A HYBRID MLP-RBF BASED SPEAKER VERIFICATION SYSTEM

Liam Kilmartin and Eliathamby Ambikairajah

Speech Research Group
Department of Electronic Engineering
Regional Technical College, Athlone, Ireland

ABSTRACT - A two stage neural architecture model is proposed in this paper for the task of speaker verification. This model operates solely in the time domain and hence removes the need for any computationally demanding pre-processing of the input speech signal. The first stage in the model is a Multi-Layer Perceptron (MLP) based non-linear speech predictor. This non-linear prediction process is non-recursive in nature and is carried out for a complete utterance. When the weights in the MLP have converged after several epochs of training, they are then applied as inputs to the second stage of the model. This second stage is a Radial Basis Function (RBF) classifier which will output a decision as to whether the utterance originated with the true speaker or not. The results obtained from initial experiments with this model were promising but a slight change in the traditional learning process of the MLP stage causes a great improvement in the results obtained. This new model was tested with a small database of speakers. It was found that the model could distinguish correctly all of the utterances used to train it but also could generalise to distinguish correctly all of the utterances in a previously unseen test set.

INTRODUCTION

The process of speaker verification uses the intrinsic individuality of a person's voice to determine if they are who they claim to be. A large number of speaker verification systems have been developed and their performances explored. In general, most of these systems do not operate in the time domain and thus require a large amount of computationally demanding pre-processing of the speech signal before the verification process itself can be undertaken. Indeed the performance of many of these systems based on LPC analysis, time warping or hidden Markov modelling have proven to be quite impressive.

In recent years, the application of neural networks in many areas of speech processing and understanding has become increasingly important. Many applications based on connectionist models have been developed for the process of speaker verification (Oglesby and Mason, 1991, Bennani et al., 1990). However the main problem which still remains for such neural network based systems is that a large amount of pre-processing still must be carried out before the information is applied to the neural network.

The model which is proposed in this paper attempts to remove the need for such pre-processing of the input speech signal. The two stage model is designed to operate in the time domain alone and performs well without the need for time warping of any type. Figure 1 shows a diagram of the proposed model.

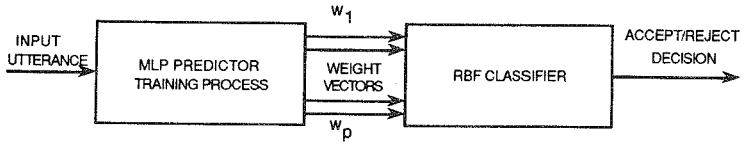

Figure 1. Block diagram of the proposed hybrid MLP-RBF speaker verification model

The model stages are both connectionist in their nature. The first stage is a Multi-Layer Perceptron (MLP) neural network which operates in the time domain and is used to extract the speech parameters which are used in the verification stage. The MLP is trained to operate as a non-linear speech predictor for the utterance spoken by the unidentified claimant. The second stage then implements the speaker verification process based on a Radial Basis Function (RBF) classifier using the weights of

the MLP as its inputs. The RBF classifier was previously trained to accept the weights produced by a true speaker utterance applied to stage one and to reject all other speakers.

BACKGROUND

The basic premise behind the operation of this model is that the MLP is in some way capable of learning the underlying speaker dependent trends in the speech utterance. There has been a large amount of work carried out in the use of neural architectures for the purpose of time series prediction and in particular speech prediction. The ability of an MLP to predict a chaotic time series quite well was illustrated by work carried out by Lapedes and Farber (1987). This work showed the ability of an MLP to predict, after training, a chaotic time series generated by the Mackey-Glass equation. The MLP used to predict the chaotic series had quite a low complexity.

Further work (Lowe and Webb, 1989) examined the ability of a connectionist model to predict the time series of a speech utterance. It was shown that this model could quite easily be trained to predict the speech waveform in a non-recursive mode. However, after the training process was completed, it was attempted to operate the model in a recursive manner. In this mode, the predicted waveform value was fed back as an input for the prediction of the next sample. The resultant time series was often chaotic in nature but in some cases bore some relationship to the original speech. This was particularly true of the voiced sections where the chaotic waveform seemed quasi-periodic in nature and seemed to have the same fundamental frequency and pitch period as the vocal tract of the original speech. Two very important points were discussed in this work. Firstly, an MLP was generally of not a sufficient complexity to model the time varying parameters of the speech production model during the production of an utterance. Because of this, the model could never work well in a recursive mode and hence as a low bit rate speech coder. However, because of the similarities between the original speech and the chaotic series produced by the recursive prediction, it was apparent that the connectionist model was learning the operation of the underlying speech production mechanism rather than producing an exact non-linear predictor. It is this property of the network which is used by the system proposed in this paper.

