers experienced some difficulty in maintaining a constant level and over-correction following a minor drop or rise in level was common.

It has been shown in previous experiments [9] that level variation is significant in causing recognition errors, and that performance can be improved by employing some form of amplitude normalisation. In the present experiments no such normalisation was used and it is therefore likely that some of the higher error-rates in the results may be attributable to variations in speech level. No experiments were performed to explicitely investigate this effect on the new data.

Vocoder-Introduced Variation

The value of MID obtained for the synthetic speech (speaker KK) indicates that significant variation is introduced during the filtering and digitizing process performed by the vocoder. This was confirmed by an experiment using human speech. An artificial recording of list SB was prepared by duplicating a single example of each digit (spoken by speaker RH) ten times. Since the copying was performed digitally with the PCM recorder, the resulting recording consisted of ten identical examples of each digit. The value of MID for the vocoded version of this data was 2.09, which represents a significant fraction of the average value of MID (3.88) over all of the subjects in the study.

The above results support the hypothesis that condiderable variability is introduced during the preprocessing stage and suggest that this variation accounts for a significant portion of the value of the MID parameter for all of the subjects.

These results are in agreement with unpublished results of experiments performed in this laboratory on commercial speech recognisers which use vocoder type preprocessing. In these experiments the recogniser was first trained on one example of each digit, and then repeatedly tested on a fixed prerecorded 100 digit test set. The experiment was performed under controlled conditions using digital recordings. It was observed that the

number of errors varied from one trial to the next. This was attributed to variability introduced during the preprocessing stage, largly due to the low temporal resolution of the vocoder analyser.

CONCLUSIONS

The purpose of these experiments was to calibrate a group of speakers who have produced speech data for use in speech technology research. This has been achieved. Since the calibration is based on the performance of a textbook speech recognition algorithm, the results should reflect the expected performance of comparable commercial speech recognisers. The results can also be interpreted in terms of speaker consistency.

Alternative methods for speaker calibration, based on measurements of the distance between corresponding spectra in different word-patterns, have been considered and compared with the mean error-rate criterion. It has been shown that a rank-ordering of the subjects based on the MID/MOD parameter correlates well with a ranking based on expected error-rate.

In addition, further evidence has been presented which highlights the inadequacy of the particular type of vocoder used in these experiments as a preprocessor for automatic speech recognition.

It is anticipated that the results presented will be useful both as an aid for interpreting the outcome of future experiments and for selecting data and speakers for testing new recognition algorithms.

REFERENCES

1. N R Dixon and H F Silverman, "What are the Significant Variables in Dynamic Programming for Discrete Utterance Recognition?", Proc IEEE Int. conf. Acoustics, Speech and Signal Processing, 1981, 728-731.

2. G R Doddington and T B Schalk, "Speech Recognition: turning theory to practice", IEEE Spectrum, September 1981, 26-32.

3. J C A Deacon, R L Pratt, R K Moore, M J Russell and M J Tomlinson, "RSRE Speech Database Recordings (1983), Part I: Specification of Vocabulary and Recording Procedure", RSRE Report No. 83010, December 1983.

4. M J Russell, R K Moore, M J Tomlinson and J C A Deacon, "RSRE Speech Database Recordings (1983), Part II: Recordings made for Automatic Speech Recognition Assesment and Research", RSRE Report No. 84008, May 1984.

5. H Sakoe and S Chiba, "Dynamic Programming Algorithm Optimisation for Spoken Word Recognition", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol ASSP-26, Feb 1978, 43-49.

6. R K Moore, "Systems for Isolated and Connected Word Recognition", NATO ASI Series, Vol. F16, "New Systems and Architectures for Automatic Speech Recognition and Synthesis", Ed. R De Mori and C Y Suen, Springer-Verlag, Berlin Heidelberg 1895, 73-143.

7. M J Russell, J C A Deacon and R K Moore, "Some Implications of the Effect of Template Choice on the Performance of an Automatic Speech Recogniser", Proc Institute of Acoustics Autumn Conf., Vol 6, Part 4, 1984, 287-292.

8. R S Vonusa, J T Nelson, S E Smith and J G Parker, "NATO AC/243 (Panel III RSG10) language data base", Proc. NBS Speech I/O Stand. Workshop, 1982.

9. "The BBC Microcomputer Speech System User Guide", Acorn Computers Ltd., 1983.

10. M J Tomlinson, Results of unpublished experiments on the effect of variations in speech level on the performance of a 'text book' speech recogniser.

SPEAKER	ERROR RATE	VARIANCE
CB	0.09	0.09
GB	0.01	0.01
JB	2.76	9.01
SB	1.07	1.83
TB	0.18	0.33
BC	2.20	2.97
NC	0.24	0.07
RC	0.84	2.29
PG	6.20	16.74
SG	0.09	0.24
KH	0.02	0.02
RH	0.04	0.03
AJ	0.13	0.17
DJ	3.82	9.57
KK	0.00	0.00
MK	0.18	0.24
SL	0.00	0.00
AM	0.98	2.93
BM	0.77	1.94
NM	0.06	0.07
RM	0.01	0.01
HN	0.32	0.52
JP	1.22	2.20
KP	1.54	5.72
SP	0.77	0.94
DR	0.89	1.27
GR	0.00	0.00
JR	11.25	27.44
MR	1.94	5.46
TR	0.43	0.86
DS	3.93	11.76
ES	2.94	6.47
SS	1.03	2.98
WS	1.92	3.28
MT	0.44	1.83
AW	2.30	3.75
FW	0.37	0.56
IW	0.14	0.17
MW	1.57	1.60
RW	0.01	0.01
SW	0.05	0.08
MEAN	1.29	3.06

TABLE 1: Mean and variance of error-rate for each subject in the study, ordered by subjects surname. The values were computed for a 300 digit test set (tables 1A, 1B and 1C) over 5000 randomly chosen reference sets from the 100 digit training set (table SB).

