

SPEAKER IDENTIFICATION USING RADIAL BASIS FUNCTIONS

M.W. Mak, W.G. Allen and G.G.Sexton

University of Northumbria at Newcastle, U.K.

ABSTRACT

This paper describes a text-independent speaker identification system based on Radial Basis Function (RBF) networks. Both text-dependent and text-independent speaker identification experiments have been conducted. The database contains 7 sentences and 10 digits spoken by 20 speakers over a period of 9 months. LPC-derived cepstrum coefficients are used as the speaker specific features. The results show that RBF networks offer fast learning speed and good generalization even in text-independent mode. Moreover, a robustness test has been carried out which demonstrates that RBF networks provide sufficient information to produce a 'no match' decision in speaker identification applications.

1. INTRODUCTION

Speaker recognition can be divided into speaker verification and speaker identification. The objective of a speaker verification system is to verify whether an unknown voice matches the voice of a speaker whose identity is being claimed. In speaker identification, we want to identify an unknown voice from a set of known voices. Speaker verification systems are mainly used in security access control while speaker identification systems are mainly used in criminal investigation. This paper will concentrate on speaker identification.

Throughout this paper, the term *text-dependent* means that the test set and training set use the same context. The term *text-independent* means that the test set and training set use different contexts.

The data processing of a speaker recognition system usually involves two steps: feature extraction and pattern classification. The effectiveness of various features for speaker recognition has been examined. For example, Atal [1] finds that cepstrum coefficients provide the highest identification score while the area coefficients provide the lowest. For the pattern classification phase, various methods have been used. For example, Furui [2] uses the dynamic time warping technique and Soong [3] uses vector quantized codebooks to store the features. Recently, Oglesby and Mason [4-5] investigate the possibility of using neural networks as the pattern

classifiers in speaker recognition tasks. They use Cepstrum coefficients as speaker specific features to train the Backpropagation networks, and show that the recognition performance is comparable to that of the vector quantization approach. Oglesby and Mason [6] also report a speaker verification system based on Radial Basis Function networks. As they use spoken digits during the training and recognition phases, their system is text-dependent.

We have reported earlier a comparison between the RBF and Backpropagation networks in speaker identification [7]. The system uses spoken digits from 10 speakers, and therefore is text-dependent. This paper extends our previous experiments to text-independent speaker identification. Moreover, the robustness of RBF classifiers in producing 'no match' decisions will be examined. Section 2 gives a brief introduction to RBF networks and shows the methods of finding the network parameters. Section 3 describes the procedures and results of the text-dependent and text-independent speaker identification experiments.

2. RBF NETWORKS

Radial Basis Function (RBF) networks [8] can be viewed as a feedforward neural network with a single hidden layer. An RBF network with n inputs, M hidden units and K outputs is shown in Fig. 1. Each hidden unit is a non-linear function (usually Gaussian) with output related to the distance between the input vectors and the centroid of the basis function. The output layer forms a linear combiner which calculates the weighted sum of the outputs of the hidden units. Note that a multiple output RBF network (Fig. 1) can be considered as the combination of several single output RBF networks (Fig. 3). Therefore, the output of an RBF network (with single output unit) is given by:

$$R(\mathbf{x}) = w_0 + \sum_{i=1}^M w_i \Phi_i(\|\mathbf{x} - \mathbf{c}_i\|) \quad (1)$$

where w_i ($i=1, \dots, M$) are the network weights, \mathbf{x} is the input vector, \mathbf{c}_i are the function centres and $\Phi_i(\|\cdot\|)$ are the non-linear functions and w_0 is the bias. If Gaussian-shaped basis functions are used as the non-linearity, (1) can be written as

$$R(\mathbf{x}) = w_0 + \sum_{i=1}^M w_i \exp\left\{-\frac{1}{2\sigma_i^2}(\|\mathbf{x}-\mathbf{c}_i\|)^2\right\} \quad (2)$$

where σ_i are the function widths which control the influences of each Gaussian function over a small region with centroid at \mathbf{c}_i . One advantage of the RBF networks over the Backpropagation networks is that the parameters in the hidden layer (σ_i, \mathbf{c}_i) and the parameters in the output layer (w_i) can be calculated separately. This leads to a fast learning algorithm as described below.

