ISOLATED WORDS, MULTISPEAKER SPEECH RECOGNITION WITH MULTILAYER NEURAL NETWORKS

Luyuan Fang

Artificial Intelligence Systems Section, CSSB
Telecom Australia Research Labs
Box 249, Clayton Vic. 3168

ABSTRACT - A multilayer neural network for speech recognition is described here. This neural network is trained with the back propagation algorithm. The network can recognize different types of speakers. For multispeaker speech recognition, a 95% rate of correct recognition is achieved.

INTRODUCTION

Many researchers have used neural networks in speech recognition. A few representative systems can be found in (Lippmann 1990, Waibel et al 1989, Kohonen 1988, McDermott et al 1990, Iso and Watanabe 1990). Most of these systems use time delayed neural networks. These networks achieved highly promising results. In this paper, we present preliminary results for multiple speaker speech recognition with multiple neural networks.

Different types of speakers may have different features in their speech signals. Even after normalization, the speech signals of different types of speakers may not be separable. This makes the speech recognition task very difficult for multiple speakers. However, if we can identify the type of a speaker, we can utilize the special features of this type of speaker in recognition. This would help to improve the accuracy of recognition. This approach was also employed in the successful SPHINX system (Lee 1989) in which a number of code books are used for different types of speakers.

We have applied the same idea to speech recognition using neural networks. We have trained our neural networks for different types of speakers. We also trained a neural network for identifying the type of speakers. These networks were then combined to form a speech recognition system. The system identifies the speaker type and outputs the result of the neural network for the corresponding speaker type. Preliminary results show that this approach exhibited better performance than training a single network with patterns from many types of speakers.

NETWORK ORGANIZATION

Our speech recognition system consists of a simple signal processing subsystem and a number of neural networks. The signal processing system digitizes the speech signal and computes the accumulated energy in 20 frequency bands for each word. The outputs are 20 floating point numbers which represent the accumulated energy in the frequency bands. The signal processing system is very primitive. Hence we only conducted experiments on recognizing ten digits spoken by several types of speakers. The purpose is to test the feasibility of multiple speaker speech recognition using multiple neural networks. So this primitive system just managed to meet our simple demand.

The neural network consists of N+2 subnets where N is the number of speaker types. In our present experiment, we choose N=2 which is sufficient to demonstrate the feasibility of our approach. N subnets are trained to recognize the speech by N types of speakers. Another subnet is trained to identify the speaker type. A multiplexer subnet is used to select the output of the subnet according to the speaker type identified. A block diagram of the system is shown in Figure 1.

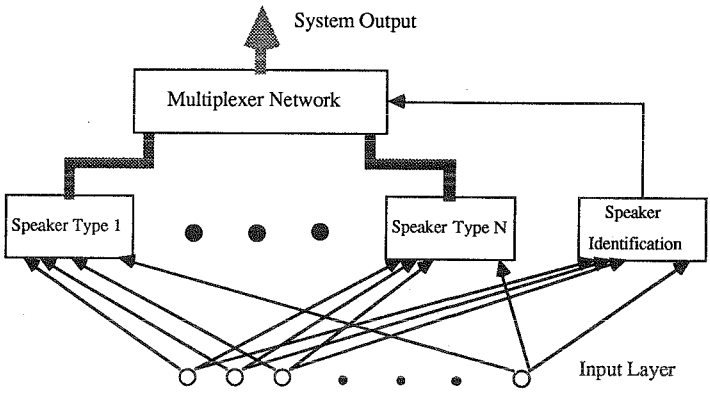

Figure 1. The organization of the speech recognition system.

Each subnet consists of three layers, including an input layer of 20 nodes (which is shared by the N subnets), a hidden layer also of 20 nodes, and an output layer of 10 nodes (each of which represents a digit). In each subnet, every node in the second layer is connected to every node of the input layer. Each node in the output layer is also connected to every node of the second layer. Although two hidden layers are needed to form arbitrary regions for classification, we found that one hidden layer is sufficient for our purpose. A subnet is shown in Figure 2.

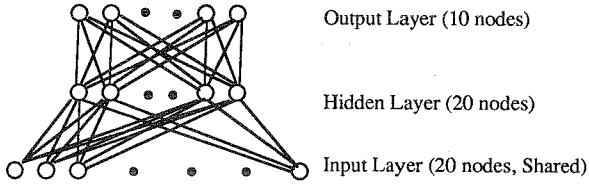

Figure 2. The structure of a subnetwork.

