Comparison of Multilayer and Radial Basis Function Neural Networks for Text-Dependent Speaker Recognition

R.A. Finan*, A.T. Sapeluk* and R.I. Damper†

*School of Engineering,
University of Abertay Dundee,
Scotland DD1 1HG.
Emails: {r.a.finan|a.t.sapeluk}@tay.ac.uk

†Department of Electronics and Computer Science,
University of Southampton,
England SO17 1BJ.
Email: rid@ecs.soton.ac.uk

ABSTRACT

This paper compares the use of multilayer perceptrons (MLPs) trained on back-propagation and radial basis function (RBF) neural networks for the task of text-dependent speaker recognition. 10 classifier networks were generated for each of 20 male speakers using randomly-generated training sets consisting of 6 true speaker utterances and 19 false speaker utterances (one from each of the false speakers). The resulting networks were then used to assess verification and identification performance for each of the network architectures. The results clearly indicate that the choice of true and false speaker utterances used in the training set has a crucial effect on the success of the classifier. The overall superiority of performance reported in general for RBF networks over MLPs would appear to be due to the reduced sensitivity of the former to a poor training set when compared to the performance of an MLP for the same training set. When both networks are presented with their 'best' training sets, however, the RBF network still significantly out-performs the MLP.

Keywords – speaker recognition, text-dependent recognition, multilayer perceptrons, back-propagation, radial basis functions.

1. Introduction

For many years now, there has been a growing interest in the use of voice as a means of recognising or confirming a person's identity. The reason for this is that a person's voice is considered a biometric identifier, as are fingerprints, retinal patterns and DNA. It is a characteristic that is supposed to be intrinsic and unique to a person and, as such, should not be reproducible by anyone else. Furthermore, it benefits from the fact that the person to be identified does not have to carry a card or a key that can be duplicated or stolen. Also a biometric identifier does not have to be remembered like the personal identification number (PIN) for an automatic teller machine (ATM) card.

In text-dependent speaker recognition, it is assumed that the speaker is cooperative, and wishes to be recognised. This is most often the case in security applications where a person may identify themselves using their voice to gain restricted access to premises or sensitive information. Some common examples of security applications are voice-activated locks, access to restricted computer data and voice verification for telephone-banking and ATM transactions. This contrasts sharply with text-independent speaker recognition where there are no constraints on the speaker's vocabulary. Since text-dependent speaker recognition models the speaker for the nominated text (e.g. their password) only, it has, in general, lower error rates than text-independent speaker recognition which must model the speaker's characteristics for a variety of speech sounds [1, 4].

Although text-dependent speaker recognition would appear to be a straightforward task, it has yet to be realised on a practical everyday level. The main reason for this is that the prime purpose of speech is to convey a message.

Therefore, it is the message that is the most important (albeit not the only) information in a speech signal. The speech signal not only carries the intended message of the speaker, but also implicit information concerning their identity, the language they are speaking and their accent, as well as their emotional and physical state. As these aspects are secondary to the message being conveyed, it is difficult to extract them from the speech waveform: the message and the speaker's characteristics are non-linearly and inter-dependently encoded in the speech waveform. As yet it is impossible to extract the characteristics that determine a person's voice from their speech waveform with total reliability.

Text-dependent speaker recognition can be divided into two categories: verification and identification. In speaker verification, the object is to confirm a person's identity using their voice. This would be the case when someone uses a card or access code that they alone should possess, and they are asked to confirm their identity by using a special password. It is a true or false scenario because there are only two possible outcomes: either it is the supposed speaker or an impostor. In speaker identification, the task is to identify the speaker as one of a group of N possible speakers, in which case there are N possible outcomes. As N increases, the likelihood of making a false identification also increases. For this reason, speaker identification for a large population is generally considered to be more difficult than verification. If there is also the possibility that the speaker comes from outside the group, there are $(N+1)$ possible outcomes, further increasing the chances of a false identification.

Recently, an increasing number of researchers has been examining the applicability of artificial neural networks to both text-independent and text-dependent speaker recognition [2]. Early work was based largely on the multilayer perceptron (MLP) architecture and variants thereof. More recently, however, attention has turned to the use of radial basis function (RBF) networks [13]. The switch to the RBF approach has been driven by reports of superior performance in recognition over MLP networks [14, 5, 11]. In most studies, a single MLP or RBF network was used to represent a speaker. In this paper, we suggest that it is useful to obtain a range of MLP and RBF networks for each speaker, and to assess the way they vary in performance. The variation is determined by the choice of true and false speaker utterances used in the training set. We confirm that RBF networks are more robust in dealing with poor training sets than are MLPs. We further show that if both networks are given their 'best' training set, the RBF network still out-performs the MLP.

