EXPERIMENTS WITH MASK-PERCEPTRONS FOR SPEECH RECOGNITION

Adam Kowalczyk [*], Herman Ferrá [*†] and Gordon Jenkins[*]

[*] Artificial Intelligence Systems Section
Telecom Australia, Research Laboratories

[†] Department of Mathematics
Monash University

ABSTRACT - The paper discusses results from a series of experiments on isolated word recognition using neural networks (multilayer perceptrons). It shows that high recognition accuracy in simple tasks can be achieved with very crude signal processing. It also shows that suitable incorporation of some classical pattern recognition techniques (distributed representation of network output with rows of Hadamard's matrix and optimised quantisation of input) can provide significant improvement in the system performance.

INTRODUCTION

Automatic recognition of spoken words is one of the important domains used for the application and testing of artificial neural networks. These provide the ability for a simple parallel hardware implementation of a speech recognition system. Although a number of successful experiments have been reported in the literature (e.g. Gold & Lipmann, 1988; Lang, Waibel & Hinton, 1990), it is clear that a lot more work has to be done before artificial neural networks can become practical in speech recognition. In particular, in this paper we intend to show that classical pattern recognition techniques (coding theory and vector quantisation, in our case) could possibly, if incorporated appropriately, offer a worthwhile enhancement to the performance neural networks.

In a series of introductory experiments conducted at Telecom Research Laboratories recently, the particularly simple structure of multi-layer perceptrons with mask hidden units was used for the task of isolated word recognition. We should stress that the prime objective at that stage was to test and refine a particular neural network algorithm, to eventually build a reliable speech recognition system. Thus, the signal processing element was kept very simple (and did not include any dynamic information of the speech waveform, in particular), yet still able to provide interesting results.

In this paper we are especially interested in mask perceptrons which are virtually a kind of single slab high order network or classical polynomial classifiers (Duda & Hart, 1969). Our special interest in mask perceptrons stems from their potential for simple and efficient implementation in digital hardware (Kowalczyk, Aumann & Cybulski, 1991) and the existence of relatively efficient generation techniques (c.f. Cybulski, Ferrá,Kowalczyk & Szymanski, 1989).

In order to simplify hardware implementation, mask perceptrons assume quantised inputs, necessitating the resolution of the discretisation when dealing with continuous inputs, usually by employing classical vector quantisation results and information theory (c.f. IEEE Trans. Inf. Theory 28 (2), 1982; Kowalczyk & Szymanski 1989). On the other hand the use of distributed output encoding (spread spectrum techniques in transmission theory) could simplify in some cases the network structure by reducing the required number of outputs for a given problem. Additionally the network may utilise the error correcting capabilities of the code to improve network performance in the presence of noise (Chiueh &Goodman, 1988).

NEURAL NETWORK BACKGROUND

From the early days of machine learning research, the potential advantages of polynomial classifiers (high order networks) for pattern recognition was recognised (c.f Duda & Hart, 1973). However, the high order networks were, in practise, not used much, mainly due to lack of efficient techniques to cope with the combinatorial explosion in the number of high order terms. In the current wave of neural network activities, a number of researchers have turned their attention to high order networks again, motivated by a variety of reasons including an increase in memory capacity, the capability of high order terms to embed prior knowledge about the properties of a domain of interest like shift and scale invariance (Giles & Maxwell, 1988), the optical implementation of associative memory (Psaltis, Park & Hong, 1988), the universal capability of the structure to implement any predicate (Minsky & Papert, 1969). Our personal interests are additionally fueled by the recent availability of efficient training procedures based on empirical selection or even generation of useful terms and the relative simplicity of possible VLSI implementations, especially in the form of modulo perceptrons which are natural adaptations of single slab high order networks to the case of a limited set of available weights (Cybulski, et al., 1989, Kowalczyk, Aumann & Cybulski, 1990).

