SPEAKER IDENTIFICATION WITH PROJECTION NETWORKS

P. Castellano and S. Sridharan

Signal Processing Research Centre
School of Electrical and Electronic Systems Engineering
Queensland University of Technology

ABSTRACT - This study compares four connectionist approaches to text-independent Automatic Speaker Identification. It concludes that projection networks such as the Higher Order Neural Network (HONN), the Moody-Darken Radial Basis Function (MD-RBF) network and the Logicon Projection Network (LPN) consistently outperform standard Multi-Layered Perceptron (MLP) networks. The difference in performance is at least 11% according to the criteria of ASI threshold and percentage of correctly recognised speech vectors. In this study, the LPN, capable of creating both open and closed boundary regions for data points, is superior to both the HONN and the MD-RBF on both criteria (Mean threshold: 90.55%, mean percentage of correct classifications: 97.2%). Since the LPN's dominance is almost universal across all speakers considered, results need not be confirmed by the additional use of one, or several, of the remaining three classifiers.

INTRODUCTION

The selection of efficient classifiers is important for Automatic Speaker Identification (ASI). These should exploit inter-speaker differences and minimise loss of information contained in parametric representations of speech. Artificial Neural Networks (ANNs) are increasingly being applied to the ASI problem. Experiments conducted by Oglesby and Mason (1991) have shown that RBF ANNs outperform an MLP in this role.

This paper introduces the Logicon Projection Network to ASI. It compares its performance to that of the previous two ANNs and to the Higher Order Neural Network.

BACKGROUND

ASI is a three step process consisting of feature extraction, pattern matching and adjudication. In the feature extraction step, a parameterisation technique is used to produce vectors of acoustic features from the speech signal. Pattern matching is accomplished by training and then testing a supervised feed-forward ANN with distinct acoustic vector databases. The present work has adopted the binary pair classification method advocated by Rudasi and Zahorian (1991). In their approach and for N speakers, N*(N-1)/2 small ANNs are trained, each to separate two of the N speakers. Each of these ANNs are independent of the others, as well as of the training data of the non-relevant classes. Since many classifiers need be trained the approach requires more work than if a single classifier was used (Castellano and Sridharan, paper A, 1994). However, in the present case, individual classifiers need not be retrained should more speakers be added to the database. The adjudication step relies both on a consideration of the percentage of correctly identified acoustic vectors, for a given speaker, and on the setting of a classification threshold below which identification is deemed to have failed.

DATA

Reflection coefficients used in this study (Deller *et al.*, 1993). This assumes that the vocal tract may be modelled as a connected set of lossless acoustic tubes. At each junction, part of the speech wave is propagated through and part is reflected back. The coefficient

$$r_i = \frac{A_{i+1} - A_i}{A_{i+1} + A_i} \qquad (1)$$

is the reflection at the ith junction where A_i and A_{i+1} are the cross-sectional areas of two adjoining tubes.

The speech signal was contained in recordings of free conversation from twenty male speakers. It is text-independent. The signal was digitised at 10 kHZ. It consisted of 50 second segments with silent parts removed. The segments were divided into 1.2 second frames which were first high frequency pre-emphasised. This was done with a transfer function of $1\text{-}0.98z^{-1}$. A 256 point Hamming window and an analysis filter of order 15 were than applied. For each speaker, 100 vectors were set aside for ANN training and 100 others for testing.

SYSTEM ARCHITECTURES

STANDARD MLP: Data patterns are circulated forward through a network which, for practical problems, has at least one hidden layer fully connected with one input and output layer. These patterns are not transformed by any projection prior to being processed. The transfer function at each PE is typically a sigmoid expressed as:

$$f(u) = \frac{1}{(1+e^{-u})} \qquad (2)$$

A back-propagation algorithm is used to feed back the error arising from computing the difference between an actual and desired (target) output. Following this, weights are updated. A single hidden layer architecture was retained here. This is because a one hidden layer architecture is optimum for ASI (Oglesby and Mason, 1990).

