

A Committee of Neural Networks for Automatic Speaker Recognition (ASR) Systems

¹Viresh Moonasar and ^{1,2}Ganesh K Venayagamoorthy, *Member, IEEE*

¹*Department of Electronics Engineering, M L Sultan Technikon, Durban, South Africa*

²*Applied Computational Intelligence Laboratory, University of Missouri-Rolla, MO 6409, USA*
moonasv@telkom.co.za & gkumar@ieee.org

Abstract

This paper describes how the results of speaker verification systems can be improved and made robust with the use of a committee of neural networks for pattern recognition rather than the conventional single-network decision system. It illustrates the use of a supervised Learning Vector Quantization (LVQ) neural network as the pattern classifier. Linear Predictive Coding (LPC) and Cepstral signal processing techniques are utilized to form hybrid feature parameter vectors to combat the effect of decreased recognition success with increased group size (number of speakers to be recognized).

1 Introduction

Technological innovations and the information technology era have created a huge need for on-line security. The reluctance of on-line shoppers using their credit cards over the Internet has been a major factor for the slow take-off of e-business. Speaker recognition, which is the process of automatically recognizing who is speaking based on unique information inherent in speech signals, is a method that may be adopted to enhance security over the Internet and other security applications. Speaker verification accepts or rejects the identity claim of a speaker - is the speaker the person they say they are or not? ASR should be contrasted with speech recognition where the goal is to identify the words spoken by a user.

The extracted feature parameter is the key aspect of any successful speaker verification system [1]. This is the inherent critical information that is present in each speaker's voice sample. Feature parameters, obtained using specific signal processing techniques [2], are the basis of determining the speaker's identity [3]. No serious single scientific protocol has been able so far to evidence the existence of a fixed, robust, non-modifiable,

individual voice characteristic that could be extracted from a speech signal and indicate without doubt the speaker's identity. Therefore, hybrid feature vectors are used to optimize the benefits extracted from individual signal processing techniques. The use of an LPC-Cepstral hybrid feature parameter vector has proved successful for this speaker verification application [1].

The extracted feature vectors from each training sample are used as inputs to a neural network. Neural Networks are trained to "learn" and then recognize each subject's feature parameter characteristic [2]. LVQ neural networks have been used to perform the pattern recognition task. LVQs are closely related to Self-Organizing Maps (SOM), developed by Teuvo Kohonen. It is an algorithm that effectively maps similar patterns (pattern vectors close to each other in the input signal space) onto locations in the output space [4]. Learning Vector Quantization is a supervised version of SOM particularly suitable for statistical pattern recognition.

This paper illustrates the introduction of a Committee of Neural Networks instead of a single recognition network. The final committee decision will be based on majority voting of the member networks. Using several individual networks rather than a single neural network optimizes the output of the committee network. Each member of this network is a complete LVQ neural network.

A block diagram of the committee arrangement is shown in Figure 1. All the training vectors are presented to each of the individual LVQ networks. The decision block is not required during the training phase. The test feature vectors are also input to each of the member networks. Each network classifies the pattern independently and the net output is the majority decision vote of all members.

2 The Recognition System

Figure 2 illustrates a block diagram of a conventional recognition system. The analogue voice samples are

sampled into digital format. Characteristic feature parameters are extracted from each subjects voice samples. The processed signal parameters are then presented to a pattern classification network to either train or test the system. A single LVQ network represents the pattern classification and memory template blocks. Figure 3 shows how a committee of LVQ networks, as described in Figure 1 now replaces the single network system.