2. Neural Networks for Speaker Recognition

Two different neural network architectures were applied to the text-dependent speaker recognition problem: the multilayer perceptron and the radial basis function neural network. Both are described in detail in [10].

The MLP architecture using back-propagation learning [3] is one of the most popular neural networks. It consists of at least three layers of neurons: an input layer, one or more hidden layers and an output layer. The hidden and output layers have a non-linear activation function. Back-propagation is a supervised learning algorithm that uses two passes through the network to calculate the change in network weights. In the forward pass, the weights are fixed and the input vector is propagated through the network to produce an output. An output error is calculated from the difference between actual output and the desired output. This is then propagated backwards through the network, making changes to the weights as required.

The RBF neural network [13] has both a supervised and unsupervised component to its learning. It consists of three layers of neurons – input, hidden and output. The hidden layer neurons represent a series of *centres* in the input data space. Each of these centres has an activation function, typically Gaussian. The activation depends on the distance between the presented input vector and the centre. The further the vector is from the centre, the lower is the activation and *vice versa*. The generation of the centres and their widths is done using an unsupervised k-means clustering algorithm. The centres and widths created by this algorithm then form the weights and biases of the hidden layer, which remain unchanged once the clustering has been done. The output layer (which has non-linear activations) is trained by back-propagation.

3. Speaker Data

The speaker database was formed by 20 male speakers with a common (East of Scotland) accent saying the word "Allenwood" 20 times. The utterances were recorded over two sessions in order to incorporate some time variance. The speech was recorded with a 16-bit A/D card at a sampling rate of 16 kHz, with a high-order low-pass filter with 8 kHz cut-off frequency to prevent aliasing. The recordings were made in ambient background noise conditions typical of a quiet computer laboratory. Each utterance was end-point detected by hand.

For presentation to the neural networks, a series of linear prediction coefficients was generated for each of the 400 utterances using the autocorrelation method [12]. The frame length was 20 ms (320 samples) using a Hamming window, overlapping by 50%. The order of the linear predictor was 12. Cepstral coefficients were then generated from the linear predictor coefficients [6]. The speech was presented to the networks as a sequence of 4 cepstral vectors, each of length 12. The presentation of 4 cepstral vectors at each instant allowed the networks to incorporate some short-term temporal speech information as well as static information.

4. Experiments

The purpose of the experiments was twofold. The first aspect was to verify that RBF networks did in fact provide consistently better results than an MLP network for text-dependent speaker recognition. The second purpose was to investigate the effect of training-set variation on the performance of the two networks.

Each speaker had their own MLP and RBF network. Each network had 2 output nodes, one indicating the likelihood that the input vector belongs to the true speaker and the other the likelihood that it belongs to an impostor – although only the first of these was actually used in testing. Target values during training were $[+1,-1]$ for a true speaker frame and $[-1,+1]$ for an impostor frame.

To investigate the effect of training-set variation, 10 training sets were created for each speaker, totalling 200 training sets. Each training set consisted of 6 true speaker utterances and 19 false speaker utterances (one from each of the possible impostors). These utterances were chosen randomly for each training set. For the verification tests, this meant that there were 14 true speaker test utterances for each network, and 280 true speaker test utterances for the 20 networks in total. There were 361 impostor test utterances per network and, hence, $361 \times 20 = 7220$ impostor test utterances in total. For the identification tests, there were 14 true speaker test utterances per network and, hence, 280 true speaker tests in total.

The number of training patterns, N_T, used to train each network was typically 1250 (depending upon utterance length). In the case of the MLP, there was a single hidden layer of 64 nodes and a tanh activation function was used. In line with usual practice, the number of hidden nodes, learning rate, momentum etc. were set empirically. The RBF network used $0.25N_T$ hidden nodes. The nearest-neighbour width heuristic used 2 nearest neighbours.

The score for a test utterance for a given network was obtained as follows. Each frame of the utterance was presented to the network and its output (i.e. for the first one of the two output nodes mentioned above) found. The average output value across all frames of the utterance was then computed, and taken to be the score. No use was made in this study of any measure of dispersion, such as the standard deviation, of the scores.