SPEAKER	ERROR RATE	VARIANCE
KK	0.00	0.00
GR	0.00	0.00
SL	0.00	0.00
GB	0.01	0.01
RM	0.01	0.01
RW	0.01	0.01
KH	0.02	0.02
RH	0.04	0.03
SW	0.05	0.08
NM	0.06	0.07
CB	0.09	0.09
SG	0.09	0.24
AJ	0.13	0.17
IW	0.14	0.17
MK	0.18	0.24
TB	0.18	0.33
NC	0.24	0.07
HN	0.32	0.52
FW	0.37	0.56
TR	0.43	0.86
MT	0.44	1.83
BM	0.77	1.94
SP	0.77	0.94
RC	0.84	2.29
DR	0.89	1.27
AM	0.98	2.93
SS	1.03	2.98
SB	1.07	1.83
JP	1.22	2.20
KP	1.54	5.72
MW	1.57	1.60
WS	1.92	3.28
MR	1.94	5.46
BC	2.20	2.97
AW	2.30	3.75
JB	2.76	9.01
ES	2.94	6.47
DJ	3.82	9.57
DS	3.93	11.76
PG	6.20	16.74
JR	11.25	27.44
MEAN	1.29	3.06

TABLE 2: Results from TABLE 1 ordered by increasing mean error-rate.

SPEAKER	M.I.D.
KK	2.57
MR	3.54
SS	3.55
TR	3.55
RH	3.55
SL	3.56
RC	3.58
CB	3.59
DR	3.63
NM	3.65
GB	3.67
MT	3.69
NC	3.69
SP	3.74
RW	3.75
PG	3.75
SW	3.76
HN	3.76
RM	3.77
JP	3.78
DS	3.84
SG	3.86
GR	3.88
SB	3.92
AW	3.94
TB	3.97
KP	3.99
KH	4.01
MK	4.09
AJ	4.12
JB	4.12
DJ	4.13
FW	4.13
ES	4.16
AM	4.22
IW	4.23
JR	4.30
MW	4.31
BM	4.48
WS	4.51
BC	4.68
MEAN	3.88

TABLE 3: Value of MID for each subject in the study, ordered according to increasing MID. The values of MID were computed by aligning each digit in the training set (table SB) with the 30 digits from the same class in the test set (tables 1A, 1B and 1C)

SPEAKER	MID / MOD
KK	0.31
RM	0.42
KH	0.44
GR	0.44
HN	0.45
SW	0.45
MR	0.45
IW	0.46
SS	0.46
SP	0.46
TB	0.47
GB	0.47
RW	0.47
RC	0.47
SL	0.47
CB	0.48
MT	0.48
NC	0.48
RH	0.48
BM	0.48
PG	0.49
TR	0.49
NM	0.49
AJ	0.49
JP	0.50
KP	0.50
FW	0.50
JB	0.51
SG	0.51
MK	0.52
DJ	0.52
SB	0.52
AM	0.53
AW	0.54
DR	0.54
BC	0.54
WS	0.54
MW	0.54
DS	0.55
ES	0.55
JR	0.58
MEAN	0.49

TABLE 4: Value of MID/MOD for each subject in the study, ordered according to increasing MID/MOD. The values of MID / MOD were computed by aligning each utterance in the training set (table SB) with all utterances in the test set (tables 1A, 1B and 1C).

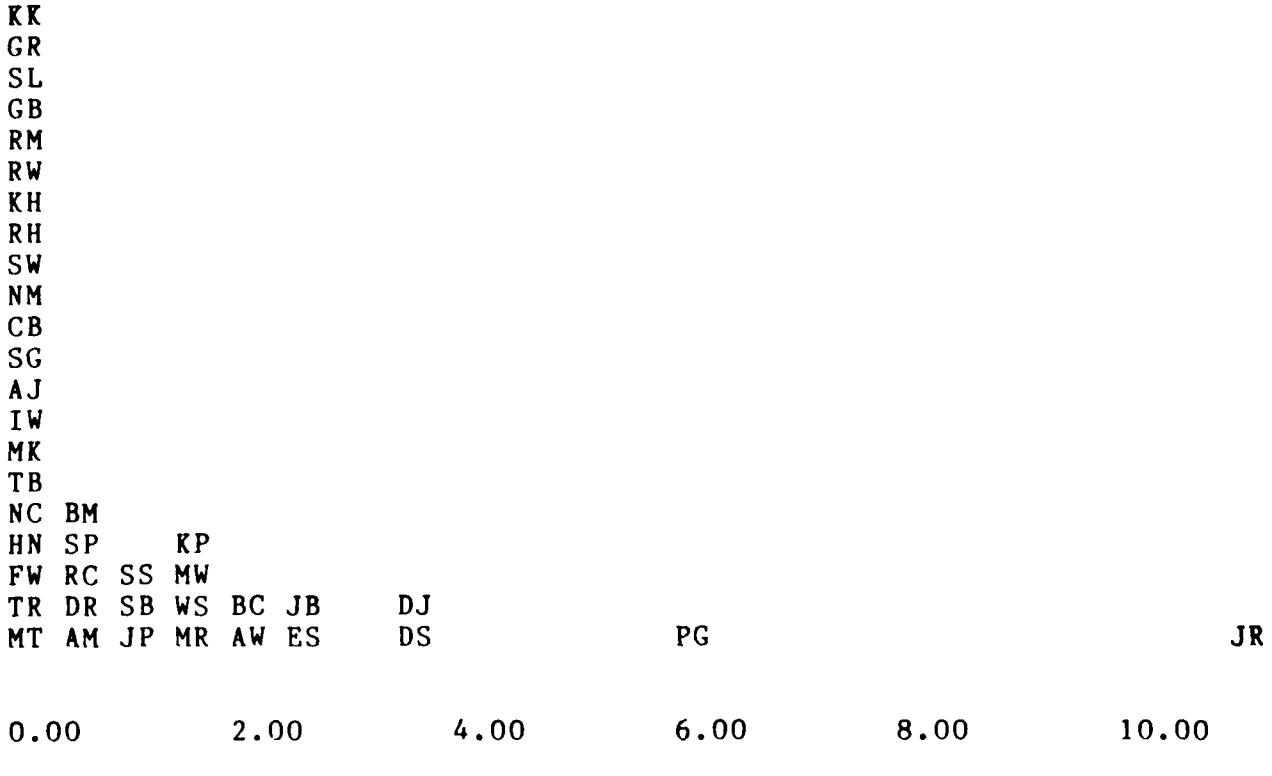


FIGURE 1: Histogram showing the distribution of mean-error rates over all subjects in the study.