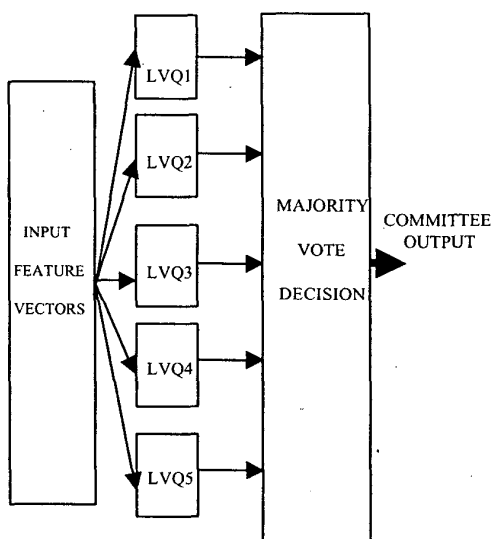


Figure 1: Committee of Neural Networks

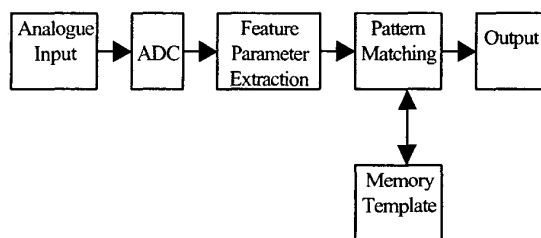


Figure 2: Conventional Recognition System

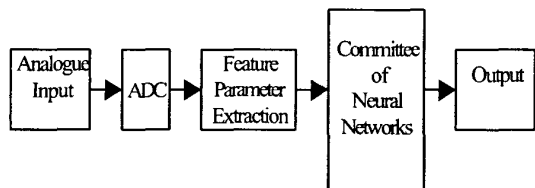


Figure 3: Recognition System Using a Committee of Neural Networks

3 Feature Parameters

Careful selection of feature parameters is critical to the neural network learning. Hence, raw voice data needs to undergo some sort of pre-processing. Our previous papers have shown improved results by using a hybrid feature vector comprising of LPCs and Cepstral coefficients than by using just a single feature alone [5]. Table 1 summarizes the recognition success rates for the different feature parameters used on a 10-speaker group.

Table 1: Recognition Success per Feature Parameter

Feature Parameter	Recognition Success Rate
Power Spectral Density (PSD) points	45%
Linear Prediction Coefficients (LPCs)	70%
Hybrid of LPCs and Complex Cepstrum (Cceps) - LCeps	90%

The complex Cepstral features shown above are derived from the 100 point LPC coefficients of the speech samples and not the raw recorded waveform. Thus, it can be considered a hybrid of LPC and Cepstral parameters. Calculating the Cepstral features from the raw waveform for our application was computationally too intensive (16000 samples per second * approximately 3 seconds per speech sample = 48000 samples per speech recording compared to 100 point LPCs). LPC coefficients are very good representations of the original waveform since the original waveform can be reconstructed from these coefficients. The Cepstrum was therefore calculated from the LPC coefficients.

Figures 4 and 5 show plots of the LPCs and Complex Cepstrum of a particular voice sample respectively. Detailed information on the different feature parameters is described in [6]

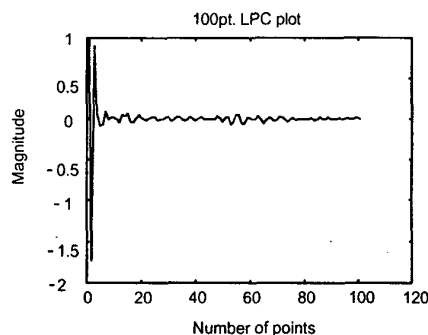


Figure 4: 100 point LPC plot.

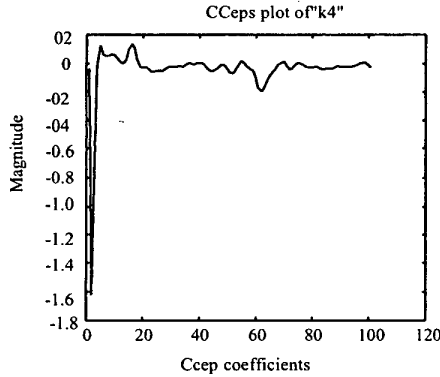


Figure 5: 100 point Complex Cepstrum plot

4 Artificial Neural Networks used for Speaker Verification

The members of the neural network committee are Learning Vector Quantization (LVQ) neural networks. Basics of Artificial Neural Networks and their components are detailed in [1], [3 -6]. Self-organizing networks learn to detect regularities and correlations in their input and adapt their future responses to that input accordingly. Competitive networks learn to recognize groups of similar input vectors in an unsupervised manner [7].