HONN: This network's architecture differs from that of an MLP in that a higher order (functional-link) layer follows immediately after the input layer. Data vectors are mapped into the resulting higher order space. It has been claimed that a flat HONN with data projection of the type

$$[x_i, x_i x_{i+1}, x_i x_{i+2}, x_i x_{i+1} x_{i+2}] \text{ (tensor)} \qquad (3)$$

for each term x_i in the input vector, is an efficient classifier for ASI (Castellano and Sridharan, paper B, 1994). This projection will be retained for this study. HONN enables faster training times than standard MLPs and eliminates the occurrence of local minima during gradient descent.

RBF: The network consists of radially symmetric hidden PEs. A cluster of training vectors is associated with each hidden PE. Each cluster has a centre which is a vector in the input space and a distance measure to determine how far an input vector is from the centre. The Moody-Darken (MD) version of this network (with one hidden layer) has been found to outperform the MLP in Automatic Speaker Verification which is closely related to ASI (Oglesby and Mason, 1991). The MD-RBF's activation function is given by

$$l_i = \sqrt{\sum_{j=1}^{N} (X_i - y_{ij})^2} \qquad (4)$$

where y_{ij} is given by K-means clustering for inputs X_i. The transform function (Gaussian) is :

$$T_i = e^{\frac{l_i - \|X\|^2}{a_i^2}} \qquad (5)$$

where the a_is are given by a nearest neighbour heuristic and govern the receptive "width" of each cluster in the hidden layer. Input vectors must be transformed so that they are of constant norm. To accomplish this, the MD-RBF projects these vectors to a hyper sphere of radius $\|X\|$ (NeuralWare Inc., 1993). This is done in a space of one higher dimension than the input space. This extra

dimension allows the preservation of information concerning each input's magnitude. MD-RBF is an efficient classifier when ample training data is available to provide a good estimate of cluster centers and widths and where data density approximates a Gaussian.

LPN: This network will be tested here, for the first time, in the context of ASI. With LPN, as in the RBF case, data vectors are projected onto a sphere of one higher dimension than the vectors (NeuralWare Inc., 1993). Unlike RBF, the LPN algorithm deals both with closed decision regions (hyper spheres or hyper ellipses) and open decision regions (hyper planes). This leads to a very streamlined network. More than one hidden layer is rarely required. During learning, open regions can become closed and vice versa. The hyper planes which partition the input space intersect the sphere, creating hyper spherical decision boundaries. This produces an approximate solution for weights and thresholds which can be very close to the network's global minimum error. The projection may be expressed as (NeuralWare Inc., 1993):

$$Y = |Y|\left(\frac{2X/R_0}{1+(|X|/R_0)^2}\ ,\ \frac{1-(|X|/R_0)^2}{1+(|X|/R_0)^2}\right) \qquad (6)$$

where
X is an N dimensional input vector,
Y is an (N+1) dimensional projection vector,
R_0 is the inner radius of the sphere onto which the Xs are first projected from the input space.
|y| is the radius of a separate Projection sphere, centered on the previous sphere, which scales the input vectors and $R_0 << R$.

The magnitude of the weight vectors on the connections between the projected X's and any one node in the hidden layer is |Y|. The transformed inputs Y become inputs to a back propagation network with an additional PE in the input layer. Gradient descent refines the initial solution. This step is typically accomplished without the plateaus which lengthen standard back propagation and local minima which seriously undermine that algorithm's effectiveness (Wilensky and Manukian, 1992).

EXPERIMENTAL STUDIES AND RESULTS

A preliminary study was conducted to select ANN architectures for optimum ASI performance, given the binary classification approach chosen. Results are shown in Table 1.

Table 1. Best architecture for four ASI classifiers.

	MLP	HONN	MD-RBF	LPN
Learning rule:	Delta-rule	Delta-rule	Delta-rule	Delta-rule
Learning rate (fixed):	0.1	0.1	0.1	0.1
Momentum (fixed):	0.9	0.9	0.9	0.9
Epoch size:	16	1 (no option)	16	20
Training iterations:	4500 (approx)	3500	6000	6000
Hidden layer size:	20	0	50	45
Projection:	N/A	2nd together with 3rd order tensor	Moody-Darken	Hyper-spherical and hyper-planar

Training times:

The training of the MLP was plagued by local minima on the error's gradient descent surface. It was therefore not possible to determine an ideal number of training iterations for the binary classification problem. Hence, for each speaker pair, the MLP required two training phases. The first phase was conducted solely to determine the best number of training iterations for the given pair. The network was than re-initialised and that number governed the length of the second training phase. The procedure added both complication and time to the training step. The duration of one iteration is shorter for a flat HONN than for MLP. That a HONN is faster to train than an MLP is well documented (Hush and Salas, 1989). MD-RBF and LPN networks were also faster to train than the MLP (typically less than 20 seconds given a SPARCstation10 platform.)