NETWORK FOR ISOLATED WORD RECOGNITION

The neural networks considered in this paper are adjusted to the specific domain used in experiments: isolated word recognition. The specific architecture is shown in Figure 1. The heart of the system is a mask perceptron, i.e. a single slab high order network (Giles & Maxwell, 1987) with binary inputs. It consists of three layers: a layer of "input quantisers" (**IQ**'s) converting continuous outputs of frequency filters to a string of bits, $x_1, x_2,...,x_n$, a hidden layer of logical conjunction units (higher order monomials, $x_{j_1} x_{j_2} \cdots x_{j_k}$) and the output layer (connected via links with real weights, $w_{i(j_1,...,j_k)}$, to some units of the previous two layers for the purpose of "normal summation" of the weighted activations of these linked units). Mathematically, the mask perceptron implements a set of polynomials:

$$y_i(x_1, x_2,...,x_n) = w_i + \Sigma_{(j_1,...,j_k)} w_{i(j_1,...,j_k)} x_{j_1} x_{j_2} \cdots x_{j_k} \qquad \text{(for } i = 1,..., m \text{ and } x_1, x_2,...,x_n \in \{0, 1\}).$$

The selection of weights, hidden units and quantisers for the mask perceptron is within the domain of suitable training.

A final processing stage is added on top of the mask perceptron in one of two ways. The first of these is known as centralised encoding. In this case the number of mask perceptron output units is equal to the number of words, with each output unit ideally assigned 1 for the word to which it corresponds and -1 for all others. In the second case, known as distributed encoding, each input pattern is assigned a row from a Hadamard matrix (c.f. Chieueh & Goodman, 1988 and MacWilliams and Sloane, 1977) whose size determines the number of output units. These patterns of ±1's have maximal Hamming distance between each other, so that for two patterns to be confused at least one of them must have a large number of errors.

The distributed encoding of the output units, if implemented, requires a final stage appearing as a "Hamming net" as well as a prior threshold level. In initial experiments a number of different perceptron output non-linearities (**ONL**'s) for the threshold level prior to the "Hamming net" were tried. These were functions as follows: **(A)** a single step $f(t) = sgn(t)$, **(B)** a two step $f(t) = sgn(t)$ for $|t| > \varepsilon$ and $f(t) = 0$ for $|t| \leq \varepsilon$ and **(C)** ramp $f(t) = sgn(t)$ for $|t| > 1$ and $f(t) = t$ for $|t| \leq 1$.The ramp function consistently showed the best performance and hence was utilised in further experiments.The "Hamming net" is used to calculate the cross-correlation of the threshold-modified perceptron results with the ideal code patterns, designated by the Hadamard matrix, for each word. The weights between "non-linearities" and "correlation" units in Fig. 1 are equal to ±1's corresponding to the designated distributed patterns.

Regardless of whether centralised or distributed encoding is used we employ a "top N selection" layer to order the activations of the previous layer by value and select the N most active outputs as the final result for a given value of N.

The speech samples consist of 20 positive real numbers each in the range 0 - 327. The IQ layer is employed to achieve a smaller range of discrete inputs. In the selection of quantisers, the problem is to choose a small number of useful thresholds from a large number of possible candidates. Two particular approaches were used here. The first being to partition the range of each filter output into uniform segments. Whilst the second technique involves the use of a generalisation of relative entropy for partitions, to the case of overlapping coverings (Kowalczyk & Szymanski, 1989).The choice of n_i thresholds for the i-th filter is made sequentially, with the objective of minimising the relative entropy for each step.

The task of training the mask perceptron involves the selection of hidden units (monomial terms) and then