The minimization of classification errors is the main objective in most pattern recognition applications. Modeling of the probability densities of the competing classes often approaches this, but since it is, in practice, often not possible to assume any proper parametric density model, the lowest error rate is obtained by concentrating on the actual discrimination between the classes. The methods based on neural networks may outperform other methods in tough problems, where the prior knowledge cannot help much in the classification and the system characteristics must be learned automatically from the data. It is advantageous that the algorithm consists of a large number of very simple units capable of learning locally.

5 The Neural Network Committee

LVQ is a method for training competitive layers in a supervised manner. They learn to classify input vectors into target classes chosen by the user. An LVQ has a competitive layer followed by a linear layer. The linear layer transform the classes found the competitive layer into classes defined by the user [7].

LVQ training requires a training set of examples of the proper network behaviour. If the input pattern is classified correctly, then the winning weight is moved toward the input vector according to the **Kohonen** rule.

$${}_i W^1(q) = {}_i W^1(q-1) + \alpha(p(q) - {}_i W^1(q-1)) \quad (2)$$

If the input pattern is classified incorrectly, then the winning weight is moved away from the input vector according to the rule:

$${}_i W^1(q) = {}_i W^1(q-1) - \alpha(p(q) - {}_i W^1(q-1)) \quad (3)$$

This is known as LVQ1 training. Other variations exist where the neighbours of the winning weight are adjusted as well as the winning weight (LVQ2 and LVQ3). The neural committee utilized comprised of 5 LVQ members. Each member of the network was trained and then tested individually. The final decision was taken as the majority vote of the individual member networks. Eight samples per subject were used to train the individual LVQ member networks. Two additional samples per subject were used to test the recognition success of the committee.

The test results per input sample were similar for each of the member networks. This was due to the fact that all member networks were identical with the same weight initialization (vector W^1 shown in Figure 6). The weight vectors of the competitive layer are calculated using the midpoint theorem. The training input vectors must contain expected minimum and maximum values in their range. The outputs were similar for each network since the training algorithm is identical in each member network. This arrangement of the committee adds no value as compared to a single network system. All individual networks behave identically to the test inputs and the committee decision is always unanimous.

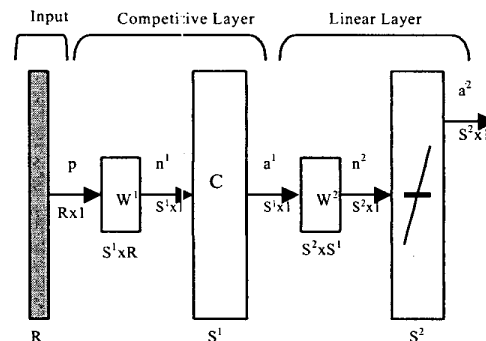


Figure 6: LVQ Network

Figure 6 shows an LVQ network with a competitive layer being followed by a linear layer. The input vector $P = (R \times 1)$ is the hybrid LPC-Cepstral feature vector per sample. This vector was scaled to 100 elements. These extracted feature vectors comprise fewer elements than the raw speech sample thus reducing the burden placed on the network.

The net input to the competitive layer is the negative distance between the prototype vectors, W^1 and the input:

$$n_i^1 = -\|W^1 - p\| \quad (4)$$

The output of the linear layer is given by the relationship:

$$a^2 = W^2 a^1 \quad (5)$$

S^1 is the number of neurons in the competitive layer. This parameter is user defined. The magnitude of W^1 is dependent on S^1 according to the relationship:

$$W^1 = S^1 \times R \quad (6)$$

Each member of the committee must process the input data using individual parameters. This results in outputs that can be cross-correlated to obtain the desired results from the committee. The S^1 parameter is therefore defined differently for each member network. S^2 is the number of target classes in the linear layer. This value is fixed at 20 (number of speakers in the group) for all member networks.