The use of an MLP as a non-linear speech predictor has also been examined (Lovell and Tsoi, 1990). The MLP was seen to operate well when predicting in a non-recursive mode. It was also proposed in this work that the weights of the MLP could be used in a speaker verification system. Some initial clustering experiments were reported which suggested that the weights of the MLP could be used as a speaker dependent parameter.

MLP PREDICTOR NETWORK

The first stage in the speaker verification process is to train the MLP neural network to act as a non-linear predictor for the utterance being applied to the system. In this way, the MLP should learn to predict the n+1 speech sample using the previous k samples. The structure of the MLP neural net is based on the traditional network (Lippmann, 1987) with one layer of hidden nodes and a single output neuron and is shown in figure 2.

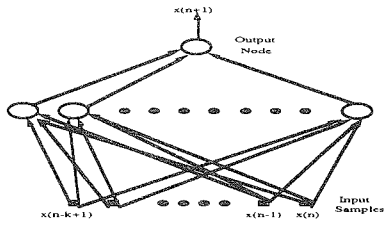

Figure 2. Structure of the MLP predictor stage

The training process by which the MLP learns to become a predictor is based on the Error Back Propagation algorithm (Rummelhart et al., 1987). The utterance to which the MLP is to be trained is first manually endpointed. In the initial stages of the development of the model, all of the internal weights of the MLP were set to totally random values. A sample within the utterance is chosen at random and the previous k samples are used to attempt to predict the next sample. This predicted

sample is the value which is produced at the output node of the MLP and it is then compared to the actual next sample value. Depending on the error produced, the weights in the MLP are then adapted. This process is randomly carried out for each sample within the utterance. The MLP is then continuously trained by repeating this whole process for several epochs until the weights converge to some fixed set of values. At this stage the MLP is acting as a non-recursive non-linear predictor for that utterance.

The results obtained using this initial procedure were promising but a slight alteration in the training procedure caused a noticeable improvement in the performance of the overall system. In the new training procedure, the weights of the MLP at the beginning of the training process are initially set to a particular set of values rather than to a random set of values. This particular set of values is the set of weights arrived at by training an MLP to predict the test utterance as spoken by the true speaker. By setting the weights to this value initially, if the utterance being predicted is that of a true speaker, the weights do not diverge that much from their initial value. If the utterance being predicted is that of an impostor, then the weights will significantly diverge from their initial values. It is in this manner that the true speaker/impostor decision can be made. It would seem that the initial set of weights lie close, in multi-dimensional space, to a local minimum for the error function for any repetition of the utterance by the true speaker but are spuriously positioned with respect to the error function for an impostor.

After several epochs of training with the complete utterance the weights will have converged to a steady set of values. At this point the training process is halted and the weights within the MLP are then stored. At this stage, the speaker dependent information should be contained within the MLP weights. These weights are used to form an input vector for the RBF classifier stage.

RBF CLASSIFIER STAGE

An RBF classifier is a connectionist model of a similar form to the MLP neural network. The form of the RBF classifier is described in figure 3. An N dimensional weight vector is applied at the input to the classifier. This vector is then compared to each of the basis functions in turn. Each basis function has an N dimensional centre associated with it and a symmetric function with an effective "radius". Depending on the distance of the test point from the basis centre compared to this "radius", a certain output is produced by each basis function. The value that is produced by the basis function depends on the form of the function. In this case the function is a normal Gaussian distribution. In a manner similar to that of an MLP, the output values of each of the basis functions are weighted and summed to produced the single classifier output. A value of '1' at this output means that the utterance has been accepted as belonging to the true speaker and a value of '0' means that it is an impostor's utterance.

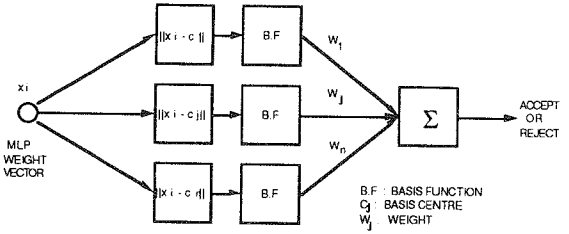

Figure 3. Structure of the RBF classifier

Unlike the MLP stage of the system, this stage is acting as a classifier and must be tuned therefore to only accept true speaker weight vectors and reject all others. This must be carried out before the RBF is used to test unknown cases. To this end, a set of training examples of both true speakers and impostors are chosen. The number of basis functions is chosen to be equal to the number of cases in the training set. The N dimensional weight vectors describing these training cases are then assigned to be the centres for the various basis functions. A certain "radius" value must be determined for all of the functions. The training process is then carried out in order to determine the values of the weights linking the output of the basis functions to the output node. This is achieved by randomly choosing cases from the training set and comparing the RBF output to the desired output. The error between these two values is used to update the weights using a standard Least Mean Squared (LMS)

algorithm (Botros and Atkeson, 1991). The complete set of training cases is continuously cycled through until the weights within the RBF converge to a steady value. The ability of the RBF to generalise to previously unseen test cases can now be evaluated.