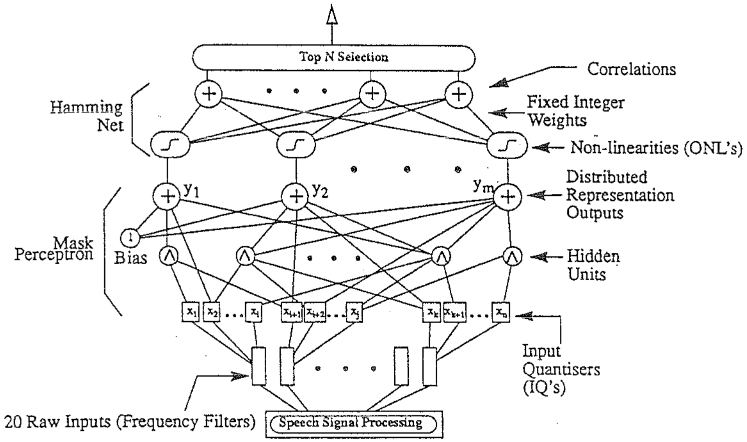


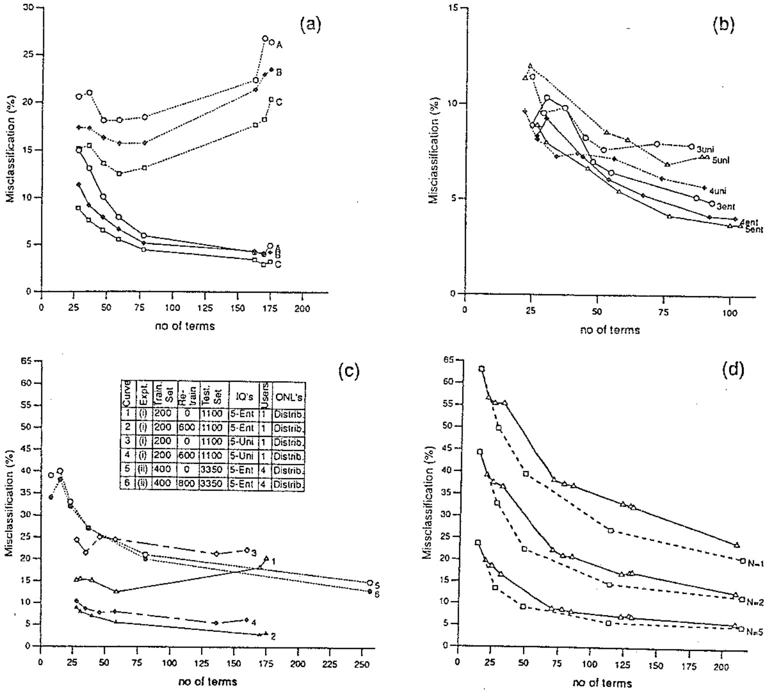
Figure1. Architecture of isolated word speech recognisor used in experiments. The "Hamming net" layer is optional and used when the distributed encoding method is implemented.

of weights. The monomial terms are selected empirically, at each stage reducing the potential mean-square-error between the desired distributed or centralised representation vector and the actual mask perceptron output for a given training set. Candidate terms are produced using a set of simple heuristics that recursively build up terms from those previously selected. Having selected a final set of monomial terms, the weights are computed using the Penrose-Moore pseudo-inverse technique. In some cases retraining was applied, meaning that the weights were recalculated for a different set of instances from the set used for term selection.

EXPERIMENTAL RESULT

To illustrate the performance of the techniques, we present in Fig. 2 results of certain experiments. They cover three different types of experiments in recognition of spoken words: (I) 10 digits by a single speaker (trained for the first 20 rep./word, and optionally retrained for an extra 40 rep./word with a test against the full set of 110 rep./ word providing 1100 instances), (ii) 10 digits spoken by four speakers (3 males and 1 female; 110 + 80 + 75 +70 = 335 rep./word giving 3350 instances; trained for the first 10 rep./word/speaker, and optionally retrained on an extra 10 rep./word/speaker with a test against the full set) and (iii) 60 words by a single speaker (40 reptn./word producing 2400 instances used in testing of which 1200 = 60×20 rep./ word was used for training). The words were spoken in an office environment. The samples were limited to 500 ms duration, band limited between 300 Hz and 3300 Hz, sampled at 8 kHz and digitised to 16 bit accuracy. The digitised information was then passed through a bank of 20 audio filters emulated by a DSP hardware board. The filter outputs were then stored and used for training on a SUN sparc workstation, in a batch type mode. In these experiments we generated the networks in a series, where each mask perceptron was constructed typically by appending new terms to those previously selected. The selection process was controlled by the adjustment of some parameters, so, in particular the network size was set automatically (c.f. Figure 2). For distributed encoding we used: ten rows of a 16x16 Hadamard matrix in experiments of types (i) and (ii), and 60 rows of a 64x64 Hadamard matrix in experiments of type (iii). In each experiment requiring ONL's (apart from Figure 2a) those of type (C) were used. (Note that for centralised encoding no ONL's are required.)

The six curves in Figure 2a represent the performance of one series of mask-perceptrons using 5 entropy



Curve	Expt.	Train. Set	Re-train	Test. Set	IQ's	Users	ONL's
1	(i)	200	0	1100	5-Ent	1	Distrib.
2	(i)	200	600	1100	5-Ent	1	Distrib.
3	(i)	200	0	1100	5-Uni	1	Distrib.
4	(i)	200	600	1100	5-Uni	1	Distrib.
5	(ii)	400	0	3350	5-Ent	4	Distrib.
6	(ii)	400	800	3350	5-Ent	4	Distrib.

