Proceedings of The Institute of Acoustics

SOME IMPLICATIONS OF THE EFFECT OF TEMPLATE CHOICE ON THE PERFORMANCE OF AN AUTOMATIC SPEECH RECOGNISER

M J Russell, J C A Deacon and R K Moore

Royal Signals and Radar Establishment, Malvern

INTRODUCTION

Researchers in the field of Automatic Speech Recognition (ASR) frequently quote 'recognition rate' as a measure of the capabilities of new ASR algorithms. The inadequacy of this method of assessment is well documented [1]: it has been shown that differences in recognition rate tend to reflect differences in test conditions, such as speaker, vocabulary or environment, rather than the relative capabilities of alternative algorithms. However, even if these variables are controlled, for example by the use of standard databases, there are other factors, more closely linked to the algorithms themselves, which can critically influence performance. One such factor is the choice of reference words, or 'templates'.

This paper is concerned primarily with the effect of reference selection on the performance of 'whole-word pattern matching' algorithms for automatic speech recognition. However, it will be seen that the arguments are equally applicable to a wide category of general pattern recognition strategies which rely on the calculation of similarities or dissimilarities between an unknown pattern and a set of known 'reference' patterns, followed by nearest-neighbour classification.

The experimenter who wishes to evaluate such a recognition system on a prerecorded database must first decide on a criterion for selecting one or more examples of patterns from each pattern class. These patterns are used by the recogniser as references. This raises several important issues. Firstly, the performance of the recogniser may be crucially dependent on the choice of reference patterns, in this case the recognition accuracy achieved will simply reflect the performance of the reference selection criterion. Secondly, the criterion for selecting reference patterns may be biased towards particular types of recognition algorithm so that the apparently superior performance of an algorithm is simply a consequence of this bias. Finally, and more tentatively, the extent to which performance depends on reference selection may vary from speaker to speaker, and this may have implications for the quantitative assessment of speaker consistency.

This paper presents an investigation of the effect of reference choice based on the calculation of distributions of error rates over a large number of randomly selected reference sets.

Experimental results are presented which show that, for the particular speech recognition algorithm which was considered, recognition accuracy is crucially dependent on the choice of reference words, but that the extent of this dependency varies significantly between speakers.

Error rate distributions are shown to be useful for assessing reference selection criteria and for comparing the performance of different recognition algorithms.

Finally, possible quantitative relationships between the degree of effect of reference selection and speaker consistency are discussed.

NOTATION AND METHOD

A nearest-neighbour classification rule with respect to a dissimilarity measure D is a simple pattern recognition strategy in which an unknown pattern u is classified as follows: the dissimilarity D(r,u) is calculated between each pattern r from a set R of reference patterns, u is then assigned to class c, where

$$D(r',u) = \min_{r \in R} D(r,u)$$

and r' is a reference pattern from class c. Such a rule will be denoted by [D,R] to emphasise that the classification of an unknown pattern depends not only on the dissimilarity measure D, but also on the reference set R.

Consider the problem of evaluating this type of classification rule using a prerecorded database S. In order to avoid the over optimistic performance which might result from resubstitution (i.e. testing on patterns which have previously been used for training) it is usual to partition the database into two disjoint sets: a set of training patterns and a set of test patterns.

A reference set is just a subset of the training set which contains one or more patterns from each pattern class. If R is such a reference set, then E[D,R] will denote the number of test patterns which are misclassified by the rule [D,R].

The purpose of this paper is to determine the extent to which E[D,R] depends on R in the special case where the patterns in S represent isolated spoken words and D is the measure of dissimilarity obtained from a 'Dynamic Time-Warping' algorithm (corresponding to case p=1 in [2]). The approach taken is to estimate the underlying distribution of values of E[D,R] by computing E[D,R] over a large number of randomly chosen reference sets R. Clearly this can be achieved by calculating E[D,R] directly for each R, however this involves unnecessary repeated calculation of the same dissimilarities. Hence, since dynamic time-warping is computationally expensive, it is better to procede as follows:

First, the dissimilarity between each training pattern and each test pattern is calculated. 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 refence set R, classification of a test pattern u according to the rule [D,R] is achieved by simply searching the ordering corresponding to u until a training pattern r belonging to R is encountered: u is assigned to the class of r. In this way several thousand reference sets can be processed in the time which would be required to process just ten or twenty reference sets directly.

EXPERIMENTAL RESULTS

The results in this section were obtained by applying the above method to the

NATO RSG10 spoken digit database [3]. This database contains recordings of isolated and connected digits spoken by nineteen speakers in four languages. Isolated digits were recorded in five tables, called SA, SB, 1A, 1B and 1C, each containing ten examples of each digit. Tables SA and SB are intended for training and tables 1A, 1B and 1C for testing. For the purposes of the present experiments the speech was digitised using a 20ms frame-rate, 19-channel vocoder.