SPEAKER DATABASE

For the purpose of developing and testing this model, a small database of speakers repeating the same utterance was taken. This consisted of a total of five different speakers all repeating the utterance "six..three..nine". These utterances were spoken in an laboratory environment and sampled using a high quality 16 bit speech acquisition system. Before the sampling process, the speech was bandpass filtered from 70 Hz to 3.4 kHz. The true speaker repeated the utterance a total of 50 times over a period of a few weeks and the four impostors each repeated the utterance a total of five times each over the same time period. All of these utterance were spoken in a normal conversational tone which often resulted in coarticulation. The complete utterances were each manually endpointed. For the purpose of these experiments, each of the three words within the utterances were roughly endpointed and hence used as utterances in their own right for the purpose of evaluating the system.

RESULTS AND DISCUSSION

The system which has been described in this paper has been tested with the above database of speakers. In order to carry out the training process of the RBF classifier section and to test the resultant model, the complete database was divided up into a training set and a test set of utterances. The training set was made up of 20 examples of the true speaker speaking the utterance and 8 cases of impostors speaking the utterance. The make up of the impostor cases, in the training set, was examined and the results obtained using examples of all four impostors showed little variance from a training set which included utterances by only two of the impostors. The test set of cases was made up of the remaining 30 utterances by the true speaker and the remaining 12 utterances by the impostors. This test set was not used in the training process and thus gave a clear indication of the RBF's ability to classify previously unseen utterances.

A large amount of experimentation was involved in determining the structure of the MLP predictor. It was found that a predictor with 10 inputs (utilising the 10 previous samples to predict the next sample) and a single hidden layer with 3 nodes was sufficient to track the desired input waveform. It was generally noted that quite a small number of epochs were necessary for the MLP weights to converge. A value of training for 6 epochs tended to result in a good predictor. The predicted waveform was played back to examine the predictor's quality and was found to be almost indistinguishable from the original speech.

There were also some parameters concerning the RBF classifier which had to be resolved. Early experiments in the training of the RBF showed that the choice of "radius" for all of the basis functions was of critical importance. However, it was generally found that a good choice for such a "radius" would be equal to the average inter-vector distance for the training set of input weight vectors. As stated before, each of the resultant 37 dimensional weight vectors for each of the examples in the training set of cases was used as the centre for one of the basis functions. A further system parameter which was of importance in analysing the performance of the system was the number of passes of the LMS algorithm needed to causes the RBF weights to converge. Figure 4 shows a plot of the reduction of the normalised mean squared error during training with each pass of the complete training set. It shows that typically 80 to 100 passes of the training set were required for convergence to occur.

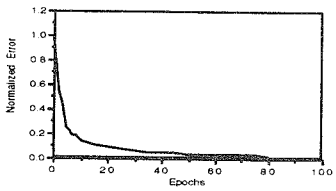

Figure 4. Variation of normalised mean squared error during training with training epoch for the utterance "three"

A further experiment attempted to compare this performance with that of an MLP neural network which had to carry out the same classification task. An MLP with 30 nodes in the hidden layer almost performed as well as the.RBF classifier. Such an MLP is extremely more complex than the proposed RBF stage and also would take a much longer time to train. In this case the MLP network took typically 300 epochs to train compared to the 80 epochs required to train the RBF and even then the MLP did not quite perform as well as the RBF classifier.

In order to examine the overall ability of the system in the process of speaker verification, the training and test sets of each of the utterances "six", "three" and "nine" were separately tested on the system. Figure 5 is a diagrammatic representation of the results of the verification process for two of the utterances.

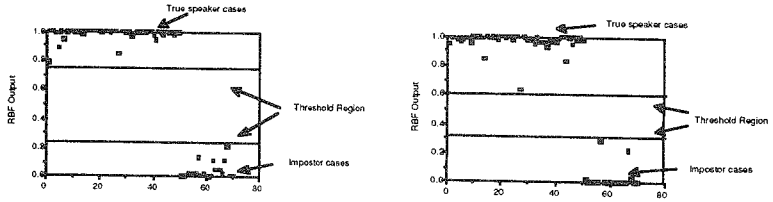

Figure 5. Verification results for the utterances "nine" and "six"

The utterances numbered 1 to 20 and 51 to 58 are the training set utterances for the true speaker and impostors respectively. The utterances 21 to 50 and 59 to 70 are the unseen test cases for the true speaker and impostors respectively.