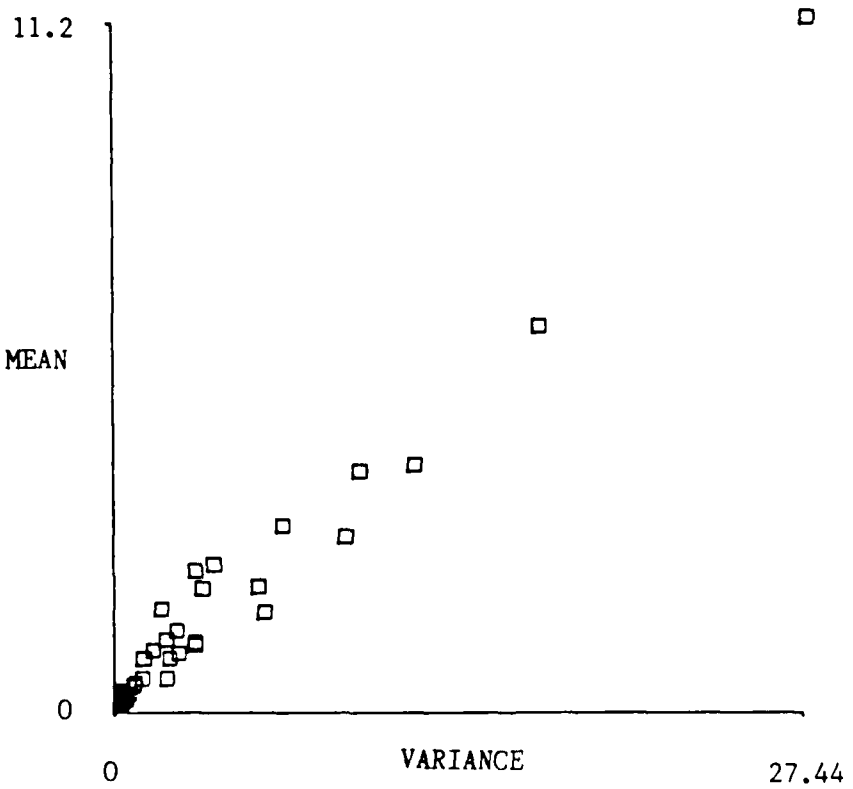


FIGURE 2: Scatter plot of mean error-rate against variance of error-rate over all subjects in the study.

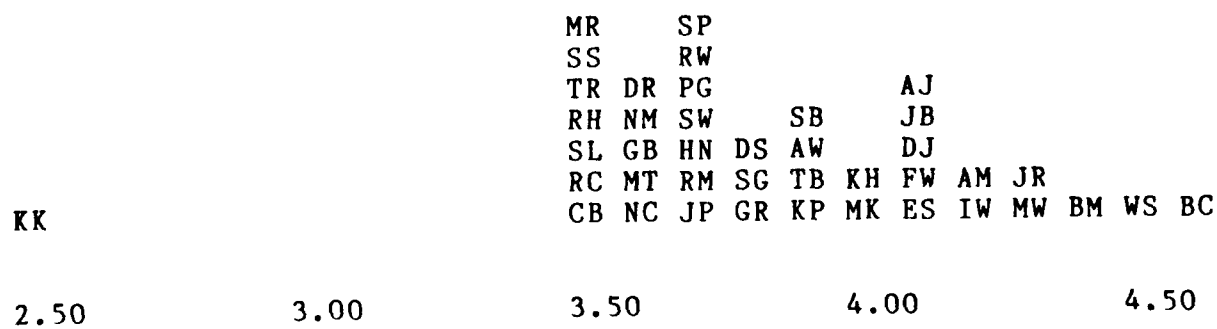


FIGURE 3: Histogram showing the distribution of values of MID over all subjects in the study.

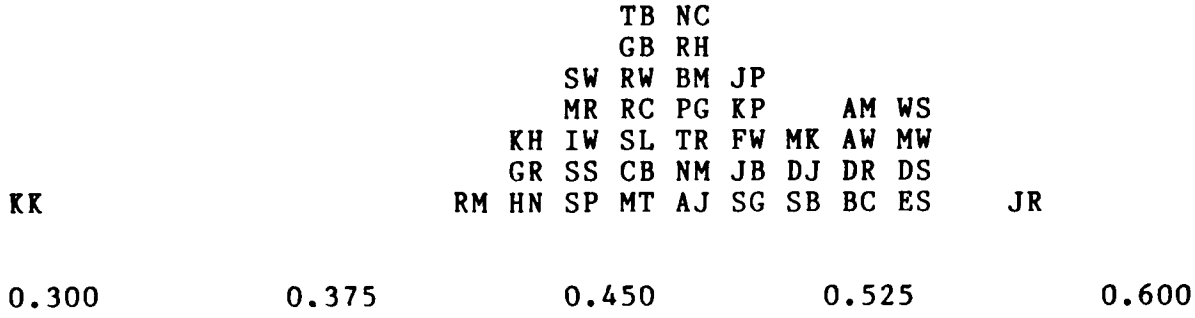


FIGURE 4: Histogram showing the distribution of values of MID/MOD over all subjects in the study

