THE APPLICATION OF ADAPTIVE CONSTRAINT SATISFACTION NETWORKS TO ACOUSTIC PHONETIC ATTRIBUTE DETERMINATION

Ian Howard and Mark Huckvale .

ABSTRACT

Speech is a signal with a complex underlying structure and considerable variability. In order to determine acoustic phonetic correlates of speech one must take this structure into account. To specify its structure adequately a priori would be very difficult. One would also have to ensure that any fixed structure imposed initially was not restrictive. Characteristics of learning machines appear useful for this type of problem because they have potential for acquiring internal structure, so less needs to be imposed in advance. The type of learning machine investigated is the multi-layer perceptron (MLP). It is shown that one may train such a system to perform very well at standard pattern recognition tasks. It is compared against a standard technique for discriminant analysis, a Bayes classifier for normal patterns.

INTRODUCTION

The aim of the work here is to investigate the usefulness of the MLP in acoustic-phonetic determination and its ability to learn structural relationships in patterns. The case of voicing determination is used here an an example. A MLP network [ref 1] may be considered as performing a trainable non-linear transformation that, when presented with an input pattern vector, generates an appropriate response. When the task involves classifying frames of a time-varying signal, such a decision process will clearly benefit from access to the signal over a wider time-window. This may be achieved by allowing access to adjacent frames or by increasing the time-window of the elements of the pattern vectors. This work is concerned with the classification of pattern vectors in isolation or in conjunction with their adjacent pattern vectors, and will not address the problem of increasing performance by demanding that the sequence of the decisions made be subject to constraint. In the case of a MLP with no hidden units, its output represents the weighted sum of the pattern vector elements passed through a non-linear function. Since the latter is monotonic the output pattern class depends on a linear combination of the input elements and a threshold. In this case the training results in the selection of suitable weights. When hidden units are introduced, the output pattern class is no longer a linear threshold function of the input vector. In the latter case one may expect increased performance since the network may exploit more complex inter-relationships between components of the pattern vector. Complex relationships must, of course, exist for this to happen.

BAYES CLASSIFIER

An established approach to classification of pattern vectors is the Bayes classifier [ref 2] which has its foundation in statistical decision theory. If one assumes a normal distribution of the pattern classes, then the optimum decision boundaries can be computed in terms of the respective mean vectors and covariance matrices [ref 3]. Provided the patterns are genuinely normally distributed, such a scheme is optimal. However, this is not always the case. One may then, of course, resort to functional approximation techniques, but they are not considered here, since our interest lies with the MLP.

Dept. of Phonetics and Linguistics, University College London, UK.

TRAINING

The training scheme involves supplying the classifiers with labelled pattern vectors. The transformation required to map the input vectors to the specified output states are then computed.

EXPERIMENT 1

The speech data was recorded anechoically together with the output of a laryngograph [ref 4]. The latter was required by a reference voicing analyser [ref 5]. This provided a convenient and acceptably good method of annotating a very large amount of speech data for voicing, with the minimum of human effort. The material was the "Rainbow" passage [ref 6] and it was recorded for five adult male speakers each with two repetitions. The first repetition was used as training data and the other as the testing data. A 19 channel vocoder [ref 7] was used to generate the pattern vectors. Clearly other features could have been used, but optimal feature selection is not the problem addressed here. The pattern vectors were formed by sampling the vocoder channels in 10 ms frames. The samples were the top 50 dB of a logarithmic scale, to help normalize amplitudes. There were, in total, 14616 frames of training data and 16395 frames of test data.

METHOD

The following schemes for voicing determination were examined. In each case there was only one output unit. 1) Bayes normal classifier with 19 input elements. 2) MLP with 19 input elements and no hidden units. 3) MLP with 19 input elements and 19 hidden units. 4) MLP with 19 input elements and two layers of 19 hidden units. 5) MLP with adjacent frames (3*19 input elements) and no hidden units. 6) MLP with adjacent frames (3*19 input elements) and (3*19) hidden units. Each was first trained on the training data and then run on the test data. The MLP training algorithms were run until there was virtually no more improvement in performance, with the same number of training cycles being used in each case. The algorithms were all written in C and ran under Unix on a Masscomp MC5500 series computer.