Classification performance:

Each speaker was tested against each of the remaining nineteen, in turn. For each speaker, a mean percentage of correctly classified speech vectors, as well as a mean threshold, were computed. Irrespective of which ANN was investigated, all thresholds were greater than 50 per cent. (Thresholds were chosen to be the lowest individual speaker scores.) Hence no speaker was erroneously classified. The performance of any one ANN, on a speaker by speaker basis, was consistent. Hence, for a given ANN, mean thresholds and percentages of correct classifications, computed over all input vectors and speakers, could be considered good criteria of ASI performance. These criteria were computed for all classifiers, as is illustrated in Table 2.

Table 2. Mean thresholds and percentages of correct classifications (over all speakers and data).

	MLP	HONN	MD-RBF	LPN
Mean percentage of correct classifications	82.8	93.9	95	97.2
Mean threshold (%)	71	83.65	85.45	90.55

From Table 2, it is apparent that HONN, MD-RBF and LPN outperformed the MLP by more than 11 per cent, irrespective of which criterion is considered. The former three are projection networks (see Table 1). Figures 1 to 3 illustrate the difference in performance between these ANNs, on a speaker by speaker basis.

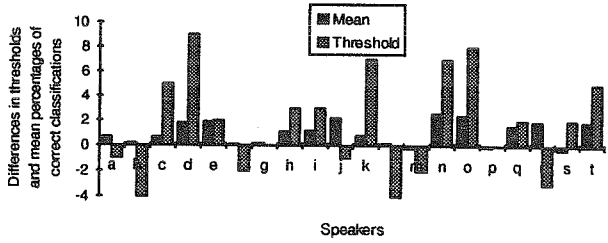

Figure 1. Differences between RBF and HONN threshold and mean percentages of correct classifications for speakers a to t.

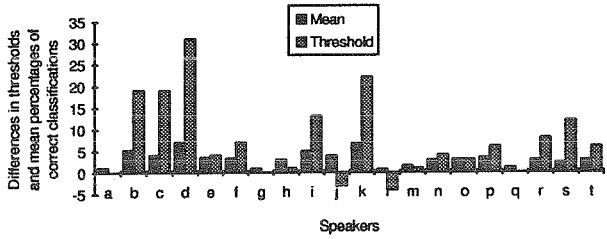

Figure 2. Differences between LPN and HONN threshold and mean percentages of correct classifications for speakers a to t.

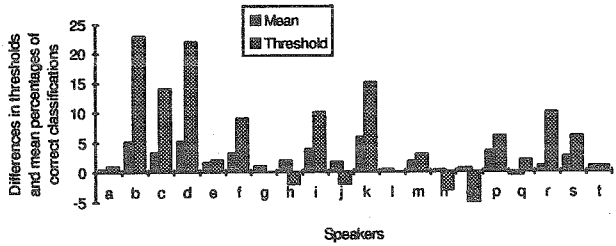

Figure 3. Differences between LPN and HONN threshold and mean percentages of correct classifications for speakers a to t.

The discriminatory capability of a HONN is generally greater than that required for the task at hand. Indeed, a single HONN PE, with second order links, is able to provide solutions to problems described by any second order equation (e. g. hyperboloid or hyper ellipsoid) (Hush and Salas, 1989). Because of this, HONN may require more data vectors than other ANNs and will not necessarily provide the best classification performance. However, the MLP requires the most amount of data to produce quality solutions (Hush and Salas, 1989). Another shortcoming affecting MLP methods is that their performance is relatively poor when hidden layers are small (Hush and Salas, 1989) which is a requirement for binary classification.