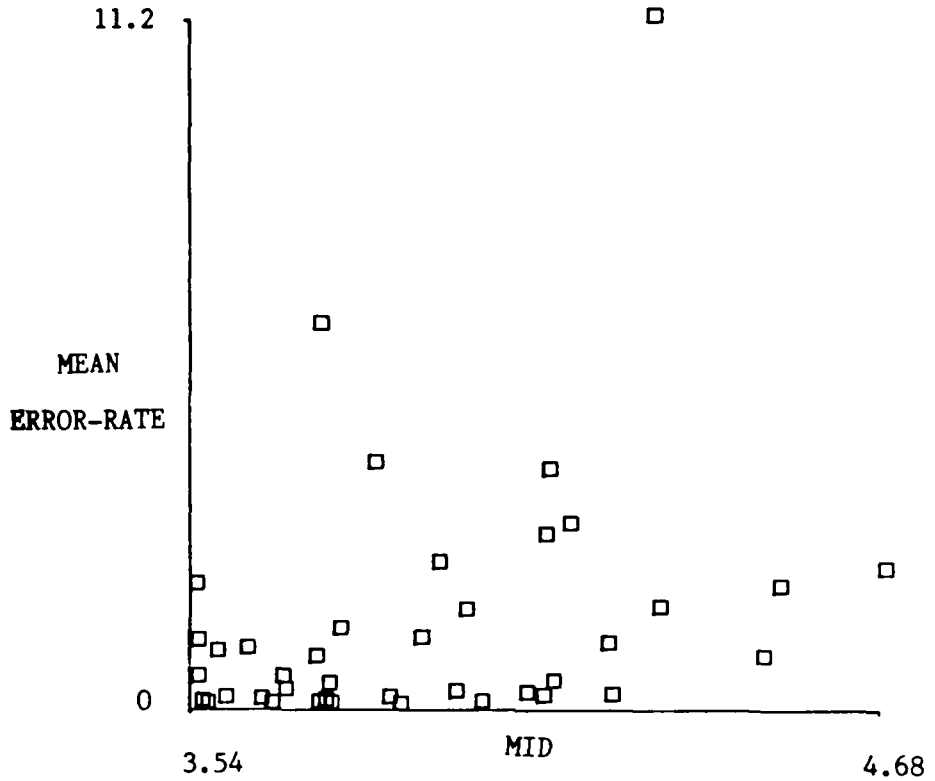


FIGURE 5: Scatter plot of mean error-rate against MID over all subjects in the study.

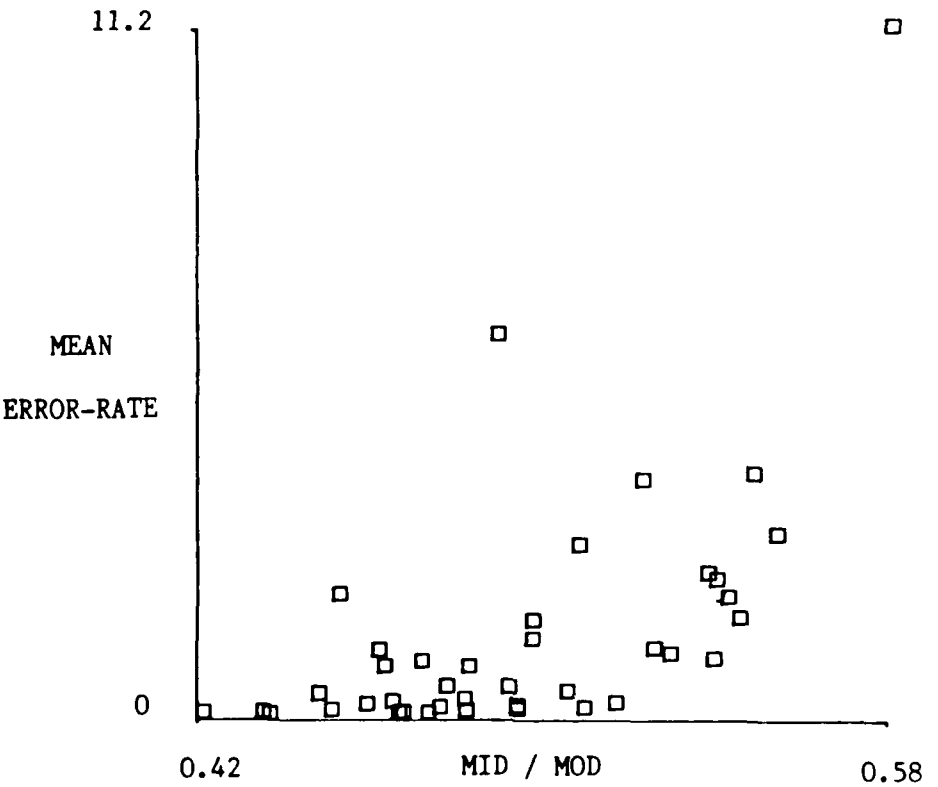
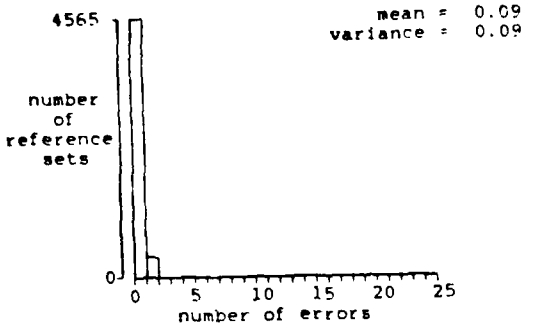


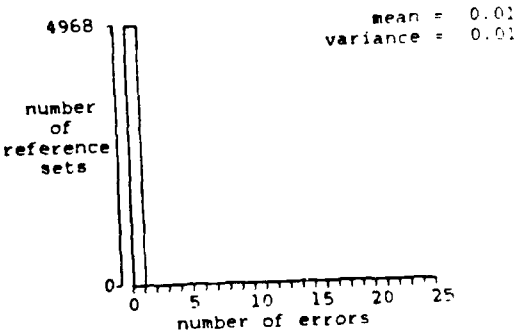
FIGURE 6: Scatter plot of mean error-rate against MID/MOD over all subjects in the study.

APPENDIX

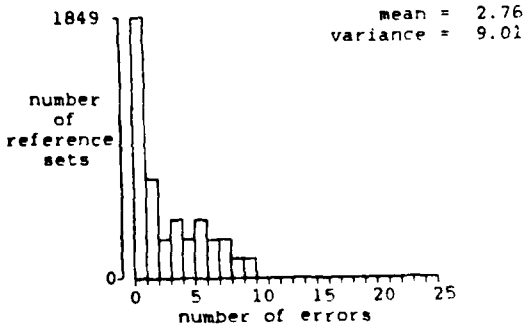
Histograms showing the distribution of error-rates for each subject in the study. The error-rates were computed for the 300 word test set (tables 1A, 1B and 1C) over 5000 randomly chosen reference sets from the 100 digit training set (table SB).