EVALUATION OF RESULTS

The results are displayed in the form of receiver operating characteristics (ROC). This is a plot of the number of correct classifications against the number of false alarms, for many threshold values. This provides a convenient visual method of comparing performance. The results appear in fig.1. The pattern vectors were not normally distributed, and consequently the Bayes classifier for normal patterns did not operate as well as a more general version may be expected to. It does, however, give a rough indication of the performance that may otherwise be considered "reasonable". Hidden units did not improve performance of the MLP. The MLP with only 19 input elements worked better than the Bayes classifier. The MLP with adjacent frames worked best of all, but this is to be expected since it has access to more information.

EXPERIMENT 2

Observation of the labelling performed by the MLPs suggested that they were, in fact, producing better results than the reference. This was due to the fact that the laryngograph output does not always indicate the presence of voicing whilst observation of the speech waveform indicates that voiced excitation is indeed present. This clearly invalidates the ROC test procedure. Rather than resort to manual correction, one solution was to degrade the speech data. This would

ensure the recognition performance would be poor compared to the reference, and so permit more meaningful comparisons to be made. The speech was degraded to 0dB with uniform density random noise.
In order to hopefully permit the MLP to show its ability to perform complex transformations of the input vector, thus showing up hidden relationships, it was necessary to have input pattern vectors with suitable relationships between their elements. A smoothed version of the vocoder output, generated by applying a 0.2 s window to the former, was used to give features which reflect signal input characteristics over a longer time-scale. Such data is not of direct value to the classification process.

METHOD

This time, the following experiments were run on the noise corrupted speech data: 1) Bayes normal classifier with 19 input elements. 2) MLP with 19 input elements and no hidden units. 3) MLP with 19 normal input elements, 19 smoothed input elements and no hidden units. 4) MLP with 19 normal input elements, 19 smoothed input elements and 19 hidden units. 5) MLP with adjacent frames (3*19 input elements) and (3*19) hidden units.

RESULTS

Fig. 2 shows that the normal Bayes classifier performed about as well as the MLP, in the single vector input case. This was due to the fact that the pattern vectors were now more normally distributed than they had been previously, and consequently the decision function calculated was more appropriate to the classification task. The best results were obtained when adjacent frames were employed, with the smoothed vector case slightly worse. Hidden units did not improve performance.

CONCLUSIONS

The MLP worked better then Bayes normal classifier. The former does not make asumptions about the form of the probability densities of the pattern vectors. Performance of the MLP was highest when adjacent frames were incorporated. Performance of the MLP improved over the single pattern vector case, with the addition of a smoothed pattern vector. Hidden units did not improve performance. This suggests that any complex hidden structure in the data vectors was not significant compared to that more directly available. Thus the MLP, when provided with appropriate pattern vectors, can perform frame labelling for voicing determination at least as well as a normal Bayes classifier.

ACKNOWLEDGEMENTS

This work was supported by Alvey grant MMI/056 and MRC studentship RS-85-2.

REFERENCE

[1] Rumelhart, D.E., Hinton, G.E., and Williams, R.J., I.C.S., Report ICS-8506, University of California, San Diego, (1985).
[2] Tou,J.T., and Gonzales, R.C., Pattern recognition principles, Addison-Wesley, (1974).
[3] Atal, B.S., and Rabiner, L.R., IEEE trans. ASSP, Vol 24-3 June 1976.
[4] Fourcin, A.J. and Abberton, E.R.M., Med. and Biol. Illust. 21, 172-182 (1971).
[5] Hess, W., and Indefry, H., Proc. ICASSP-84, 1-4, (1984).
[6] Mermelstein,P., JASA 61 p581(1977).
[7] Rabiner,L.,R., and Schafer,R.W., Digital processing of speech signals, Prentice-Hall, (1978).