There are 3 steps in calculating the parameters of an RBF network. First, the function centres \mathbf{c}_i are determined by the K -means algorithm [9]. Second, the function widths σ_i are calculated by the P -nearest neighbours rule:

$$\sigma_i = \sqrt{\frac{1}{P} \sum_{j=1}^P \|\mathbf{c}_i - \mathbf{x}_j\|^2} \quad (3)$$

where \mathbf{x}_j are the P -nearest neighbours of the centre \mathbf{c}_i . P is usually set to 2. Finally, the network weights w_i are estimated by minimizing the total-squared error

$$E = \sum_{p=1}^N \left[d(\mathbf{x}_p) - w_0 - \sum_{i=1}^M w_i \Phi_i(\|\mathbf{x}_p - \mathbf{c}_i\|) \right]^2 \quad (4)$$

where $d(\mathbf{x}_p)$ is the desired output for the input vector \mathbf{x}_p and N is the total number of training patterns. This leads to the matrix equation

$$\mathbf{A}\mathbf{w} = \mathbf{d} \text{ or } \mathbf{w} = \mathbf{A}^{-1}\mathbf{d} \quad (5)$$

where \mathbf{A} is a $N \times (M+1)$ matrix with elements $\Phi_i(\|\mathbf{x}_p - \mathbf{c}_i\|)$ and \mathbf{A}^{-1} is the pseudo inverse of \mathbf{A} . In this study, \mathbf{A}^{-1} is obtained by singular value decomposition. Equation (4) shows that the error function is parabolic in the weight space (w_i). Therefore, there is no local minimum and only one global minimum exists. The non-adaptive nature of the least squares method leads to a training scheme which can be an order of magnitude faster than the Backpropagation networks. For instance, Renals [10] finds a training time of four minutes for an RBF network versus three hours for a Backpropagation network.

It has been shown [7] that selecting the function centres by the K -means algorithm gives better generalization than selecting the centres randomly from the training samples. It implies that the locations of function centres are vital to the performance of RBF networks. For example, Fig. 2 shows the decision surfaces of an RBF network in solving an XOR problem. In Fig. 2a, the function centres are far apart from each other so that a large section of the decision surface between the peaks and troughs has a value around 0.5. Therefore, it is

difficult for the classifier to determine which class a testing sample should belong to if it falls in this flat region. However, the function centres in Fig. 2b are wisely chosen so that the region of flat surface or the region of uncertainty is reduced dramatically.

Moreover, the Gaussian-shape basis functions ensure that an RBF network will not produce a high output for the samples which are far away from the centroids (see Fig. 2). This helps to reduce the extrapolation errors which occur when test data fall beyond the range of the training data and are misclassified. Moreover, each hidden node represents a small region in the feature space, i.e. the effect of each hidden unit is completely local. These characteristics enable the RBF networks to generalize extremely well, even in complicated tasks such as text-independent speaker identification. This will be shown in the next section.

3. EXPERIMENTS

The configuration of the speaker identification system is shown in Fig. 3. Each speaker has his/her own personalized network. Each RBF network has 12 inputs and 1 output with the output being active for the features associated with that speaker. The final decision is made by a winner-takes-all unit such that the speaker identity corresponds to the network with the highest output.

3.1 Speech Database

The speech database contains 4 utterances of 7 sentences and 10 digits spoken by 20 speakers (10 males and 10 females) over a period of 9 months. The contents of the database are shown in Table 1. The speech signals were band-limited at 50Hz and 3.5kHz. The filtered speech was then sampled at 8kHz by a 14 bit A/D converter. The silent portions and pauses in the utterances were removed manually and the resulting speech signals were pre-emphasised by a filter with transfer function

$H(z) = 1 - 0.94z^{-1}$. Then, a 50% overlapping Hamming window with a window size of 32ms is applied, i.e. 32ms per frame. 12th order LPC-derived Cepstrum coefficients are calculated for each frame. These form the feature vectors for the RBF networks.