Figure 2. Result of experiments in recognition of 10 digits by a single speaker, (a) and (b), by one and four speakers (c) and of 60 words by a single speaker (d).

selected IQ's, trained for experiment type (i) using the distributed encoding technique. Different ONL's, (A), (B) with $\varepsilon = 0.2$ and (C) were used highlighting the improvement possible by a correct choice of these nonlinearities. The solid curves display the performance of networks additionally retrained (against 600 instances).

Figure 2b represents the performance of six different series of mask perceptrons, trained and retrained for experiment type (i) using the centralised encoding technique. Different numbers of uniform-and entropy selected QL's have been used, as indicated in the figure.

In Figure 2c we display the performance of three different series of neural networks (three pairs of trained-only/ trained-retrained curves: 1 / 2, 3 / 4 and 5 / 6) developed for experiments of types (i) and (ii). We used here different sets of 5 QL's for each series, as indicated in the table.

Figure 2d shows results for two series developed for experiments of type (iii). We have allowed several error counting methods here, namely a misclassification occurs if the correct alternative is not included in the "top N selection" for N = 1, 2 and 5 respectively. Solid lines are for a series using centralised encoding and for the series using distributed encoding we use broken lines. In each case the same 15 uniform QL's were used.

DISCUSSION OF RESULTS

In Fig.2a we observe that the ramp non-linearity (C) produces consistently the least misclassification. This result could be attributed to the observation that (C) preserves the largest "amount of information" contained in the mask perceptron outputs. However, when in some experiments we passed unchanged perceptron outputs to the Hamming net, the results were much worse than with any of these non-linearities; thus some non-linearity is needed, although the optimal shape is still to be determined.

Figures 2b and 2c show that training with carefully selected IQ's gives an improvement of the order of 20-50% over the uniform assignment. Notice that initially for a small number of IQ's the benefit of the entropy thresholds is not apparent, but that as the number of accepted monomials increases, the curves with entropy selected thresholds converge to a smaller misclassification than their corresponding uniform counterparts. We must point out that these curves have been obtained using retraining on the mask perceptron, and while the same conclusion can generally be made for the case of no retraining, we did encounter a case in which the opposite was true. More work needs to be done to resolve this problem and other selection techniques such as *k*-means clustering etc. should be tried.

We observe that retraining typically improves the performance of the neural network, in some simple cases even significantly (c.f. curves for single speaker in Figures 2a and 2c). However, in the case of 4 speakers (Fig. 2c) the improvement after retraining was negligible. This is perhaps due to some improved term generation heuristics that were used in this case, a relatively good level of generalisation was achieved during training. Comparison with the other two curves in this figure obtained for the much simpler task of recognition for a single user using less efficient term generation heuristics seems to support such an explanation. This aspect obviously requires further investigation.

Referring to Fig. 2d there are several interesting points to be made. Firstly, it is apparent that the misclassification using the distributed encoding technique is always less than or equal to that of the centralised encoding technique independently of the number of mask perceptron monomials used. This can be explained in the following way: the task of the mask perceptron is to approximate as closely as possible the representation string of the word which the input instance denotes, in many cases this will not be done exactly leading to several possible candidate representation strings being alternatives to the correct one intended. The task of the "Hamming net" is to correlate this error infected pattern with each true representation string. The greater the difference between these true representation strings the greater the chance of correlation with only one of the true strings. Thus the Hamming distance between pairs of distributed encoded strings is 32 bits in our case, contrasting the Hamming distance between pairs of centrally encoded strings being only 2. This also explains why the distributed encoding curves appear to converge to a minimum faster than their corresponding centralised encoding counter-parts. We also observe that, as the number of terms increases the larger mask perceptron becomes more accurate in its own classification task and so less is left for the Hamming net to do.

