# ON THE ASYMPTOTIC PERFORMANCE OF NEAREST NEIGHBOUR PATTERN CLASSIFIERS IN SPEECH RECOGNITION

Simon J. BUTLER * and Frantz CLERMONT **

* Speech, Hearing and Language Research Centre
Macquarie University
Sydney, New South Wales 2109, Australia

** Computer Sciences Laboratory
Research School of Physical Sciences
The Australian National University
Canberra, ACT 2601, Australia

ABSTRACT : When distance measures based on Linear Prediction are used in Nearest Neighbour speech recognisers with a large number of training samples, it is found that the recognition performance is independent of the distance measure used. This contrasts with the case of small training sample sizes, in which performance is highly sensitive to choice of distance measure. The "asymptotic nearest neighbour equivalence" of this class of distance measures is explained and demonstrated in a vowel recognition experiment.

## INTRODUCTION

The Nearest Neighbour (NN) method of pattern classification is widely used in speech recognition systems where it is often referred to as "template matching". The popular Dynamic Time Warping (DTW) approach is an example of NN-classification which is the basis of many commercial speech recognisers. A key research issue in NN-speech recognition has been the choice of distance measures and their effect on recognition performance. A number of distance measures have been theoretically derived, which are naturally suited to the speech problem by virtue of their spectral matching properties. There are, however, few theoretical results which directly indicate what recognition performance is to be expected of the various distance measures.

In this respect, the speech community has relied on empirical evaluations of distance measures in NN-speech recognisers, which are trained and tested on carefully designed speech databases. A class of related distance measures which are derived from the Linear Prediction (LP) method of signal analysis (Figure 1) has been the subject of numerous evaluations, and it is the performance of these measures which is addressed here. Some of these distance measures use the linear prediction coefficients themselves (LPC) whereas others use one-to-one, non-linear transformations of them.

Empirical studies have achieved relatively consistent results, whether it be for vowel recognition (Paliwal and Rao, 1982), consonant recognition (Clermont, 1982) or word recognition (Davis and Mermelstein, 1980). The "Cepstral" distance measure (CEP_EUC), for instance, yields the best performance amongst the distance measures indicated in Figure 1. Apparently, the recognition accuracy is highly dependent on choice of distance measure. An important limitation of these empirical studies is the small number of speech samples used in training. It is important, for instance, to know to what degree the high cost of collecting additional training data can be offset by improved performance. While a small number of studies have examined large training sample behaviour (e.g. Paliwal and Rao, 1983), they do not contrast a variety of different distance measures. This paper addresses the relative performance of the class of LP based distance measures when a large number of training samples are available.

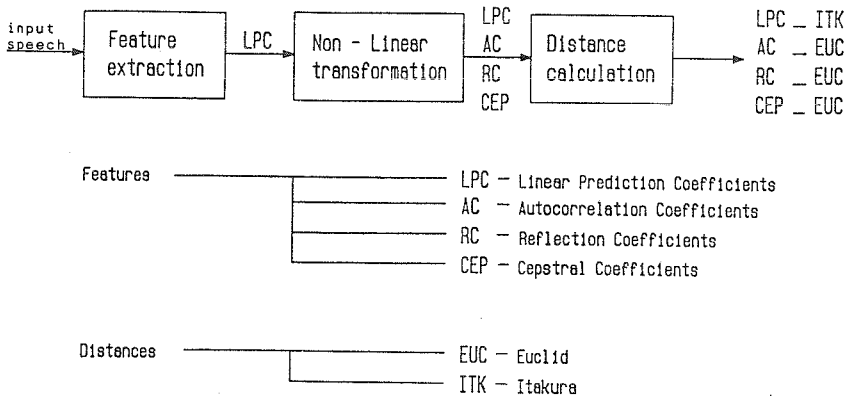


Figure 1: The Class of Distance Measures based on Linear Prediction

## OPTIMAL PERFORMANCE OF SPEECH CLASSIFIERS

The feature extraction process of a speech recogniser, such as LPC, provides the basic information that is necessary to discriminate between classes of speech sounds. An important property of the non-linear transformations used in NN-speech recognition (Figure 1) is that they are one-to-one and therefore involve no loss of information. On intuitive grounds, therefore, the classification accuracy of speech using the transformed LP coefficients should be identical to that using the LP coefficients themselves. This assumes an ideal or optimal pattern classifier which can use all of the information in the features. The optimal classifier is usually defined as the Bayes classifier which can be derived from the probability density functions of the features for each class. An example of optimal classification compared with NN-classification using one reference template per class appears in Figure 2. In Figure 2(a), the reference templates have been chosen so that the NN decision boundary ($d_{NN}$) coincides with the optimal decision boundary ($d^*$), achieving optimal performance. If the features are transformed (Figure 2(b)), however, the NN boundary no longer coincides with the optimal decision boundary. Thus, whilst the optimal performance is invariant to transformation, small sample NN-classification performance is not.