3.2 Text-dependent Speaker Identification

There are 3 text-dependent speaker identification experiments (TD-1, TD-2 and TD-3). Each experiment is divided into two phases: training and recognition. During the training phase, each RBF network is trained independently using the K -means algorithm, P -nearest

neighbour rule and the least squares method (see section 2). The K -means algorithm is applied on the patterns from each speaker independently because this clustering scheme gives better generalization [7]. Therefore each network contains an equal number of function centres from each speaker. During the recognition phase, five digits (selected from the test set) are fed to each network successively. The classification is correct if the network which gives the highest average output is the speaker's own network. The process is repeated for all possible combinations of the five digits out of 10 digits. By averaging the number of correct classifications, the average identification accuracy is obtained. Moreover, the identification confidence is defined. The identification confidence is the difference between the two highest outputs among the K networks (for K speakers). Since the desired output $d(\mathbf{x}_p)$ in Equation (4) is either a '0' or a '1', the theoretical minimum and maximum output values of an RBF network is '0' and '1' respectively. Therefore, the identification confidences are bounded between these values, with '0' meaning no confidence at all and '1' meaning complete confidence.

Experiment TD-1 uses the first and second utterance of 10 digits from 10 speakers (5 male and 5 female) as the training set. The test set is formed by the third and fourth utterance of 10 digits from the same speakers. The identification performances for different numbers of function centres are shown in Table 2. Note that the identification accuracy increases with the number of function centres. This implies that the number of function centres affects generalization and classification accuracy of RBF networks. However, the improvement in identification accuracy is negligibly small when the number of function centres is larger than 300. For example, the identification accuracy only improved by 1.2% when the number of function centres is increased by 100% (from 300 to 600). Moreover, the training time increases dramatically with the number of function centres. For example, it takes 25 minutes to train an RBF network with 300 centres on one T-800 Transputer, but takes 73 minutes to train an RBF network with 600 centres.

Experiment TD-2 increases the population size to 20 and uses the same number of function centres per speaker as experiment TD-1 (except 60 centres per speaker). Table 3 shows the identification results. The results for 200, 400, 600 and 800 centres in Table 3 should be compared with the results for 100, 200, 300 and 400 in Table 2, respectively. Note that identification accuracy is lower when the population size increases. Moreover the least-squares method fails to find a solution when 800 function centres are used. This suggests that other techniques such as the gradient method [11] should be used when the number of training samples is too large. However, similar to the Backpropagation algorithm, the gradient method suffers

from local minima and slow convergence.

Experiment TD-3 divides the 20 speakers into two groups: registered speakers and unregistered speakers. Both of them contain 5 male and 5 female. The first and second utterance of 10 digits from the registered speakers form the training set. The test set is formed by the utterances of 10 digits from the unregistered speakers. Therefore, the robustness of the classifiers in producing 'no match' decisions can be examined. We have devised two robustness indicators in RBF classifiers: identification confidence and normalized distance D_k of test patterns (from speaker k) to the nearest function centres. D_k is defined as:

$$D_k = \frac{1}{N_k} \left\{ \sum_{p=1}^{N_k} \frac{\|\mathbf{x}_p - \mathbf{c}_i\|}{\sigma_i} \right\} \quad (6)$$

where N_k is the number of test patterns (\mathbf{x}_p) from the k -th speaker and σ_i is the width of the function centre \mathbf{c}_i , and $\|\bullet\|$ is the Euclidean distance. Therefore, D_k denotes the closeness of the test patterns to their nearest function centres with respect to the function widths. Moreover, the outputs of the networks will be nearly equal for the speech associated with the unregistered speakers. Therefore, the identification confidence will be low. Table 4 shows the identification confidence and the average normalized distance to the nearest function centre of 10 registered and 10 unregistered speakers. 300 function centres per network were used. Note that the average identification confidence for the registered speakers is 4.5 times that of the unregistered speakers. Moreover, the average distance to the nearest function centre of the unregistered speakers is larger than the registered speakers. This suggests that RBF networks can be used for producing 'no match' decisions.

3.3 Text-independent Speaker Identification

There are two text-independent experiments (TI-1 and TI-2). They are described below.

