International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010
1793-8201

A Hybrid model of Neural Network Approach for Speaker independent Word Recognition

N.Uma Maheswari, A.P.Kabilan, R.Venkatesh

***Abstract*—Speech Recognition by computer is a process where speech signals are automatically converted into the corresponding sequence of words in text. When the training and testing conditions are not similar, statistical speech recognition algorithms suffer from severe degradation in recognition accuracy. So we depend on intelligent and recognizable sounds for common communications. In this research, word inputs are recognized by the system and executed in the form of text corresponding to the input word. In this paper, we propose a hybrid model by using a fully connected hidden layer between the input state nodes and the output. We have proposed a new objective function for the neural network using a combined framework of statistical and neural network based classifiers. We have used the hybrid model of Radial Basis Function and the Pattern Matching method. The system was trained by Indian English word consisting of 50 words uttered by 20 male speakers and 20 female speakers. The test samples comprised 30 words spoken by a different set of 20 male speakers and 20 female speakers. The recognition accuracy is found to be 91% which is well above the previous results.**

***Index Terms*—speech recognition, Intelligent,recognizable sound,hybrid model,neural network, Radial Basis Function, Pattern Matching**

I. Introduction

Automatic speech recognition is a process by which a machine identifies speech. It takes a human utterance as an input and returns a string of words or phrases as output. Recently, Neural network was considered as one of the most successful information processing tools that has been widely used in speech recognition. Word recognition is a process to recognize speech (in the form of word) uttered by a speaker. For human beings, it is a natural and simple task. However, it is an extremely complex and difficult job to make a computer respond to spoken commands. The conventional neural networks of Multi-Layer Perceptron (MLP) type have been increasingly in use for word recognition and also for other speech processing applications. One simple method to perform word recognition is a bottom-up approach, in which different features are extracted from the input speech signal which is in the form of word and then converted into text using neural network approach. Thus, it plays an important role in constructing a powerful word recognition system.

II. System Architecture for Word Recognition Model

Generally, word recognition process contains three steps to process the speech signal which is acoustic processing, feature extraction and ANN frame classification with pattern matching as shown in Figure 1. First, we digitize the speech that we want to recognize. In this paper, we digitize the recorded input words from the speakers and also digital filtering that emphasizing important frequency component in signal. Then we analyze the start-end point that depends the signal of the speeches. The second step is feature extraction where the linear predicted values are extracted.

The ANN frame classification technique is used to identify the output unique vector. Then the pattern matching method is applied for word recognition. Here we use neural network pattern matching method.

1) Data recording and Utterance detection
2) Pre-Filtering(preemphasis,normalization,
3) banding,etc.)
4) Framing and Windowing (chopping the data into a usable format)
5) Filtering (further filtering of each window/frame/freq. band)
6) Neural Network Frame Classification
7) Action (Perform pattern match function to get the recognized pattern)

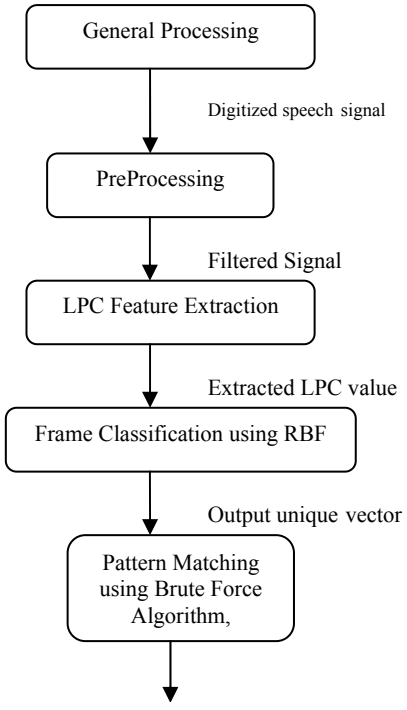

Fig.1.Architecture For Speaker Independent Word Recognition

A. Feature Extraction

For Automatic Speech recognition by computers, feature

vectors are extracted from speech waveforms. A Feature vector is usually computed from a window of speech signals(20~30 ms) in every short time interval(about 10ms). An utterance is represented as a sequence of these feature vectors. In LPC Analysis a speech recognizer is a system that tries to understand or "decode" a digitized speech signal. This signal, as first captured by the microphone, contains information in a form not suitable for pattern recognition. However, it can be represented by a limited set of features relevant for the task. These features more closely describe the variability of the phonemes (such as vowels and consonants) that constitute each word.

The feature measurements of speech signals are typically extracted using one of the following spectral analysis techniques: filter bank analyzer, LPC analysis or discrete Fourier transform analysis. Since LPC is one of the most powerful speech analysis techniques for extracting good quality features and hence encoding the speech signal at a low bit rate, we selected it to extract the features of the speech signal[5].