## ASYMPTOTIC PERFORMANCE OF NN-SPEECH RECOGNITION

In the asymptotic case of NN-classification (infinite number of training samples) Cover and Hart (1967) have established that the NN probability of error R is bounded from above by twice the Bayes error $R^*$ :

$$R^* \leq R \leq 2R^*$$

This bound, valid for metric distance measures, is independent of the particular distance used. The NN-asymptotic performance, like the optimal (Bayes) case is a function of only the probability density functions of the features. This implies that identical asymptotic performance is obtained irrespective of the distance measure or one-to-one transformation of the features used.

In one of the few theoretical papers on relative performance of distance measures, Gray et al. (1980) introduced the concept of "nearest neighbour equivalence" for those distance measures which achieve the same performance in NN systems. Since the present discussion relates to distance measures which achieve identical performance in the asymptotic limit, this paper puts forward the notion of "asymptotic nearest neighbour equivalence". According to this definition, the members of the LP class of distance measures with metric properties are asymptotic nearest neighbour equivalent.

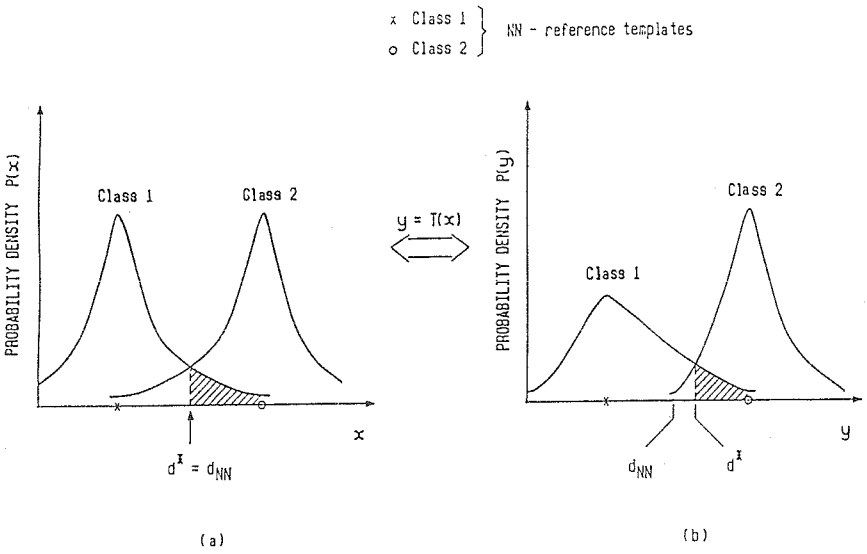


Figure 2: Decision Boundaries for Optimal and NN-Classification

## ASYMPTOTIC PERFORMANCE IN VOWEL RECOGNTITION

Of the few studies that have addressed speech recognition performance for large training samples, it appears that only Paliwal and Rao (1983) have compared different LP based distance measures. They conclude that vowel recognition performance converges to an asymptotic value when a certain number of training samples is exceeded. Furthermore, their asymptotic values are substantially different for the two distance measures considered. However, closer examination of the Paliwal and Rao data (1983, Figure 2) indicates that the final slopes of their curves are approximately 0.5 percent/sample, suggesting that convergence has not been achieved.

To evaluate the performance in the large sample case, a software system for NN-speech recognition called KNN (for Kth Nearest Neighbour) was developed. This system offers the capability to use a number of different distance measures and to perform different feature transformations on LP encoded speech data. Recognition is performed using the K-nearest neighbours where K is variable, and training data may also be clustered prior to recognition. In the present experiments, K equals 1 and clustering is not attempted.

An Australian English vowel database is used and consists of linear prediction coefficients extracted from the most stationary part of vowels in /CVd/ context. The database consists of five repetitions of eight vowels (as in hid, head, had, hard, hod, hood, hudd, herd) in seven consonantal contexts /h,b,d,g,p,t,k/ produced by four adult male speakers. The vowel data is partitioned into a training and a test set. From the training set, the required number of samples are selected randomly allowing for multiple experiments with different training data. Ten runs of each experiment are averaged to reduce the variation that occurs with different training samples.