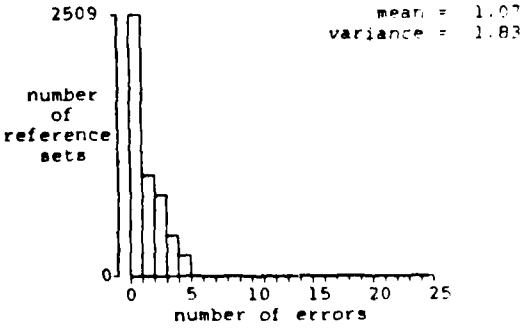
speaker CB



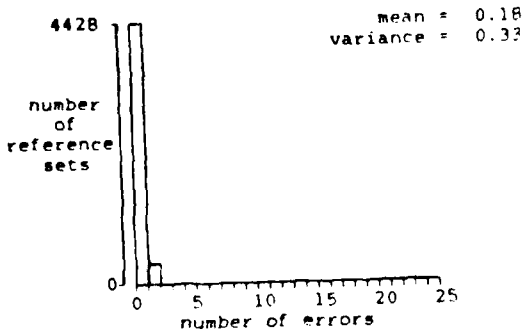
speaker GB



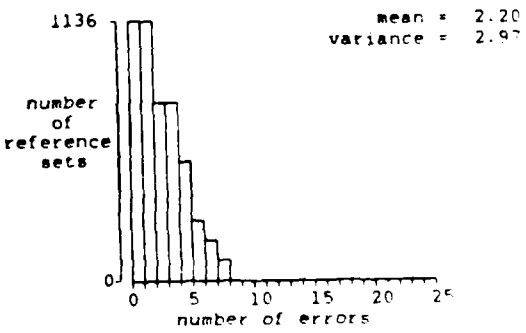
speaker JB



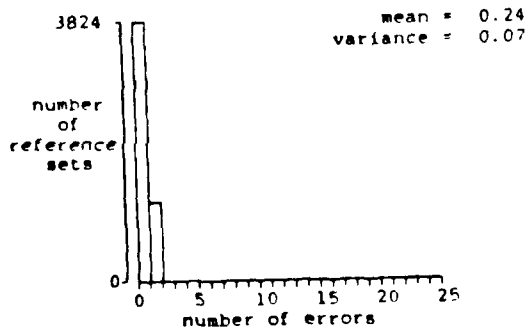
speaker SB



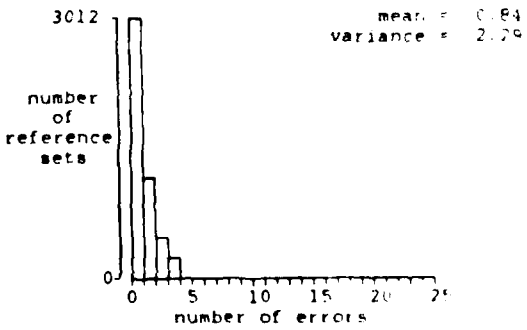
speaker TB



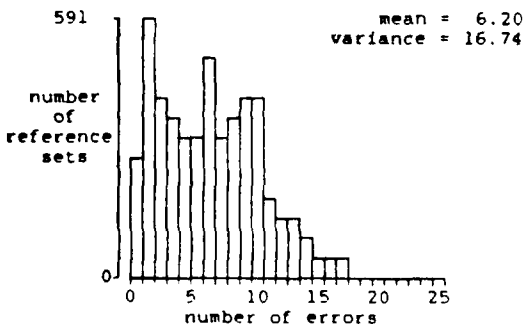
speaker BC



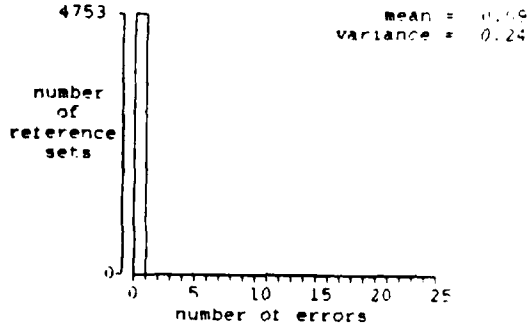
speaker NC



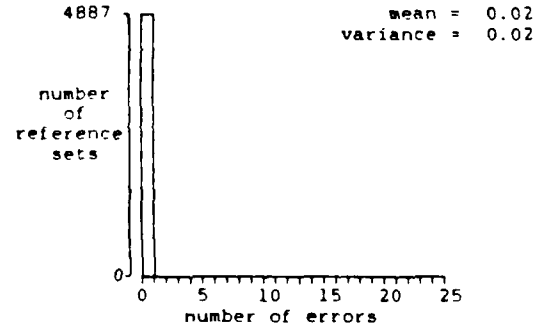
speaker RC



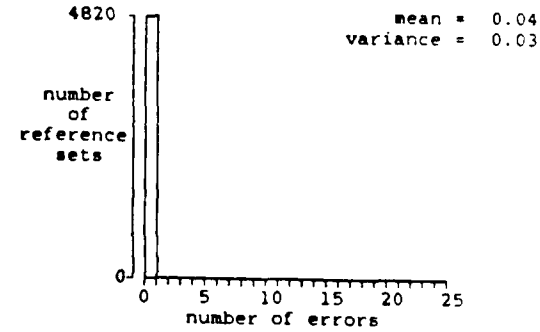
speaker PG



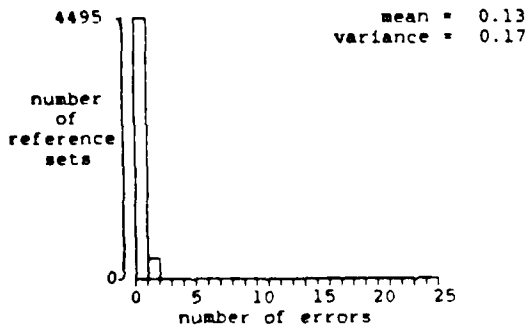
speaker SG



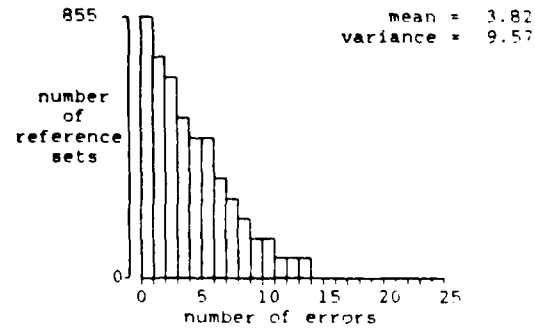
speaker KH



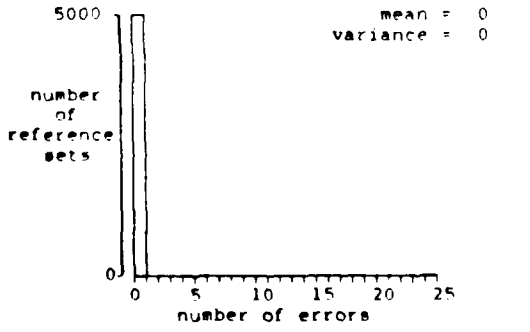
speaker RH



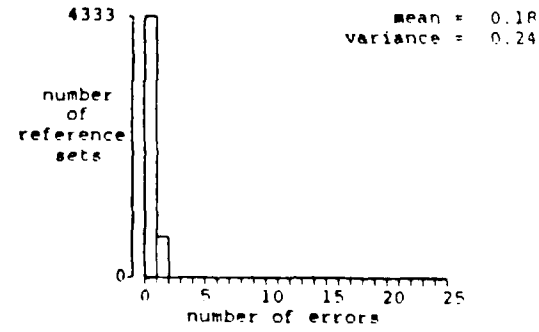
speaker AJ



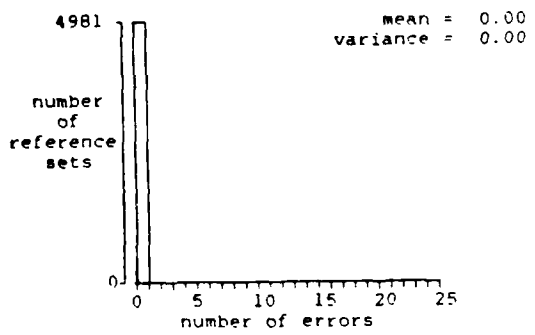
speaker DJ



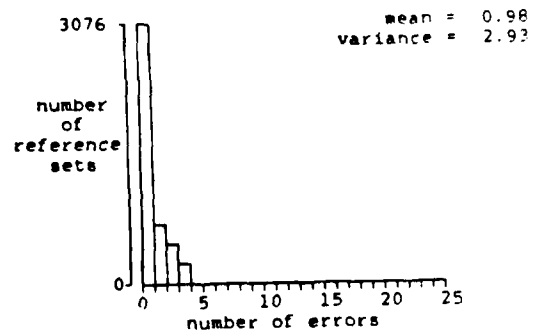
speaker KK



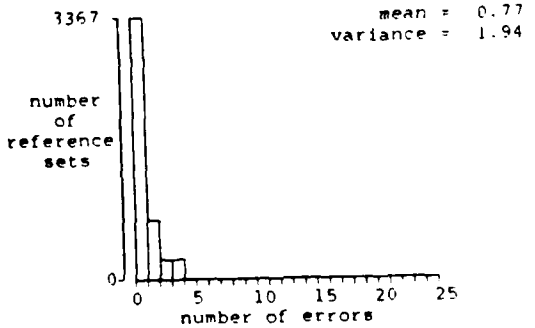
speaker MK



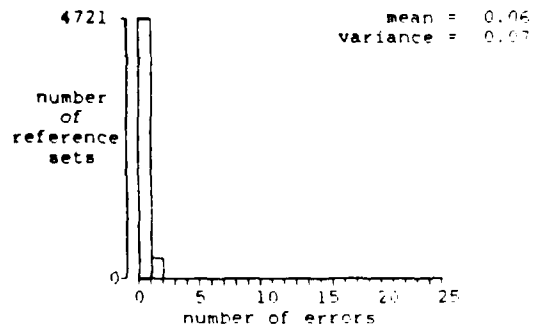
speaker SL



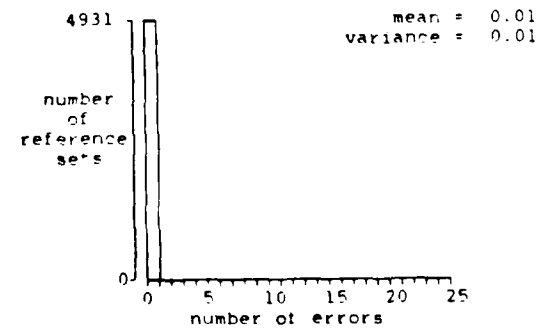
speaker AM



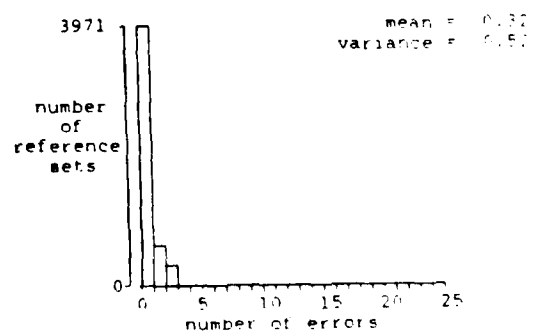
speaker BM



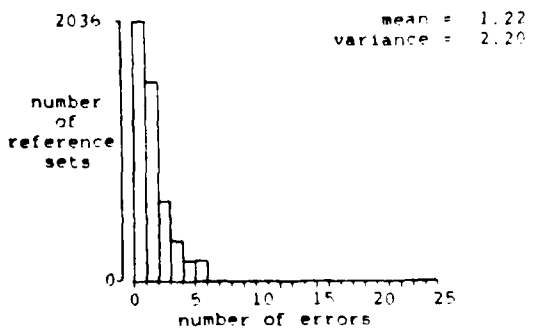
speaker NM



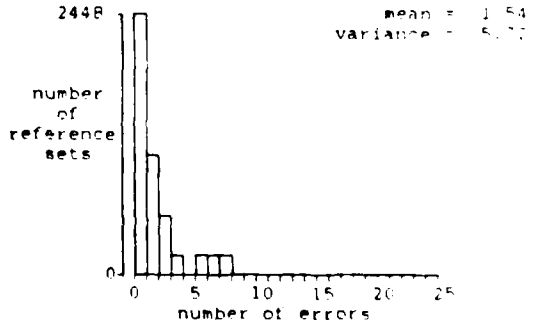