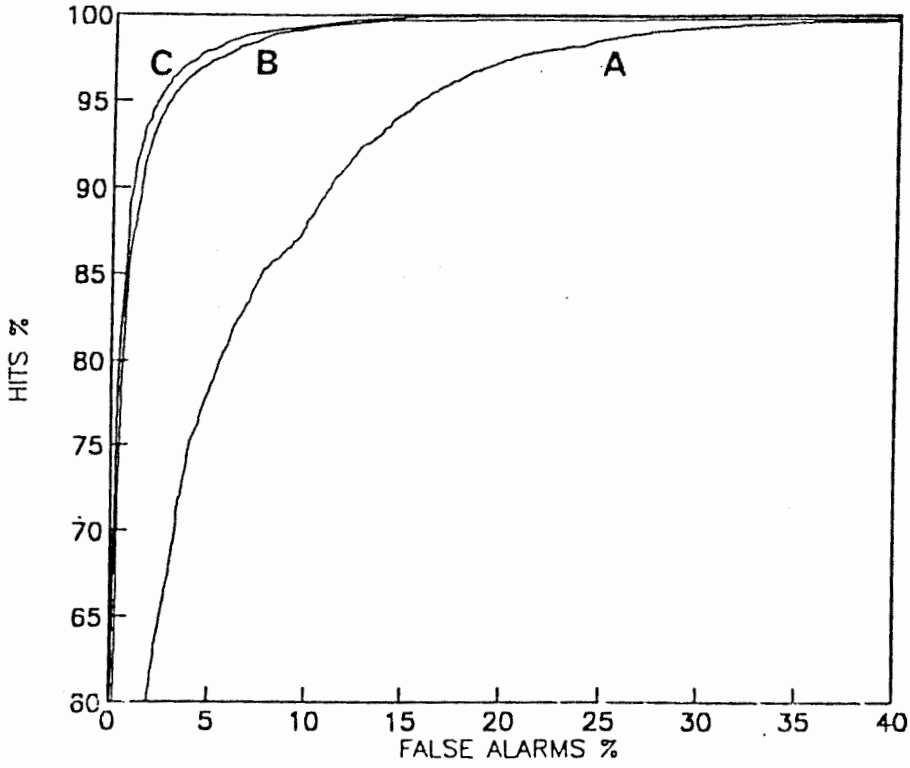

FIGURE 1. Result ROC's with anechoic speech data input for the following schemes: Curve A for Bayes normal classifier with one frame input. Curve B for MLP with one frame input, with and without hidden units. Curve C for MLP using adjacent frames, with and without hidden units.

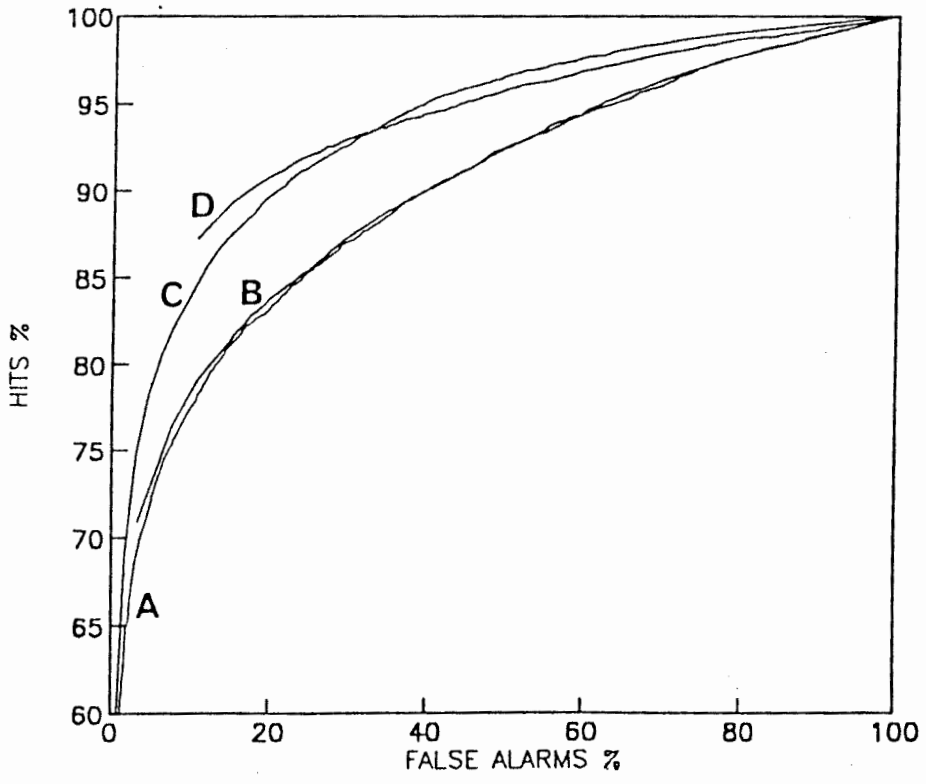

FIGURE 2. Result ROC's with speech degraded to 0 dB SNR with noise for the following schemes: Curve A for Bayes normal clasifier with one frame input. Curve B for MLP with one frame input, with hidden units. Curve C with additional smoothed frame input, with and without hidden units. Curve D MLP using adjacent frame input, and hidden units.