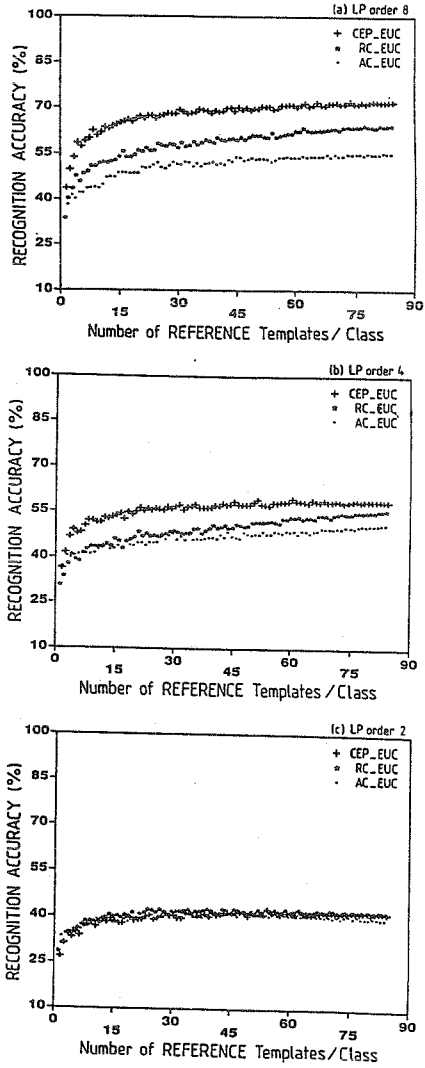


Figure 3: Large Sample Behaviour in Vowel Classification

Initial results for LP of order 8 with a number of distance measures (Figure 3(a)) are comparable with previous studies such as Paliwal and Rao (1983). There are noticeable differences between the performances of different distance measures even for 84 reference templates per class. Most importantly, however, there is no indication of convergence of recognition performance. Apparently, the rate of convergence of NN speech classifiers is very slow.

One factor which strongly influences the small sample behaviour of pattern classifiers is the intrinsic dimensionality of the data (Kanal, 1974). An investigation was therefore undertaken to examine the NN-speech recognition performance under conditions of reduced intrinsic dimensionality. In the case of speech data, the intrinsic dimensionality may be reduced by lowering the order of LP analysis. Reduced order recognition experiments are likely to achieve lower performance, but nevertheless are useful in verifying the theory of asymptotic nearest neighbour equivalence.

Vowel recognition for LP of orders 4 and 2 are given in Figures 3(b) and 3(c). In the case of order 2, the recognition performance converges very quickly and the different distance measures have the same asymptotic performance. For order 4, convergence is achieved by 84 reference templates for the Cepstral (CEP_EUC) distance measure. In addition, the other distance measures are clearly approaching the same asymptotic limit as the number of templates becomes large. These results are consistent with the previously advanced notion of asymptotic nearest neighbour equivalence of the LP based class of distance measures.

## DISCUSSION

The performance of any speech recogniser is dependent on the information available from feature extraction (in the present case LP). Whilst NN classification is suboptimal, its asymptotic performance is at worst twice the optimal performance. The results of Figure 3(a) show, however, that for a typical speech recognition experiment, the convergence is so slow that asymptotic performance is never attained. That asymptotic nearest neighbour equivalence of the LP based distance measures has not been previously observed is indication that current speech recognition methodology is performing poorly with respect to the optimal performance. It is worth noting that progress in reducing the error rates of speech recognisers over the last decade has not reached previously held expectations. The results presented here suggest an explanation for this in terms of the intrinsically slow convergence of NN classifiers. Whilst choice of distance measure has a substantial effect on both non-asymptotic performance and rate of convergence, it is not at all clear to what degree performance can be improved through the development of new distance measures.

## REFERENCES


Clermont, F. (1982) "Word recognition based on phoneme template: comparison of several parametric representations, distance measures and time alignment techniques for recognition of English consonants", STL-Quarterly Report, July 1982, Speech Technology Laboratory, Santa Barbara, California.

Cover, T.M. and Hart, P.E. (1967) "Nearest neighbour pattern classification", IEEE Trans. Inf. Theory, 13, 21-27.

Davis, S.B. and Mermelstein, P. (1980) "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. Acoust. Speech, Signal Processing, 28, 357-366.

Gray, R.M., Buzo, A., Gray, A.H. Jr and Matsuyama, Y. (1980) "Distortion measures for speech processing", IEEE Trans. Acoust. Speech, Sig. Proc., 28, 367-376.

Kanal, L. (1974) "Patterns in pattern recognition: 1968-1974", IEEE Trans. Inf. Theory, 20, 697-722.

Paliwal, K.K. and Rao, P.V.S. (1982) "Evaluation of various linear prediction parametric representations in vowel recognition", Signal Processing 4, 323-327.

Paliwal, K.K. and Rao, P.V.S. (1983) "Application of K-nearest-neighbour decision rule in vowel recognition", IEEE Trans. Patt. Analysis Mach. Intelligence 5, 229-231.