B. ANN Frame Classification and Pattern Matching

The Hybrid model is a combination of Radial Basis Function approach neural network (frame classification) and the pattern matching method. pattern matching is the act of checking for the presence of the constituents of a given pattern.Brute Force algorithm is used for Patter Matching.

Brute-force pattern matching algorithm

The brute-force pattern matching algorithm compares the pattern **P** with the text **T** for each possible shift of **P** relative to **T**, until either

- a match is found, or
- all placements of the pattern have been tried

Brute-force pattern matching runs in time **O(nm)**

Example of worst case:

- **T = aaa … ah**
- **P = aaah**
- may occur in images and

DNA sequences

- unlikely in English text

The algorithm for **BruteForceMatch(T, P)**is presented as follows:

- **Text T** of size **n** and pattern **P** of size **m is given as input.**
- **The** starting index of a substring of **T** equal to **P** or −1 is obtained as output
- if no such substring exists shift **i** of the pattern is tested from 0 to n-m
- **if j = m,match at I is found**

C. .Radial Basis Function Approach

This neural network is the most powerful pattern classifying one which considered to be separated any pattern by constructing any hyper planes among the different classes of patterns. In RBF initialization of the centers taken place in unsupervised manner looking at the data pattern which I have used is called modified kmeans algorithm. Upon spreading the centers with relevant to the data set then it trained with the supervised manner to mimic the human brain as same as back propagation which is know as extended back propagation variation of the LMS algorithm. Since it uses both the combination of while try to mimic the human brain it also uses statistical approach in the initialization process.

D. Proposed Configuration of Radial Basis Function Approach

The network receives the 40 Boolean values as a 40-element input vector. It is then required to identify the words by responding with a 30-element output vector. The 30 elements of the output vector each represent a word. To operate correctly, the network should respond with a 1 in the position of the word being presented to the network. All other values in the output vector should be 0.

In addition, the network should be able to handle noise. In practice, the network does not receive a perfect Boolean vector as input. Specifically, the network should make as few mistakes as possible when classifying vectors with noise of mean 0 and standard deviation of 0.2 or less.

III. NETWORK ARCHITECTURE

The neural network needs 40 inputs and 30 neurons in its output layer to identify the words. The network is a two-layer `log-sigmoid/log-sigmoid` network. The log-sigmoid transfer function was picked because its output range (0 to 1) is perfect for learning to output Boolean values

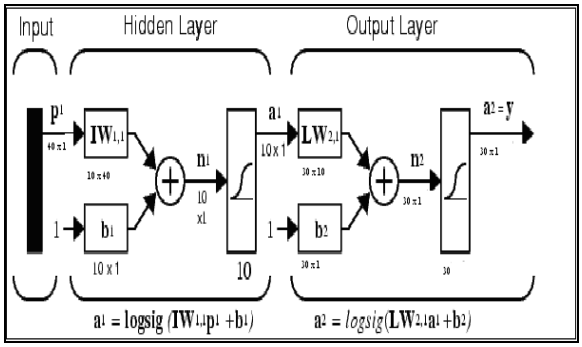

Figure 2 Network Architecture

The network is trained to output a 1 in the correct position of the output vector and to fill the rest of the output vector with 0's. However, noisy input vectors can result in the network's not creating perfect 1's and 0's. After the network is trained the output is passed through the competitive transfer function `compet`. This makes sure that the output corresponding to the letter most like the noisy input vector takes on a value of 1, and all others have a value of 0. The result of this postprocessing is the output that is actually used.

Initialization

Create the two-layer network with newff.

net = newff(alphabet,targets,25);

Training

To create a network that can handle noisy input vectors, it is best to train the network on both ideal and noisy vectors. To do this, the network is first trained on ideal vectors until it has a low sum squared error.

Then the network is trained on 10 sets of ideal and noisy vectors. The network is trained on two copies of the noise-free alphabet at the same time as it is trained on noisy

vectors. The two copies of the noise-free alphabet are used to maintain the network's ability to classify ideal input vectors.

Unfortunately, after the training described above the network might have learned to classify some difficult noisy vectors at the expense of properly classifying a noise-free vector. Therefore, the network is again trained on just ideal vectors. This ensures that the network responds perfectly when presented with an ideal letter. All training is done using backpropagation with both adaptive learning rate and momentum, with the function trainbpx. Then the corresponding output is tested using the obtained output pattern with the stored pattern using pattern matching method and the word output is displayed.

IV. Speaker Independent Word Recognition using Radial Basis Function

A.System Model

The system model for speech recognition is shown below.