speaker RM



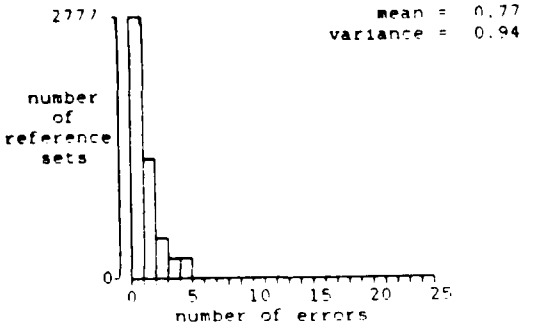
speaker HN



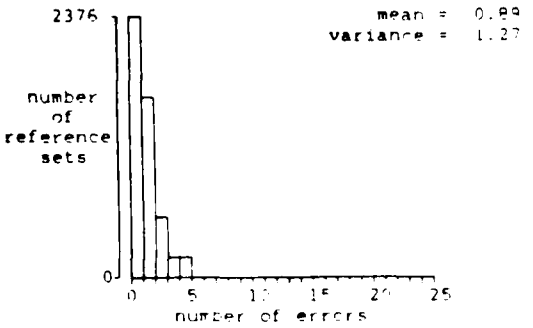
speaker JP



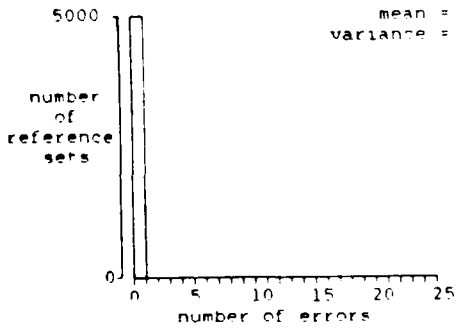
speaker KP



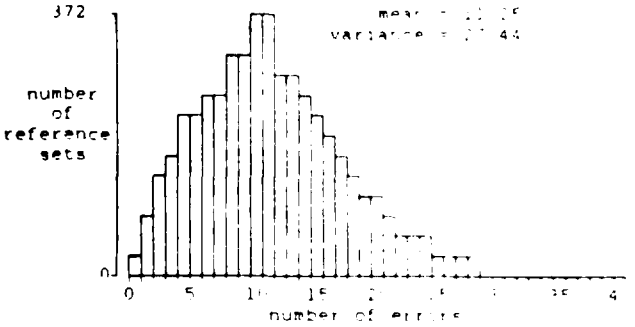
speaker SP



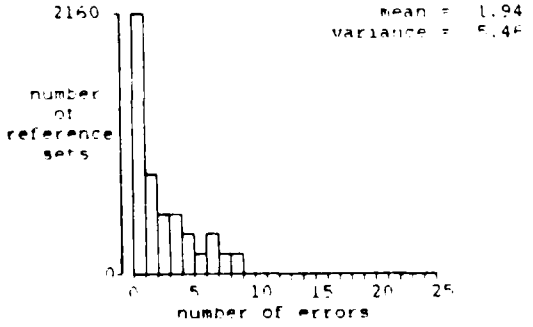
speaker DR



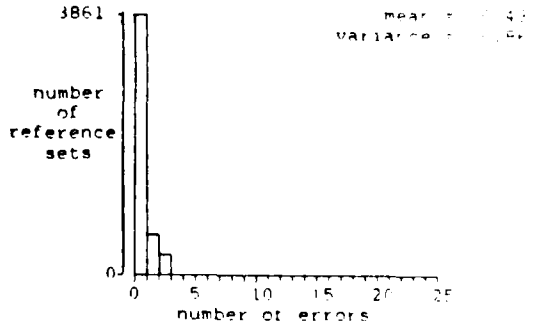
speaker GR



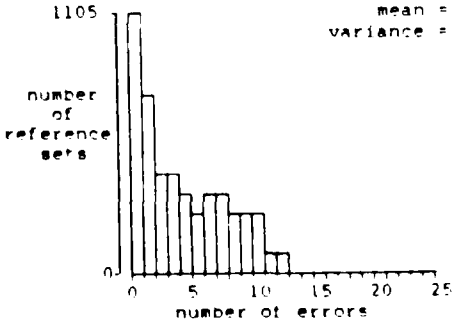
speaker JR



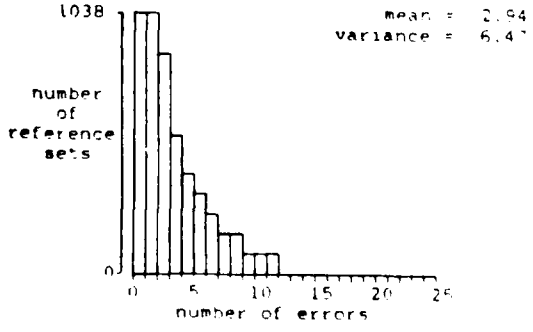
speaker MR



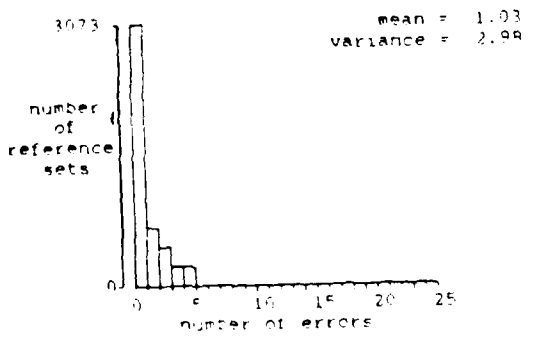
speaker TR



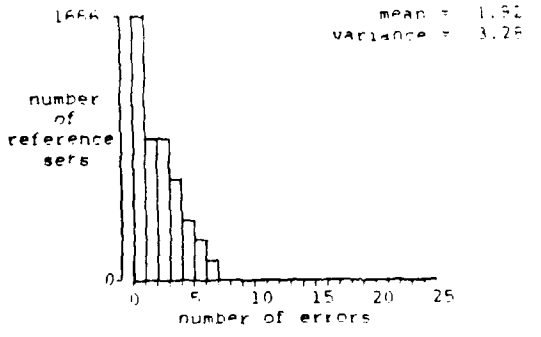
speaker DS



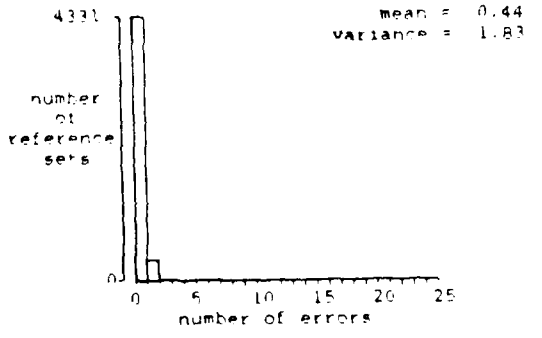
speaker ES



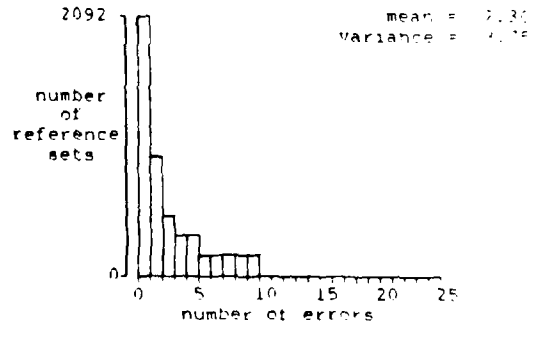
speaker SS



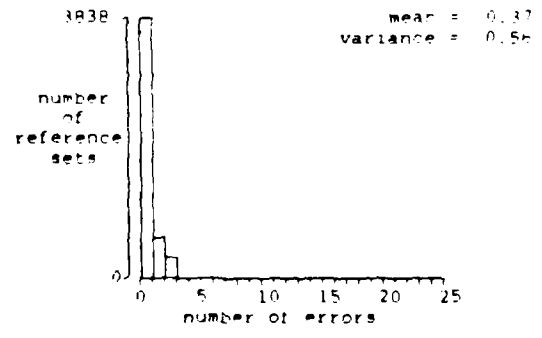
speaker WS



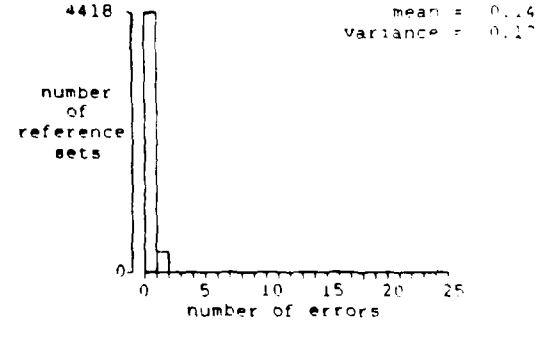
speaker MT



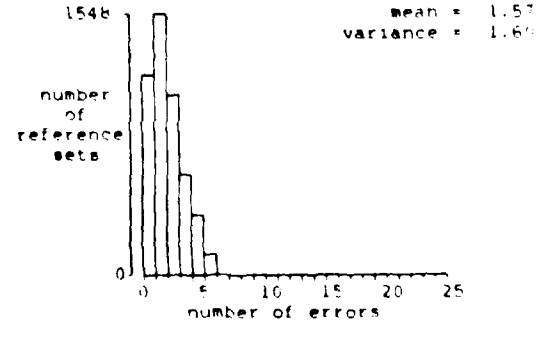
speaker AW



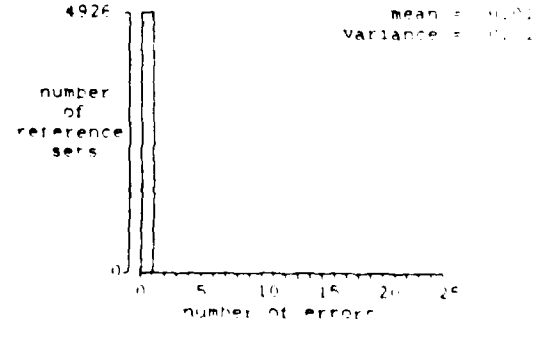
speaker FW



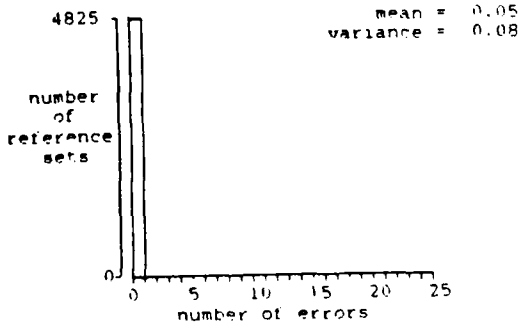
speaker IW



speaker MW



speaker RW


4825
mean = 0.05
variance = 0.08
number of reference sets
0
0
5
10
15
20
25
number of errors

speaker SW

DOCUMENT CONTROL SHEET