5. Results

The identification and verification errors for each of the two systems were calculated for both randomly-selected and best-performing training sets. In the first case, a network for each of the 20 speakers was randomly picked from the 10 created. These selected networks then formed a group for one set of verification and identification tests. This was done 100 times so that average values could be calculated. In the second test, the best-performing network for each speaker was selected by hand in order to give an idea of the potential performance for each system.

For the verification test, a threshold of 0 was taken. Any score above 0 was deemed to correspond to the true speaker and any below to an impostor. This threshold value is arbitrary and could be altered if deemed necessary, in order to trade the numbers of false acceptances and rejections so as to improve overall performance. A false acceptance occurred when an impostor was recognised as the true speaker and a false rejection occurred when the true speaker was recognised as an impostor. As stated in the previous section, the number of impostor tests for each network was 361 and true speaker tests 14. An average value of the false acceptances and rejections was then calculated for the 20 networks.

The identification test was done by comparing the outputs of all 20 networks for a particular utterance for the 280 speaker identification tests in total. The network with the highest output was considered to belong to the true speaker. No minimum difference between network outputs was imposed.

The results for the randomly-chosen training sets and the best-performing training sets for both RBF and MLP networks are shown in Tables 1 and 2, in which FA indicates the false acceptance rate and FR indicates the false rejection rate. These clearly show that the RBF networks are considerably better than the MLPs for both typical (i.e. trained on randomly-selected sets) and best-performing operation. We note in passing that the identification

	Verification			Identification
	FR (%)	FA (%)	Total Error (%)	Error (%)
MLP	5.89	0.37	0.57	2.06
RBF	5.01	0.11	0.28	1.02

Table 1: Verification and identification results for randomly-selected training sets.

	Verification			Identification
	FR (%)	FA (%)	Total Error (%)	Error (%)
MLP	1.42	0.06	0.11	0.0
RBF	0.71	0.0	0.03	0.0

Table 2: Verification and identification results for best-performing training sets.

performance is apparently better than the verification performance in Table 2. However, the number of tests is different in the two cases, making the significance of this observation uncertain.

The superiority of the RBF networks over MLPs is further borne out by the results presented in Figures 1–4. These show, for each training set, the total number of errors, values of the separation metric d' (see below) and the number of false rejections and false acceptances, respectively. In each case, graph (a) represents the results for the MLP and graph (b) the results for the RBF networks. Each training set is referenced by the index of the x-axis. Every decade represents the results of a single speaker's training sets.

In Figure 1, the total area under (a) is clearly less than that under (b), indicating that the total number of errors for the RBF network is less than that of the MLP. The RBF network had in fact only 219 errors compared to 426 for the MLP. The graphs also show that if the RBF network had trouble separating the speakers then the MLP was in general markedly worse. This confirms the commonly-held viewpoint that RBF networks exhibit better general performance than MLPs for speaker recognition.

In previous work [8, 9], the d' sensitivity index of classical signal detection theory [7] was modified to yield a measure of separability between impostor and true speaker distributions in speaker recognition. It is defined as the difference between the means of the two distributions, normalised by the geometric mean of their standard deviations. A d' of approximately 6 was found to represent good separation in this application, with higher values representing better discrimination. Figure 2 shows the d' values of the true speaker distribution against the impostor distribution, for the two networks. The average value for the RBF networks is 7.92 compared to 6.54 for the MLPs. So the RBF system created a greater gap between the true and false speaker distributions. There is also a high correlation coefficient of 0.82 between the d' values for the RBF and MLP networks. This indicates that the success of both networks is highly dependent on training sets and that, in general, a good training set for the RBF system is also a good training set for the MLPs and *vice versa*.

Figure 3 shows that the area under the curve for the MLP is only slightly larger than that for the RBF network, indicating a similar level of false rejections. The MLP had 162 false rejections while the RBF system had 142. So the RBF network did not succeed in making significant differences to the number of true speakers who were rejected. However, Figure 4 shows that the RBF network significantly reduces the number of false acceptances, with only 77 compared to 264 for the multilayer perceptron. (These remarks should be interpreted in light of the rather arbitrary threshold score of 0, i.e. there was no attempt to trade false acceptances and false rejections to produce lower total errors.)

