Proceedings of The Institute of Acoustics

TOWARDS AN INTEGRATED DISCRIMINATIVE NETWORK FOR AUTOMATIC SPEECH RECOGNITION

R K MOORE, D BEARDSLEY, M J RUSSELL and M J TOMLINSON

ROYAL SIGNALS AND RADAR ESTABLISHMENT

ABSTRACT

Whole word pattern matching using dynamic time-warping (DTW) has achieved considerable success as an algorithm for both isolated and connected word recognition. However, the performance of such an algorithm is ultimately limited by its inability to discriminate accurately between similar sounding words. This paper presents an alternative DTW approach which is able to focus attention onto those parts of a speech pattern which serve to distinguish it from similar patterns. Preliminary results indicate that discrimination between similar sounding words can be improved, and the implications for future speech recognition algorithms are discussed.

INTRODUCTION

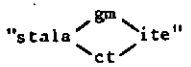
Current automatic speech recognition algorithms use non-linear time registration techniques to match an unknown speech pattern against a set of reference patterns [1,7]. The identity of the unknown pattern is determined by the resulting best match. The technique is very successful for vocabularies which contain easily distinguishable words, but for vocabularies which contain similar sounding words, errors often occur. This is not too surprising since, almost by definition, similar sounding words are difficult to disambiguate. However, in this instance, the matching algorithm is also partly to blame.

The reason is that current whole word approaches to speech pattern matching calculate a figure of similarity which gives equal weight to all parts of the patterns. Hence, when words which differ only slightly from each other, such as "stalagmite" and "stalactite", are compared, the algorithm may be swayed by irrelevant differences in regions which are unimportant. For example, if the "stala-" in "stalagmite" happens to be very similar to the "stala-" in the "stalactite" reference pattern, then it might be misrecognised. What is needed is a matching algorithm which is able to focus its attention onto those parts of a speech pattern which serve to distinguish it from other similar sounding patterns.

Rabiner and Wilpon [8] have proposed a two-pass solution to this problem. A vocabulary is divided into subsets, where each subset contains words which sound similar. On the first pass an unknown word is classified as belonging to one of these sets using a standard DTW algorithm. A second pass is then performed within the identified set using word specific weighting functions to emphasise the regions of importance. The technique succeeds in achieving an increase in recognition accuracy for difficult vocabularies such as the alphabet.

The most obvious disadvantage of this approach is the requirement for two recognition passes. They are needed because the weighting functions used on the second pass are only relevant to the particular types of difference characterised by each of the vocabularly subsets. A degradation of performance would result from the use of the weighting functions on dissimilar words. What is required, therefore, is a technique which is able to focus its recognition for similar sounding words, but which remains unaffected for words which are obviously different. Such a technique is described in this paper.

A network type data structure is derived automatically from reference speech patterns [5]. Alternative paths through the network correspond to the different constitutent patterns, but the network is so constructed that regions not contributing to the difference are integrated into common paths. In this way irrelevant variations in these regions cannot have an adverse differential effect on the recognition decision. In the case of "stalagmite" and "stalactite" the network would take the form:-



CONSTRUCTING A DISCRIMINATIVE NETWORK

In order to construct a network of the type just described it is first necessary to determine in which regions a network should branch, and secondly, to provide a scheme for combining the information in regions where the network is integrated. Both Moore [5] and Rabiner [8] have proposed the use of a standard DTW algorithm as a solution to the first of these.

Suppose that $V=[V(i);1\leq i\leq I]$ and $H=[H(j);1\leq j\leq J]$ are two samples of speech, where V(i) and H(j) are vectors which characterise the speech signals at times i and j respectively. The cumulative distance between V and H is found by dynamic programming using the recursive equation

$$D(i,j) = \min_{(p,q)\in P} [D(i-p,j-q)+d(V(i),H(j))]$$

where P is a set of simple productions and d is a metric on the set of vectors containing V(i) and H(j). Hence the distance D(I,J) between the two patterns is the sum of the local distances d(V(i),H(j)) along the optimal time registration path (see [6] for a full explanation of the terms).

If different examples of the same word are compared, then one would expect the local distances to be small. As a consequence the total distance between the patterns would be small. On the other hand, if examples of different words are compared, then the local distances would be high. If, however, two different but similar sounding words are compared, then some local distances would be small (the similar regions) and some local distances would be large (the differing regions). Hence the local distances may be used as a guide to the formation of a network.

TOWARDS AN INTEGRATED DISCRIMINATIVE NETWORK FOR AUTOMATIC SPEECH RECOGNITION

The algorithm for constructing a network from two speech patterns is thus as follows:

> For every point (i,j) on the optimal time registration path perform the following actions; if the local distance d(V(i),H(j)) is greater than some threshold T, then retain the vectors V(i) and H(j) separately, otherwise average V(i) and H(j) to form a new vector.

Obviously the discriminating power of the network depends crucially on the value of T. Too low a value will separate the speech patterns entirely, whilst too high a value will produce a single averaged pattern with no discriminating power. However, the arguments put forward in this paper suggest that between these two extremes there might be a value of T for which discrimination is better than that obtainable using the original patterns.