Experiment TI-1 uses two utterances of the 10 digits from 10 speakers as the training set (as in Experiment TD-1) and uses four utterances of 7 sentences as the test set. The aim of this experiment is to examine the performance of the RBF classifiers (working in text-independent mode) when only short training utterances are available. Table 5 shows the test results for three different numbers of function centres. Note that the identification accuracies and identification confidences are much lower than the text-dependent case when we compare Table 5 with Table 2. It is because the patterns from the sentences have never been seen by the networks before. Therefore, the test patterns are far away from the function centres. This leads to low network

outputs and low identification confidence. Note also that 400 function centres gives the best identification result. If there are too many function centres (e.g. 600), the networks begin to memorize the details of the training set (the 10 digits), and fail to develop general rules that apply to unseen data. This is similar to the overtraining phenomenon of Backpropagation networks.

Experiment TI-2 uses the first utterance of sentences 1 and 2, spoken by 10 speakers, as the training set. The test set is formed by 4 utterances from sentences 3 to 7 (see Table 1). The aim of this experiment is to compare the text-independent speaker identification performance when sentence-length utterances are used as training set and test set. Table 5 shows the text-independent identification performance when the number of sentences used in the test set varies from 1 to 5. Table 6 shows that identification accuracy approaches 100% if sufficient test patterns are available. Comparison of Tables 5 and 6 also show that identification performance can be improved by using sentence-length utterances rather than isolated digits. This is because sentences usually have longer duration and more phonetic variations.

4. CONCLUSION

A series of text-dependent and text-independent speaker identification experiments based on Radial Basis Function networks has been conducted. The results show that it is feasible to use RBF networks in a text-independent speaker identification system. For the text-dependent case, the best performance based on 5 test digits is 95.7%. For the text-independent case, the performance based on a single sentence test utterance is 83.5%. Moreover the text-independent identification performance reaches 100% when sufficient training and testing feature vectors are available. Finally, the robustness test demonstrates that RBF networks can be used for producing 'no match' decisions in speaker identification.

REFERENCES

- Atal, B.S., 1974. J. Acoust. Soc. Amer., 55 (6), June, 1304-1312.
- Furui, S., 1981. IEEE Trans. on ASSP, ASSP-29 (2), 254-272.
- Soong, F.K. & Rosenberg, A.E., 1985. Proc. ICASSP, 387-390.
- Oglesby, J. & Mason, T.S. 1989. , Proc. of the 1st IEE ICANN, 306-309.
- Oglesby, J. & Mason, T.S. 1990. Proc. IEEE Int. Conf. on ASSP, 261-264.
- Oglesby, J. & Mason, T.S. 1991., Proc. ICASSP-91, 393-396.
- Mak, M.W., Allen, W.G. & Sexton, G.G. 1993. to appear in J. of Microcomputer Applications, April.
- Broomhead, D.S. & Lowe, D. 1988. Complex Syst. 2, 321-355.
- Macqueen, J. 1967. Proc. 5th Berkeley Symp. Math. Stat and Prob., 281-297.
- Renals, S., Rohwer, R. 1989. Proc. IJCNN, 1.461-1.467.
- Robinson, A.J., Niranjan, M. & Fallside, F. 1988. Cambridge University Engineering Department, Report # CUED/F-INFENG/TR 25.

1	We were away a year ago
2	I know when my lawyer is due
3	I'm naming amazing men
4	Intermediate allowance
5	Numeric interrogation
6	Available terminals
7	Allowable clearance
8-17	zero - nine

Table 1 - Sentences and words in the database

No. of function centres	Identification accuracy	Identification confidence
100	87.7%	0.22
200	90.1%	0.22
300	94.5%	0.27
400	95.3%	0.29
600	95.7%	0.28

Table 2 - Results of text-dependent experiment TD-1 (10 speakers)

No. of function centres	Identification accuracy	Identification confidence
200	74.5%	0.17
400	89.7%	0.24
600	88.1%	0.28
800	Failure to find a solution	

Table 3 - Results of text-dependent experiment TD-2 (20 speakers)

Speakers	Mean identification confidence	Average normalized distance to nearest centres
Un-registered	0.06	1.58
Registered	0.27	1.21

Table 4 - Results of robustness test TD-3

Number of function centres	Identification accuracy	Identification confidence
300	67.9%	0.08
400	72.5%	0.09
600	52.5%	0.07