It is worth stressing that the signal processing used in these experiments was extremely simple and yet, overall, quite satisfactory accuracies were achieved (especially good in single user digit recognition). This strongly suggests that the combination of a mask-perceptron classifier with a more sophisticated feature extraction system could lead to significantly better recognition rates. As a rough guide, Lang, Waibel & Hinton (1990) use in their isolated word recognition experiments, discrimination between B, D, E and V, a carefully preprocessed 192 number spectrogram; optimal speaker independent classifier of Italian digits over telephone line, reported by Gemelo & Mala (1990) was using 240 cepstral and energy coefficients. This is obviously much more sophisticated than our 20 raw filter outputs, however we do not intend true comparison to be made here since the comparison of efficiency of different network training algorithms is dubious, unless the same data and preprocessing is used.

CONCLUSIONS

Results show the usefulness and suitability of mask perceptrons in the area of speech recognition. Such systems should be tested with different, more sophisticated feature extraction techniques, on a variety of tasks, including different aspects of isolated and continuous speech recognition, integration with speech understanding systems etc., since the structure is capable of simultaneous processing of both numerical and categorical attributes on an equal footing.

Experiments show that merely using a more efficient output representation (a distributed encoding) can in some cases significantly improve the generalisation ability of neural network classifiers. Another advantage of distributing outputs is multiplexing. Using a set of pseudo-orthogonal vectors with significant mutual Hamming distances (e.g Gold sequences; Gold 1967) instead of orthogonal Hadamard matrix rows, one can represent efficiently multiple classes of interest by a relatively small number of network outputs. Current results show that this is feasible.

A careful choice of input quantisers was found to produce significantly better results: in some cases 25-50% drop in misclassification. This warrants more effort in this direction, especially including the testing of different statistical quantisation methods in order to determine the most appropriate method for a given problem.

Distributed encoding and a careful choice of input quantisers, proved to enhance perfomance of our networks, readily lend themselves to application with other feed forward networks (e.g. back-propagation) and results for this should be obtained in the near future.

ACKNOWLEGEMENTS

We acknowledge the use of a speech digitisation system developed by Voytech Systems Inc., Regina, Canada for these experiments. The permission of Executive General Manager, Telecom Australia Research Laboratories to publish this paper is also acknowledged.

REFERENCES

Chiueh, T. and Goodman R. (1988) *A Neural Network Classifier Based on Coding Theory,* in *Neural Information Processing Systems,* Anderson, D. Z. (ed.), American Institute of Physics, New York, 1988.

Cybulski, J. L., Ferrá, H. L., Kowalczyk, A. and Szymanski, J. (1989) *Experiments with multi-layer perceptrons,* in *Proc. of the Australian Joint Artificial Intell. Conf.* (AI'89, Melbourne).

Duda, R. O., Hart, P. E. (1973) *Pattern classification and scene analysis* (John Wiley & Sons, New York).

Gemalo, R., Mana, F. (1990) *A neural approach to speaker independent isolated word recognition in an uncontrolled environment,* in *The Proceedings of INNC,* Paris 1990.

Giles, C. L., Maxwell, T. (1987) *Learning, invariance, and generalization in high-order neural networks,* Applied Optics 26, 4972-4978.

Gold, R. (1968) *Maximal Recursive Sequences with 3-Valued Recursive Cross-Correlation Functions,* IEEE Trans. Inf. Theory 14, 154-156.

Gold, B. and Lipmann, R. P. (1987) *Neural-Net Classifiers Useful for Speech Recognition,* in *Proceedings of the IEEE First International Conference on Neural Networks.*

IEEE Trans. Inf. Theory 28 (2) (1982) (A special issue on quantisation etc.).

Kowalczyk, A. and Szymanski J. (1989) *Rough Simplifications of Decision Tables,* in *Computing and Information,* Janicki, R. and Koczkodaj, W. W. (eds.) (North-Holland: Amsterdam)

Kowalczyk, A., Aumann, G. and Cybulski, J. (1990) *A Simple Architecture of High Order Network with Positive Bounded Integer Weights,* submitted.

Lang, K. J., Waibel, A. H., Hinton, G. E. (1990) *A Time-Delay Network Architecture for Isolated Word Recognition,* Neural Networks 3, 23-44.

MacWilliams, F. J. and Sloane, N. J. A. (1977) *The Theory of Error Correcting Codes* (North-Holland).

Minsky, M. and Papert, S. (1969) *Perceptrons,* (MIT Press, Cambridge: Massachusetts; edition 1988).

Psaltis, D., Park, C. H., Hong, J. (1988) *Higher Order Associative Memories and Their Optical Implementations,* Neural Networks1, 149-163.