Effect of reference selection

For each of the 54 speaker-language-test set combinations, the underlying distribution of E[D,R] was estimated by calculating E[D,R] for 50000 reference sets R, each containing one reference per class, chosen randomly from the corresponding training set SB. All but one of the distributions confirm that the number of errors does depend on the choice of reference set. However, the degree of this dependency varies greatly between speakers, as can be seen from figure 1. The distribution in figure 1a (speaker RM, language English, table 1A) shows very little sensitivity to reference selection. Almost 99% of the reference sets considered resulted in no errors and the remaining 1% each resulted in just 1 error. This contrasts sharply with the distribution shown in figure 1b (speaker MW, language English, table 1C). In this case the number of errors varies between 0 and 23 depending on the choice of reference set. It should be noted that both of these recordings were made under identical strictly-controlled conditions on the same day in the same laboratory.

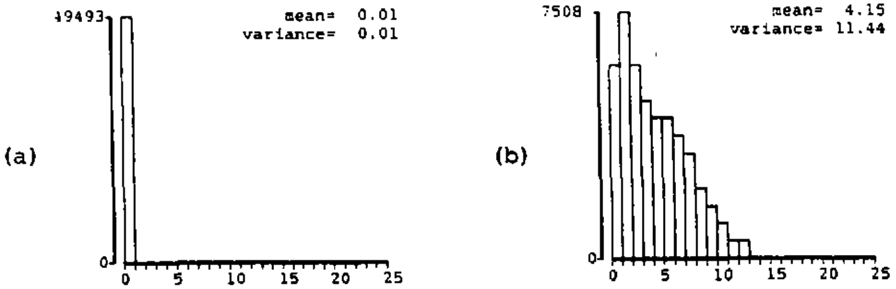

Figure 1: Observed distributions of error rates over 50000 randonly chosen reference sets for (a) speaker RM, language English, table 1A and (b) speaker MW, language English, table 1C.

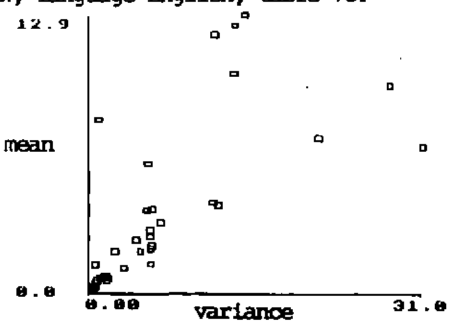

Figure 2: Scatter plot of mean against variance for the error-rate distributions associated with each of the NATO RSG10 isolated digit test sets.

Overall trends can be deduced from figure 2, which shows a scatter plot of mean against variance for each of the 54 distributions. As might be expected, the figure shows that the degree of dependency on reference selection, indicated by the variance of the distributions, increases with the average number of errors.

Evaluation of reference selection criteria

The dependency of error rate on reference selection, demonstrated in the previous section, highlights the need for reliable reference selection criteria. Using the distributions from the previous section, the performance of such criteria can be assessed. In this section two selection criteria, denoted by C1 and C2, are considered.

For each pattern r in the training set let MID(r) (mean inclass dissimilarity) denote the average dissimilarity between r and all other patterns in the training set which belong to the same class as r. Similarly let MOD(r) (mean outclass dissimilarity) denote the average dissimilarity between r and all patterns in the training set which are not in the same class as r.

Under criterion C1 (respectively C2), a class c is represented by the pattern r' in the training set such that

$$\mathrm{MID}(r') = \min \mathrm{MID}(r)$$

(respectively

$$\frac{\mathrm{MID}(r')}{\mathrm{MOD}(r')} = \min \frac{\mathrm{MID}(r)}{\mathrm{MOD}(r)} \;)$$

where the minimum is taken over all patterns r in the training set which belong to class c.

Comparison with the distributions from the previous section shows that selection of reference sets according to criterion C1 will result in the best achievable error rates for 72% of the test sets. The percentage of error rates for criterion C1 which are less than the mean of the corresponding distributions is 94%. This means that 94% of the time a reference set chosen according to criterion C1 will perform better than an 'average' reference set. There was no significant difference between the performance of C1 and that of C2.