EXPERIMENTS

In the following experiments acoustic analysis was performed using a 19 channel vocoder, producing one vector every 20ms. The metric d was Euclidean, and P was the set {(1,0),(1,1),(0,1)}. Ten examples of "stalagmite" and ten examples of "stalactite" were obtained from a single speaker (speaker-A) as training data. From these, one example of each word was selected as a reference pattern. A further 50 examples of each word were then obtained from the same speaker to form a test set. The test set was then recognised by matching it against the two reference patterns using a standard DTW algorithm. As expected no errors were made since speaker-A is known to be a consistent speaker. Therefore, in order to introduce some independent variability, 50 examples of each word were obtained from a second speaker (Speaker-B). These were then recognised using speaker-A's references and an error rate of 50% was obtained. Networks were then constructed from the reference patterns and recognition proceeded for speaker-B's test set at various values of the threshold T. Figure 1 shows how the error rate drops to 11% at an intermediate value of T. The corresponding network is shown in Figure 2. At this threshold value the recognition rate for speaker-A was still 100%.

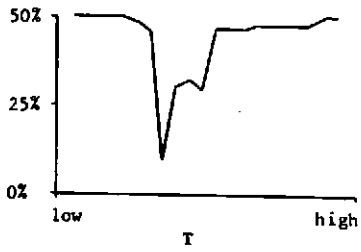


Fig.1: Error rate-vs-threshold.

Fig.2: "Stalagmite-stalactite" network.

The experiments were repeated for the difficult alphabet pair "bee-dee" using references from both speaker-A and speaker-B. Again discrimination is improved using a network. It was also apparent that speaker-B was less consistent than speaker-A, so three more experiments were conducted on speaker-B alone. Table I summarises all the results. In each case a network was able to overcome some of the irrelevant variability in the data and obtain an increase in recognition accuracy.

WORD PAIR	REFERENCE SPEAKER	TEST SPEAKER	NORMAL DTW	NETWORK
stalagmite-stalactite	A	A	0	0
	B	A	50	11
bee-dee	A	A	5	5
	B	B	32	13
	A	B	6	2
	B	A	48	14
tee-dee	B	B	28	10
kay-jay	B	B	4	2
5-9	B	B	15	11

Table I: Percentage Error rates for the network experiments.

DISCUSSION

The experiments reported in this paper have shown how an integrated discriminative network can improve the performance of an automatic speech recogniser by focusing recognition onto the important regions of a speech pattern. As a one-pass algorithm, it has a significant advantage over the two-pass solution proposed by Rabiner and Wilpon [8]; a network prepared from two similar sounding words can be used within a larger vocabulary with no modification. Similarly, further words can be integrated into an existing network with relative ease.

The experiments have also shown how the effects of cross-speaker variability can be reduced. This confirms the hypothesis, put forward recently by Gupta and Mermelstein [2], that a reference network might aid speaker independent word recognition.

A practical advantage is that a network can provide very efficient storage of reference speech patterns. Consequently computation during recognition can be reduced significantly. This aspect has been recently emphasised by Tanaka et al [9] who have demonstrated a similar scheme for handwritten character recognition.

However, perhaps the most important feature of the discriminative network is that it is derived automatically simply on the basis of a same/different contrast between speech patterns. Consequently, the different branches of a network, being determined acoustically and semantically, could be considered to constitute 'phonemic' segments. Thus the discriminative network can be seen as a step towards the phonetic networks of Jelinek [3], but without the need for vector quantisation, and also as a possible implementation of Klatt's model of speech perception [4].

TOWARDS AN INTEGRATED DISCRIMINATIVE NETWORK FOR AUTOMATIC SPEECH RECOGNITION

REFERENCES

1. J S BRIDLE and M D BROWN 1979 Proc. Inst. of Acoustics Autumn Conf. Connected Word Recognition Using Whole Word Templates.

2. V GUPTA and P MERMELSTEIN 1982 JASA 71, 1581-1587. Effects of Speaker Accent on the Performance of a Speaker-Independent, Isolated-Word Recogniser.

3. F JELINEK 1976 Proc. IEEE 64, 532-555. Continuous Speech Recognition by Statistical Methods.

4. D KLATT 1979 Journal of Phonetics 7, 279-312. Speech Perception: a Model of Acoustic-Phonetic Analysis and Lexical Access.

5. R K MOORE 1980 Unpublished Report, Dept. Phonetics, University College London. An Investigation into some Fundamental Problems in the Automatic Recognition of Continuous Speech (2).

6. R K MOORE, M J RUSSELL and M J TOMLINSON 1982 Proc. IEEE Int. Conf. ASSP, 1270-1272. Locally Constained Dynamic Programming in Automatic Speech Recognition.

7. L R RABINER, A E ROSENBERG and S E LEVINSON 1978 IEEE Trans. ASSP 26, 575-582. Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition.

8. L R RABINER and J G WILPON 1981 Bell System Technical Journal 60, 739-766. A Two-Pass Pattern-Recognition Approach to Isolated Word Recognition.

9. H TANAKA, Y HIRAKAWA and S KANEKU 1982 IEEE Trans. PAMI 4, 18-25. Recognition of Distorted Patterns Using the Viterbi Algorithm.