6. Discussion

Although others have found that RBF networks generally give better results than MLPs for speaker recognition, they may not be using the RBF network to the best of its ability. In general, the RBF network is more resilient against a bad training set than an MLP and, hence, provides better results. However, an RBF system can provide even better

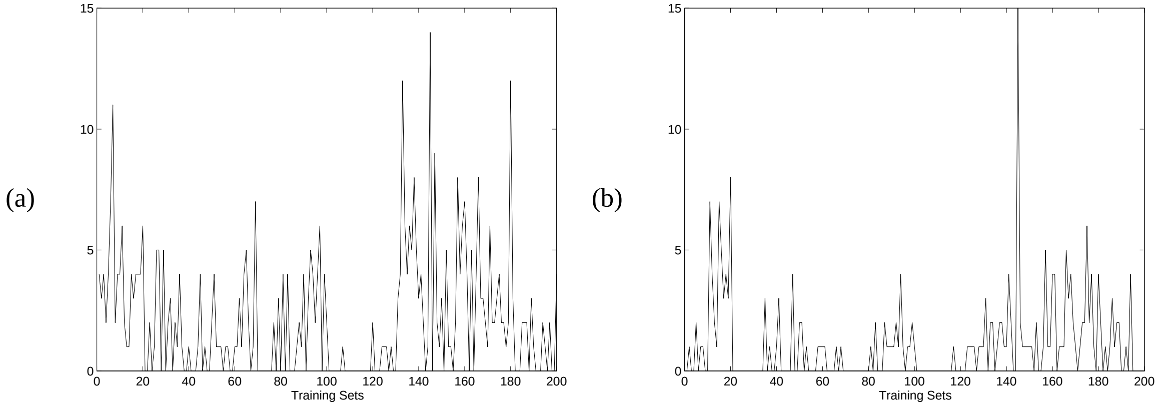

Figure 1: Total number of errors for (a) MLP and (b) RBF network versus training set index. There are 10 training sets for each of the 20 speakers.

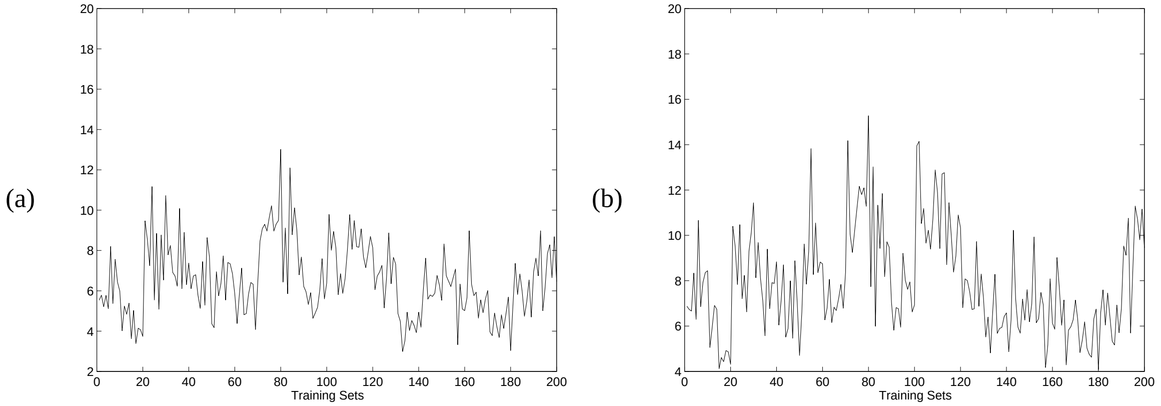

Figure 2: Separation of true and impostor speaker distributions in terms of d' measure for (a) MLPs and (b) RBF networks.

results with a suitable training set. The best means of selecting a suitable training set for speaker recognition remains an unresolved issue: however, the results of our experiments indicate that the difference in performance between the RBF and the MLP networks may in itself form the basis of a measure of the suitability of a training set. For a good training set, a significant improvement would be expected for an RBF network relative to an MLP, whereas a poor training set will not show much improvement.

References

[1] B. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, no. 4, pp. 460-475, 1975.

[2] Y. Bennani and P. Gallinari, "Connectionist approaches for automatic speaker recognition," *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland, pp. 95-102, 1994.

[3] Y. Chauvin and D. Rumelhart (editors), *Backpropagation: Theory, Architectures and Applications*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1995.

[4] G. Doddington, "Speaker recognition – identifying people by their voices," *Proc. IEEE*, vol. 73, no. 11, pp. 1651-1664, 1985.

[5] S. Fredrickson and L. Tarassenko, "Radial basis functions for speaker identification," *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland, pp. 107-110, 1994.

[6] S. Furui, "Cepstral analysis techniques for automatic speaker verification," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-29, pp. 254-272, 1981.