Algorithm comparison

A further implication of the importance of reference selection is that the results of an experiment which compares two classification algorithms will be more meaningful if the effect of reference selection is removed. This suggests that rather than comparing error rates for specific reference sets it is better to compare distributions of error rates, or at least the means of such distributions. For example, figure 3 shows the results of an experiment to determine the advantage of representing each class by two reference patterns rather than one. Figure 3a is a scatter plot of error rate for reference sets containing one reference per class (selected according to C2) against error rate for reference sets containing two refences per class (selected using a similar criterion). For 39 of the 54 test sets, both reference sets result in no errors and so comparison of the two 'algorithms' must rely on the remaining 15 results.

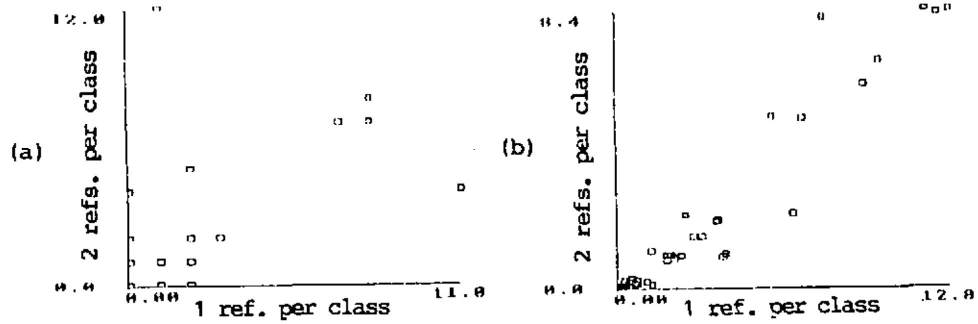

Figure 3: Scatter plot of error rate for one reference per class against error rate for two references per class for each of the NATO RSG10 isolated digit test sets. Figure (a) shows error rates for reference sets chosen according to criterion C2, (b) shows mean error rates over 10000 randomly chosen reference sets of each type.

A much clearer picture emerges from figure 3b, where mean error-rates for one reference per class are plotted against mean error-rates for two references per class. The figure suggests that the expected reduction in isolated digit recognition errors resulting from representing each class by two references is approximately 33%

Speaker consistency

The final aim of this paper is to ask if the speaker-dependent nature of the effect of reference selection can be exploited to quantitively measure 'speaker consistency'.

Initial experiments suggest that the shape of the underlying distribution of E[D,R] might provide such a measure. Figure 4 shows a comparison of observed error distributions with distributions predicted according to binomial laws with the same means. Figure 4a corresponds to an 'inconsistent' speaker (speaker MW, language English, table 1C), whereas figures 4b and 4c correspond to a 'consistent' speaker (speaker RM) speaking the digits (language English, table 1A) and the orthographic alphabet (language English, alphabet table 1C, recorded under identical conditions to the RSG10 database) respectively. For the inconsistent speaker, the shapes of the two distributions are quite different, but for the consistent speaker the match between the two distributions is good. Moreover, the fit remains good, even when the mean error rate increases due to increased vocabulary confusability.

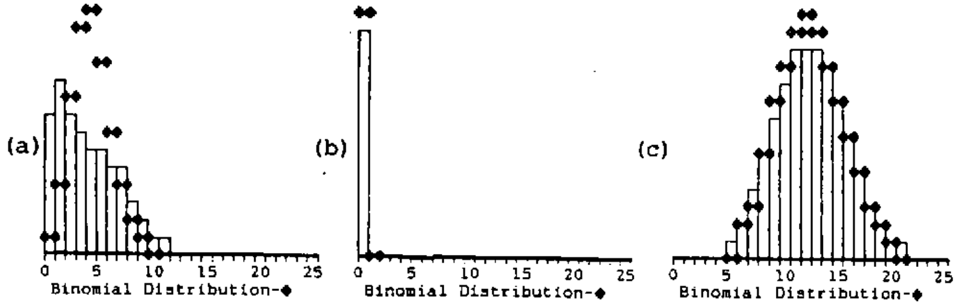

Figure 4: Comparison of observed error-rate distributions with distributions predicted according to a binomial rule with the same mean: (a) speaker MW, language English, table 1C, (b) speaker RM, language English, table 1A and (c) speaker RM, language English, alphabet table 1C.

CONCLUSION

The experiments reported in this paper confirm that the performance of whole-word pattern matching algorithms can depend crucially on the choice of reference set. The extent of this dependency has been shown to vary significantly between speakers. It has been demonstrated that distributions of error rate with respect to reference set provide a useful tool, both for evaluating reference selection methods and for comparing algorithms. Finally, the implications for assessing speaker consistency of the speaker-dependent nature of the importance of reference selection have been considered.

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, 728-731, (1981).

[2] H Sakoe and S Chiba, 'Dynamic programming algorithm optimisation for spoken word recognition', IEEE Trans. ASSP, 26, 43-49, (1978).

[3] 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.