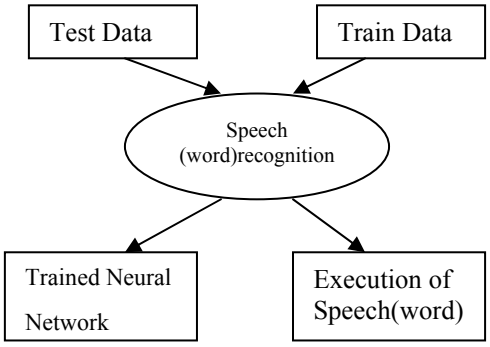

Figure 3 Data Flow Model

V. Working Principle of SPEECH MODEL

In a language, elementary sounds that distinguish meaning are called phonemes. For speech recognition, a language is built upon a set of phonemes and several other types of sounds(noises,etc).

- A lexicon is required for the language, containing the pronunciations in their phonetic form, of atleast all the words used in the training speech database.
- The pronunciation rules generate pronunciations of unknown words,based on their spelling.

The Word Recognition works as

- A frame level classification process transforms the **words** into their frame representation, using ANN frame classification.
- Pattern matching method combines with the neural network and compares the output with the target and produces the output word to be displayed.

VI. Implementation

The speaker independent word recognition is implemented in mathlab, by training the system each 40 samples from different speakers consisting of 50 words each. A test samples taken from a different set of 40 speakers each uttering 30 words. All the samples were of Indian English and recorded by Capturing the speech via sound card via DirectX . A partial output for the word about which assigned the position value of one is given in figure 4.

VII.CONCLUSION

Speech recognition has a big potential in becoming an important factor of interaction between human and computer in the near future. A system has been proposed to combine the advantages of ANN's and Pattern Matching for speaker independent word recognition. Encouraged by the results of the above described experiment, which indicate that global optimization of a hybrid ANN-pattern matching system gives some significant performance benefits. We have seen how such a hybrid system could integrate multiple ANN modules, which may be recurrent. A Neural Network with trained delays and widths and random weights classifies 91% of the wordss correctly. A further refined speech recognition system can improve the accuracy to near 100%.

References

[1] Martin D. Buhmann (2003). Radial Basis Functions: Theory and Implementations. Cambridge University. ISBN 0-521-63338-9.
[2] Yee, Paul V. and Haykin, Simon (2001). Regularized Radial Basis Function Networks: Theory and Applications. John Wiley. ISBN 0-471-35349-3.
[3] D.A.Reynolds,"An Overview of Automatic Speaker Recognition Technology", Proc. ICASSP 2002, Orlando, Florida, pp. 300-304.
[4] A. Biem, S. Katagiri, E. McDermott, and B.-H. Juang, "An application of discriminative feature extraction to filter-bank-based speech recognition," IEEE Trans. Speech Audio Processing, vol. 9, pp.96–110, Jan. 2001.
[5] B. Allen and L.R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis", Proc. IEEE, Vol. 65, No. 11, pp. 1558-1564, 1977
[6] M.R. Portnoff , Short-time Fourier analysis of sampled speech IEEE Trans.Acoust., Speech and Signal Processing, Vol. ASSP-29, pp. 364-373, 1981.
[7] J.S. Lim et al., Signal estimation from modified short-time Fourier transforms, IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-32, pp. 236-243
[8] R.Kronland-Martinet, J.Morlet and A.Grossman, Analysis of sound patterns through wavelet transformation, International Journal of Pattern Recognition and Artificial Intelligence, Vol.1(2)
[9] Marcus E. Hennecke, K. Venkatesh Prasad, and David G. Stork, Using deformable templates to infer visual speech dynamics, 28th Annual Asilomar Conference on Signals, Systems, and Computers volume 1, pages 578582, Pacific Grove, CA, November 1994 IEEE, IEEE Computer Society Press
[10] Alan L. Yuille, David S. Cohen, and Peter W. Hallinan. Facial feature extraction by deformable templates Technical Report 88-2, Harvard Robotics Laboratory, 1988.
[11] H. Sakoe and S. Chiba, Dynamic programming optimization for spokenword recognition, Proceedings of ICASSP-78, vol. 26, no. 1, pp. 43-49, 1997.
[12] Rabi G.and Lu S., (1997). "Visual Speech Recognition by Recurrent Neural Networks." Electrical and Computer Engineering, IEEE 1997 Canadian Conference on, Vol. 1, 25-28 May, Page(s): 55 -58.
[13] Bourlard, H., Kamp, Y., Ney, H., and Wellekens, C. J. (1985). "Speaker-Dependent Connected Speech Recognition Via Dy- namic Programming and Statistical Methods," in Speech and Speaker Recognition, ed. M. R. Schroeder, Basel, Switzerland: Karger, pp. 115-148.