Overall security classification of sheet .....UNCLASSIFIED.....

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the box concerned must be marked to indicate the classification eg (R) (C) or (S))

1. DRIC Reference (if known)	2. Originator's Reference	3. Agency Reference	4. Report Security Classification
	MEMORANDUM 3926		

5. Originator's Code (if known)	6. Originator (Corporate Author) Name and Location
	ROYAL SIGNALS AND RADAR ESTABLISHMENT

5a. Sponsoring Agency's Code (if known)	6a. Sponsoring Agency (Contract Authority) Name and Location

7. Title

RANK-ORDERING OF SUBJECTS INVOLVED IN THE EVALUATION OF AUTOMATIC SPEECH RECOGNISERS

7a. Title in Foreign Language (in the case of translations)

7b. Presented at (for conference papers) Title, place and date of conference

8. Author 1 Surname, initials	9(a) Author 2	9(b) Authors 3,4...	10. Date	pp. ref.
SMITH, D C	RUSSELL, M J	TOMLINSON, M J		

11. Contract Number	12. Period	13. Project	14. Other Reference

15. Distribution statement

UNLIMITED

Descriptors (or keywords)

continue on separate piece of paper

Abstract

It is well known that the performance of current automatic speech recognisers varies significantly between talkers. Hence it is essential for anyone involved in the evaluation of speech recognition systems, or the assessment of the acceptability of such systems in a particular application, to be able to 'calibrate' potential talkers.

The purpose of this memorandum is to provide measurements of the expected performance of a typical speech recogniser on a group of forty subjects who have provided recordings for speech recogniser evaluation. The group

/includes

includes all but one of the speakers from the RSRE speech database, all of the speakers who contributed towards the UK part of the NATO RSG10 spoken digit database, and a group of pilots involved in the assessment of automatic speech recognition in avionics applications at the Royal Aircraft Establishment at Bedford and Farnborough.

A method of ranking the subjects based on simpler measurements was also considered. This method potentially requires much less computation and was found to correlate well (rank correlation 0.687) with the ranking by expected error-rate.

END

DTIC

8-86