Table 5 - Results of text-independent experiment TI-1 (10 speakers)

Sentences per Test	Identification accuracy	Identification confidence
1	83.5%	0.16
2	88.3%	0.15
3	91.5%	0.15
4	95.5%	0.14
5	100.0%	0.14

Table 6 - Results of text-independent experiment TI-2 (sentence length utterances)

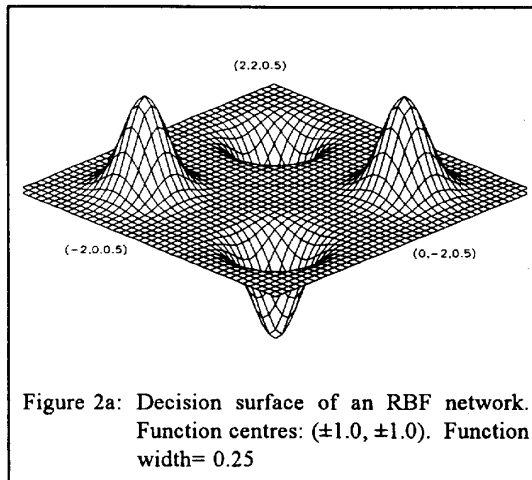


Figure 2a: Decision surface of an RBF network. Function centres: $(\pm 1.0, \pm 1.0)$. Function width = 0.25

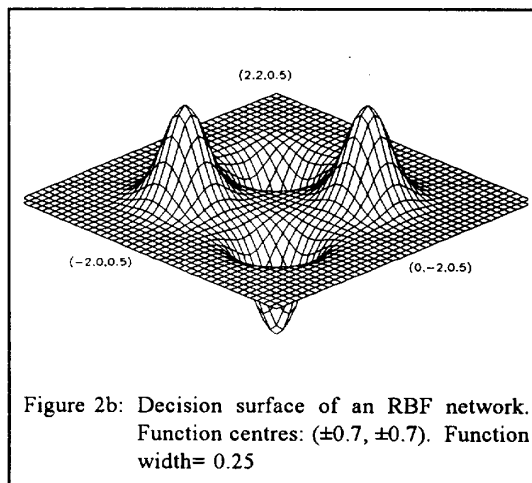


Figure 2b: Decision surface of an RBF network. Function centres: $(\pm 0.7, \pm 0.7)$. Function width = 0.25

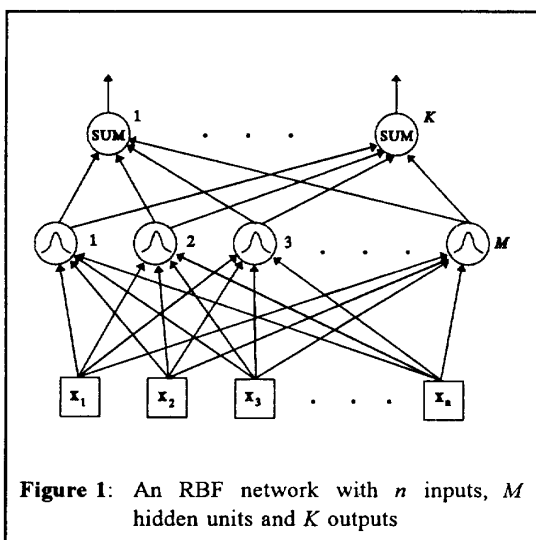


Figure 1: An RBF network with n inputs, M hidden units and K outputs

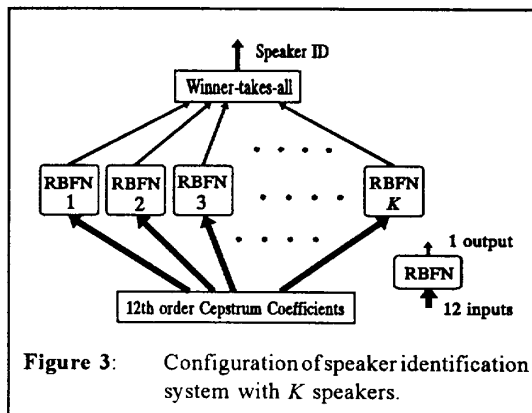


Figure 3: Configuration of speaker identification system with K speakers.