Umamaheswari .N received her M.E in Computer Science & Engineering from the Madras University, Chennai, India in 2002. Currently, She is working as an Assistant Professor in the Department of Computer Science & Engineering at the P.S.N.A. College of Engineering & Technology, Dindigul, India.Her current research interests include Artificial Intelligence, Speech Processing,

Neural Networks and Soft Computing

Kabilan A.P completed his 5- year Integrated M.S. in communication engineering in 1976, 2-year Advanced course in microwave engineering in 1978 and Ph.D. in Microwave engineering ,in 1981 all at Patrice Lumba University , Moscow. He has worked as faculty at Anna university , Chennai, I.I.T Bombay, R.A.I.T., Bombay Panimalar Engineering College, Chennai and P.S.N.A. engineering college, Dindigul, before joining B.I.T., Sathyamangalam in June,2005 as professor and Head, dept. of ECE. Currently he is working as Professor and Principal in Chettinad College of Engineering and Technology, Karur, Tamilnadu, India. He has published 8 papers in International journals, one paper in national journal and 18 papers international conferences and one paper in national conference. He is a recognized Ph.D. supervisor of Anna University and is presently guiding 7 Ph.D scholars is the areas of smart Antenna, Signal processing,Speech Processing Wireless Technology and virtual Instrumentation

Venkatesh .R received his ME in Computer Science and Engineering from Anna University Chennai in India , in 2007. He is working as an Assistant Professor in the department of Information Technology in RVS College of Engineering & Technology, Dindigul, in India..His current research interests include artificial intelligence , Neural Network , soft computing and Networks.

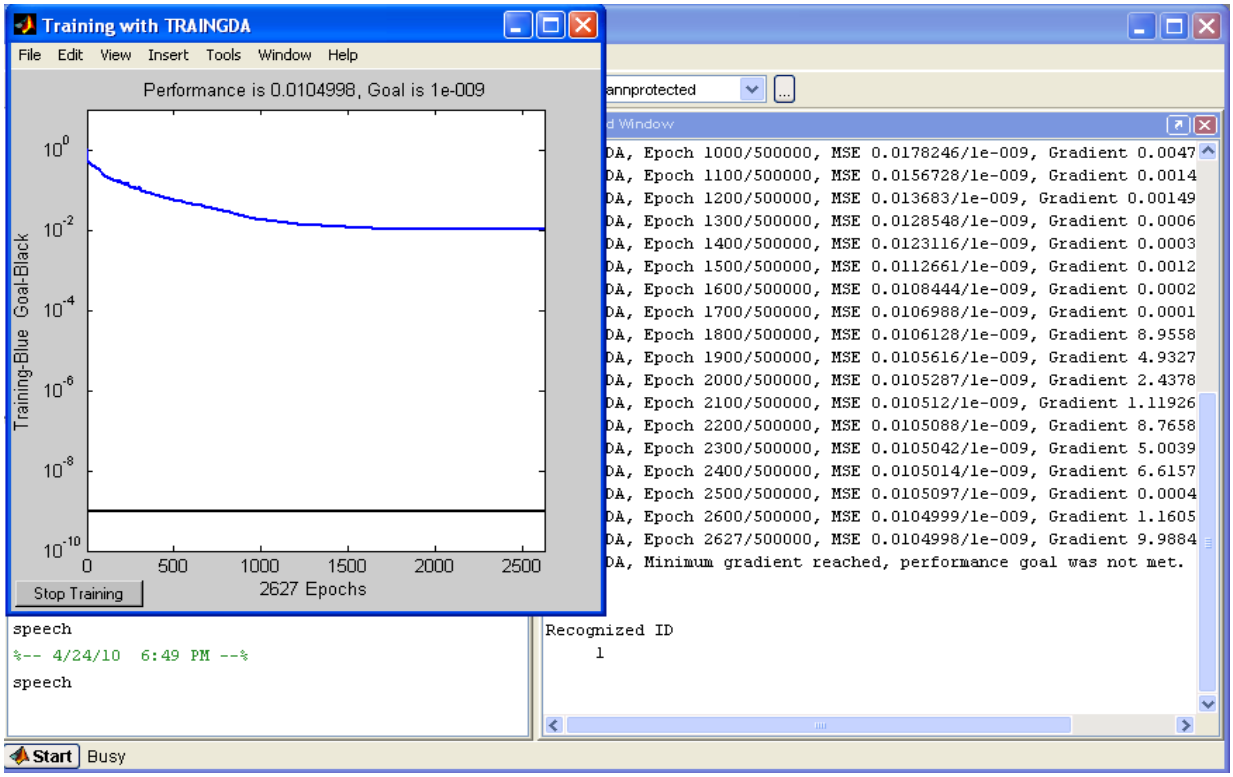

Figure 4.Implemented Result