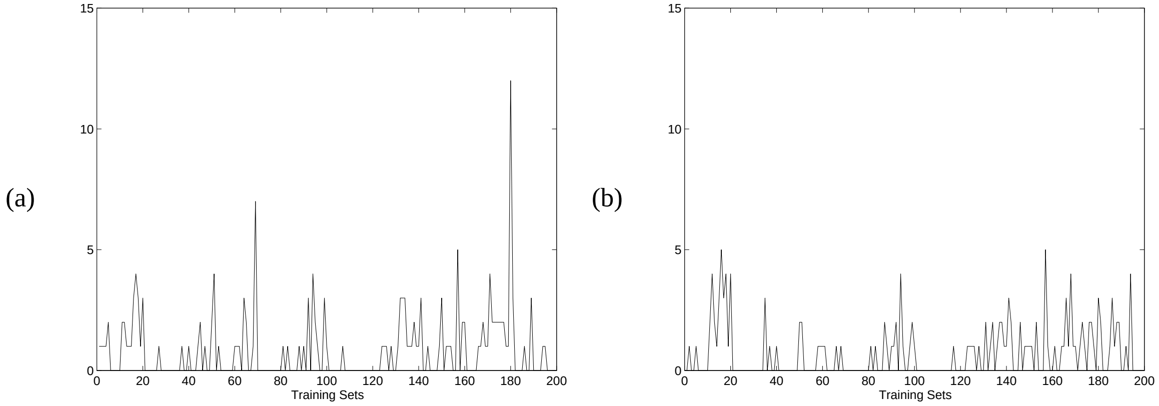

Figure 3: False rejections produced by (a) MLPs and (b) RBF networks.

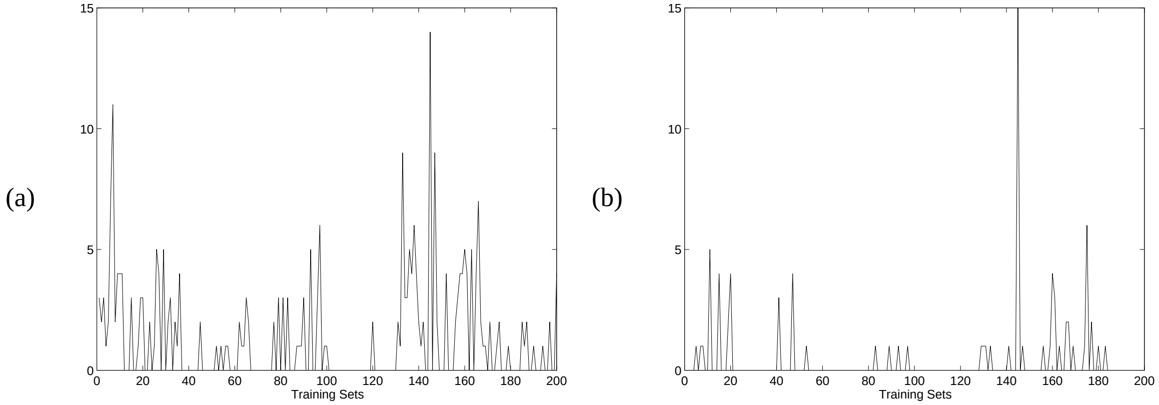

Figure 4: False acceptances produced by (a) MLPs and (b) RBF networks.

[7] D. Green and J. Swets, *Signal Detection Theory and Psychophysics*. New York: Wiley, 1966.

[8] M.I. Hannah, A.T. Sapeluk and R.I. Damper, “The effect of utterance content and length on speaker-verifier performance,” *Proc. ESCA Eurospeech ’93, Vol. 3,* Berlin, Germany, pp. 2299–2302, 1993.

[9] M.I. Hannah, A.T. Sapeluk and R.I. Damper, “The rôle of the reference template in speaker verification,” *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland, pp. 181–184, 1994.

[10] S. Haykin, *Neural Networks – A Comprehensive Foundation*. New York: McMillan, 1994.

[11] M. Mak, W. Allen and G. Sexton, “Speaker identification using multilayer perceptrons and radial basis function networks,” *Neurocomputing*, vol. 6, pp. 99-117, 1994.

[12] J.D. Markel and A.H. Gray Jr., *Linear Prediction of Speech*. Berlin: Springer-Verlag, 1976.

[13] J. Moody and C. Darken, “Fast learning using locally-tuned processing units,” *Neural Computation*, vol. 1, pp. 281–294, 1989.

[14] J. Oglesby and J. Mason, “Radial basis function networks for speaker recognition,” *Proc. IEEE ICASSP ’91, vol. 1*, Toronto, Canada, pp. 393-396, 1991.