It is clearly illustrated in these diagrams that the verification system has correctly classified all of the speakers in both the training and test sets. What is even more important than this is that the system makes quite a large distinction between the true speaker utterances and those of the impostors. In both diagrams it is clear that there is quite a large region in which a threshold could be placed to distinguish between the true speaker and an impostor. The results obtained for the utterance "three" were of a similar nature to those shown above. Overall, it can be seen that for a single word utterance the system has performed very well at its classification task.

The next experiment to be carried out was to investigate, in some way, the ability of the network to deal with a multi-word utterance such as the original "six..three..nine" utterance. Since the individual words were only generated by a rough process of endpointing, it was felt that this task could be undertaken by combining the input weight vectors for each word to form a 3x37 dimensional input vector to the RBF classifier stage for each of the original utterances. As the RBF was not being increased in its complexity but merely in the dimensionality of the basis functions, the training process for the RBF in this experiment was not noticeable longer than for the single word utterances. Figure 6 shows the results obtained in this experiments. The results show that, as for the single word utterance, the system correctly classifies all of the cases in the training and test sets. The combination of all of the weights at the input vector stage has resulted in an even wider region in which the distinction threshold could be placed.

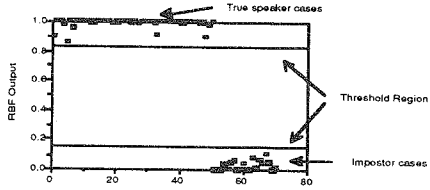

Figure 6. Verification results for the complete utterance "six three nine"

CONCLUSIONS

The model which has been proposed in this paper for the purpose of speaker verification has been tested on a small database of speakers. However, the results obtained show that the system is able to distinguish correctly between true speakers and impostors both in its own training set but, more importantly, also with an unseen test set. The system has been extensively tested on single word utterances but also experiments indicate that it will be equally as good in operating with multi-word utterances.

It was noted that the system is capable of rejecting impostor utterances even if there had been no utterance by that impostor in the original training set for the system. This illustrates that the RBF classifier stage has trained very well to recognise the true speaker and rejecting all other input weight vectors. This suggests that the decision surface for the RBF stage is probably similar to a plateau shape with very steep sides. The true speaker input weight vectors would lie in the plateau region with all other possible input vectors lying beyond the steep transition region around the plateau. Since the weight vectors produced by the MLP predictor do not vary that much from their initial values for a true speaker, there would not seem to be a problem with overtraining the RBF system. In many RBF classifiers, the ability of the classifier to generalise to unseen cases is reduced if the training process is carried on for too many epochs. This occurs because the RBF trains too much to its training set. However, in this application, the process of over training may in fact be an advantage in steepening the transition region between the well defined accept region and the reject region.

Further work is being carried out on this system to evaluate the effect of varying the MLP predictor topography and also the number of weights passed to the RBF stage. Also, it is hoped to apply a complete continuous multi-word utterance to an enlarged MLP predictor and to examine the ability of the RBF to classify the resultant weight vectors correctly.

ACKNOWLEDGEMENTS

This research has been funded by EOLAS, the Irish Science and Technology Agency, under the Applied Research Programme 1990-1992. The authors would also like to thank Prof. D. J. Wilcox of the Department of Electronic Engineering, University College, Galway for providing facilities during the duration of the research.

REFERENCES

Bennani, Y. et al. (1990) *A connectionist approach for automatic speaker identification*, CH2847-2/90/0000-0265 1990 IEEE, 265-268

Botros, S. M. and Atkeson, C. G., (1991) *Generalization properties of radial basis functions,* Advances in Neural Information Processing Systems 3, 707-713

Lapedes, A. and Farber, R. (1987) *Nonlinear signal processing using neural networks: Prediction and system modelling,* Los Alamos National Laboratory , LA-UR--87-2662 DE88 006479

Lippmann, R. P. (1987) *An introduction to computing with neural nets*, IEEE ASSP Magazine 0740-7467/87/0400-0004 1987 IEEE, 4-22

Lovell, B. C. and Tsoi, A. C. (1990) *Speaker verification using artificial neural networks,* Proceedings of the Third Australian Speech Science and Technology Conference 1990, 298-303

Lowe, D. and Webb, A. (1989) *Adaptive networks, dynamic systems and the predictive analysis of time series,* Proceedings of the First IEE International Conference on Neural Networks, 1989

Oglesby, J. and Mason, J. S. (1991) *Radial basis function networks for speaker recognition,* CH2977-7/91/0000-0393 1991 IEEE, 393-396

Rummelhart, D. et al. (1987) *Learning internal representations by error propagation,* Parallel Distributed Processing, Rummelhart, D. and McClelland, J. L. (Eds), MIT Press 1987, 318-362