NETWORK TRAINING

The Back Propagation (BP) algorithm was used for training our neural networks. BP uses a gradient descent search to minimize the cost function which is the mean squared error between the desired outputs and the actual outputs. Let w_{ij} be a connection weight between neuron *i* and neuron *j* and let *f* be the cost function. Then the weight is updated according to the learning rule

$$w_{ij}(t+1) = w_{ij}(t) + \eta \Delta_j x_i$$

where η is the learning rate. The weights are updated iteratively by propagating errors backward through layers until they converge at which time the cost function is reduced to an acceptable value.

During the training phase, each subnet is trained by the speech patterns of a specific type of speaker. In our preliminary experiments, we identify two types of speakers, namely, male speakers and female speakers. The sample patterns are spoken by two male speakers and one female speaker. Another neural network is trained to identify the type of the present

speaker. The multiplexer network is directly constructed. These subnets are then connected as shown in Figure 1.

It would take much longer time to train the neural networks if the 200 patterns spoken by each speaker were used in training. Also this would cause the networks to become overtrained. To overcome these problems, we average ten patterns of the same digit from a speaker. This results in a 20 fold reduction of training patterns. Experiments show that the performance does not degrade. The experimental results given in the next section are based on the averaged training patterns.

Since the outputs from the signal processing systems are often large floating point numbers, the neural networks do not learn well on these patterns without preprocessing of the input data. Hence we scale the patterns such that the first 8 frequency bands (which often contain high energy) are scaled down by a factor of 10. The networks can then be trained successfully.

TEST RESULTS

In the experiments performed, two types of speakers are identified. We recorded ten digits spoken by male and female speakers. Each digit was spoken about 30 times. The recording was performed under normal office environment. No sound proof room was used. The data collected are thus noisy. The accumulated energy in the eight low frequency bands is comparatively large. So they were scaled down to 10% of the original values. The multilayer neural networks were correctly trained on the scaled data.

Training would have taken extremely long time if we had used all the sample patterns (nearly 1,000 of them). Instead, we averaged every ten samples of the same digit spoken by the same speaker to produce one training pattern. So the total number of training patterns was reduced to about 100. This decreased significantly the training time. The learning rate used was 0.1 and the momentum term was 0.9.

The results are summarized as follows.

Speaker	Recognition Rate	Training Time
m1	96%	1'12"
m2	97%	1'15"
f1	97%	1'11"
mfs	87%	31'10"
mfm	95%	8'35"
spkid	95%	4'43"

In the above table, m_i stands for the i-th male speaker and f1 stands for the female speaker. *mfs* stands for *male and female speakers with a single multilayer network*. *mfm* stands for *male and female speakers with multiple subnetworks*. *spkid* stands for *speaker identification network*.

CONCLUSIONS

A neural network for multiple speaker speech recognition has been developed. The grouping of many types of speakers and the used of multiple subnetworks for speech recognition can improve the accuracy of recognition according our experiments. Preliminary study has demonstrated the feasibility of this approach.

We have also used ad hoc techniques, such as averaging sample patterns and scaling selected input attributes, to train the multilayer neural network. These techniques have greatly improved the training speed.

ACKNOWLEDGEMENTS

The permission of the Executive General Manager of Telecom Australia Research Labs to publish this paper is acknowledged. The author also wishes to thank the many colleagues at Telecom Research Labs for helpful discussions.

REFERENCES

Iso, Ken-Ichi and Watanabe, T. (1990) *Speaker-Independent Word Recognition Using A Neural Network Prediction Model*, IEEE Int. Conf. Acoust., Speech and Signal Proc., April 1990.

Kohonen, T. (1988) *The Neural Phonetic Typewriter*, IEEE Computer, March 1988.

Lee, K-F, Hon, H-W., Hwang, M-Y., Mahajan, S. and Reddy, R. (1989) *The SPHINX Speech Recognition System*, IEEE Int Conf. Acoust., Speech and Signal Proc. 89, 445-448.

Lippmann, R.P. (1990) *Review of Neural Networks for Speech Recognition*, in Readings in Speech Recognition (eds. A. Waibel and K-F. Lee), (Morgan Kaufman: San Mateo,CA), 374-392.

McDermott, E., Iwamida, H., Katagiri, S. and Tohkura, Y. (1990) *Shift-Tolerant LVQ and Hybrid LVQ-HMM for Phoneme Recognition*, in Readings in Speech Recognition, (eds. A. Waibel and K-F. Lee), (Morgan Kaufmann: San Mateo,CA), 425-438.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang, K.J. (1989) *Phoneme Recognition Using Time-Delay Neural Networks*, IEEE Trans. Acoust., Speech and Signal Proc., March 1989.