HONN's ability to partially remove invariance in data accounts for its superiority over the MLP in the present case. While the MD-RBF outperformed HONN, this was not necessarily the case in terms of individual identification thresholds.(An explanation lies in the clustered nature of acoustic parameter distributions with corresponding densities only very roughly Gaussian in nature.) The relatively high threshold results provided by the HONN have been reported elsewhere (Castellano and Sridharan, paper B, 1994). It may be advantageous to use both ANNs to arrive at a solution (for the purpose of confirming results). The receptive width of clusters in the RBF provided it with a classification advantage over the previous two networks. Widths are slightly greater than adjoining cluster centers providing a smooth fit to the problem space. This contrasts with pure back-propagation type networks where the problem space contains regions with not training data, leading to extrapolation errors (Leonard and Kramer, 1991).

In the present study, the LPN's early discrimination strategy, based on dividing the projection space into open and closed boundary regions, provided the most accurate classification (verified both for ASI scores and thresholds (Figures 2 and 3). This, together with results from Table 2, suggests that an optimum connectionist solution for ASI may be obtainable, using LPN as sole classification method.

CONCLUSION

While the nature of acoustic parameter sets has a strong influence on Automatic Speaker Identification outcome, the classification tool employed must also be carefully considered. This study compared the performance of four connectionist classifiers in the context of text independent ASI. It found that an MLP performed consistently worst than newer projection type networks such as the Higher Order Neural Network, the Moody-Darken Radial Basis Function Network and the Logicon Projection Network. The latter ANNs were able to map input into different problem spaces were superior discriminant solutions were arrived at faster than in the initial space. Their speaker identification thresholds (criterion one) and mean percentages of correct classifications (criterion two) were both at least 11 per cent superior to those of the MLP. This suggests that some form of data projection should precede the main ANN classification stage in ASI.

The MD-RBF was superior to the HONN, across most speakers, according to criterion one. This dominance was less apparent in relation to the second criterion so that ASI classification may benefit from a joint use of these networks. This is true particularly since acoustic parameter densities are very rough Gaussian approximations.

The ability to create both open and closed boundary regions for data points gave the LPN an edge in classification performance over both the HONN and MD-RBF. The edge was 2.2 percent according to criterion one and 5.1 percent according to the second. Hence, the use of an LPN, as sole classifier, may the best solution for connectionist ASI.

ACKNOWLEDGMENT

This research was supported by a grant from the National Institute of Forensic Science.

REFERENCES

Castellano, P. and Sridharan S. (A) (1994) "A Two Stage Fuzzy Decision Classifier for Speaker Identification", Proc. Fifth Australian Conference on Speech Science and Technology.

Castellano, P. and Sridharan, S. (B) (1994) "Text-Independent Speaker Identification with Functional-link Neural Networks", Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, 111-114.

Deller, Jr. J. R., Proakis, J. G. and Hansen, J. H. L. (1993) Discrete-Time Processing of Speech Signals, p. 301, (Macmillan Publishing Co., New-York).

Hush, D. R. and Salas, J. M. (1989) "Classification with Neural Networks: A Comparison", Proc. ISE '89 - 11th Annual Symposium, Economic Growth Through Science and Engineering, 107-114.

Leonard, J. A. and M. A. Kramer (1991) "Radial Basis Function Networks for Classifying Process Faults", IEEE Control System Magazine, Vol. 11, 31-38.

Oglesby, J. and Mason, J. S. (1990) "Optimisation of Neural Models for Speaker Identification", Proc. ICASSP, Vol 1, 261-264.

Oglesby, J. and Mason, J. S. (1991) "Radial Basis Function Networks for Speaker Recognition", Proc. International Conference on Acoustics, Speech and Signal Processing, Vol 1, 393-396.

Rudasi, L. and Zahorian, S. A. (1991) "Text-Independent Talker Identification with Neural Networks", Proc. International Conference on Acoustics, Speech and Signal Processing, Vol 1, 389-392.

Wilensky G. D. and Manukian, N. (1992) "The Projection Neural Network", International Joint Conference on Neural Networks, Vol 2, 358-367.

(1993) Neural Computing - A Technology Handbook for Professional II/Plus and NeuralWorks Explorer, pp. 209-226 and 265-275, (NeuralWare Inc, Technical Publications Group).