6 Results

The LVQs were trained with the following training parameters:

Table 2: Network Parameters

lr (learning rate)	0.001
me (maximum epochs)	15000
Number of speakers/subjects	20
Training samples per subject	8
Test samples per subject	2
Total Test Samples	(20x2)=40

The committee of networks does improve the overall recognition success to 95%. In this case, with these specific speech samples, an increase in the number of members, from three LVQ networks to five, did not improve the recognition success rate any further.

There is a significant increase in performance for LVQ5. The committee has produced a success of 95% compared to 77.5% obtained when using LVQ5 on its own.

Table 3: Recognition Success of Individual LVQ Networks

Member Network	S^1	Recognition Success
LVQ1	20	90.0%
LVQ2	30	92.5%
LVQ3	40	90.0%
LVQ4	50	90.0%
LVQ5	60	77.5%

Although the overall success rate is similar for each network in Table 3, the success rate per speaker sample differs. Particular samples are identified correctly by certain networks and incorrectly by the others.

Table 4: Recognition Success of the Committee of Neural Networks

Committee	Network Members	Recognition Success
3-member	LVQ1, LVQ2, LVQ3	95%
4-member	LVQ1, LVQ2, LVQ3, LVQ4	95%
5-member	LVQ1, LVQ2, LVQ3, LVQ4, LVQ5	95%

Choosing the architecture of the committee more carefully can enhance the performance of the system further. LVQ 5, with S^1 equal to sixty, has an inferior success rate as compared to the other individual networks. It should not be chosen as a member of the committee, based on its individual performance, and should be substituted with another member network that can add greater benefit.

Practically, one should compare the speed of computation, too, not only ultimate accuracies. A relative difference in accuracy of a few percent can hardly be noticed in practice, whereas tiny speed differences during actual operation are very visible. A single LVQ network with only thirty neurons in the hidden layer produces 92.5% accuracy while a five-member committee with a total of two hundred neurons in the competitive layer increases the output to 95%. Very few applications would compromise such a large expense of resources for this slight gain in accuracy.

An alternative method of changing the order in which the training samples are presented to each network was attempted. In this case, the number of neurons in the competitive layer, S^1 , was kept constant for all networks. This did not affect the recognition success rate per sample.

The order in which the training samples are presented to the individual networks does not matter for this application.

7 Conclusions

The use of artificial neural networks in voice recognition in our work has so far proved a fair amount of success, especially with the hybrid LVQ network. The performance of the system can be improved even further with a committee of neural networks as described in this paper but with the tradeoff of increased number of computations to carry out. With the faster processors coming into the market, this will not be a major issue.

The most significant factor affecting automatic speaker recognition performance is variation in the signal characteristics from trial to trial (inter-session variability and variability over time). Variations arise from the speaker themselves, from differences in recording and transmission conditions, and from background noise. There are also long-term changes in voices. It is important for speaker recognition systems to accommodate to these variations and this paper has demonstrated the use of a committee of neural networks to some extent achieve robustness in speaker recognition.

8 References

- [1] V.Moonasar, G.K.Venayagamoorthy, "Speaker Identification using Combined Feature Parameters as Inputs to an Artificial Neural Network Classifier", *Proceedings of Africon 99*, October 1999.
- [2] P.Krauss, L.Shure, J.N.Little, "MATLAB Signal Processing Toolbox User's Guide", The Mathworks Inc., 1996.
- [3] G.K.Venayagamoorthy, V.Moonasar, "Voice Recognition Using Neural Networks", *Proceedings of South African Symposium on Communications and Signal Processing*, pp. 29-32, September 1998.
- [4] T.Kohonen, *Self-Organizing Maps, 2nd Edition*, Berlin: Springer-Verlag, 1997.
- [5] V.Moonasar, G.K.Venayagamoorthy, "Artificial Neural Network Based Speaker Recognition Using A Hybrid Technique for Feature Extraction", *Proceedings of the ANNs In Engineering Conference (ANNIE 2000)*, November 2000, St. Louis, Missouri, USA
- [6] J.S.Carmona, "A Hybrid System with Symbolic AI and Statistical Methods for Speech Recognition", 1995.

[7] T.Kohonen, *Self-Organization and Associative Memory, 2nd Edition*, Berlin: Springer-Verlag, 1987.