SPEECH RECOGNITION USING RADIAL BASIS FUNCTION NEURAL NETWORK

Dr.R.L.K.Venkateswarlu
Professor and Head, Department of Information Technology
Sasi Institute of Technology and Engineering, Tadepalligudem, INDIA.
rlk.maths@gmail.com

Dr. R. Vasantha Kumari
Principal
Perunthalaivar Kamarajar Arts College,PUDUCHERRY 605 107
Vasunthara1@gmail.com

G.Vani Jayasri
Lecturer, Department of Information Technology
Sasi Institute of Technology & Engineering, Tadepalligudem, INDIA.
Vani.jsri@gmail.com

Abstract-**In this paper a novel approach for implementing isolated speech recognition is studied. While most of the literature on speech recognition (SR) is based on hidden Markov model (HMM), the present system is implemented by Radial Basis Function type neural network. The two phases of training and testing in a Radial Basis Function type neural network has been described. All of classifiers use Linear Predictive Cepstral Coefficients. It is found that the performance of Radial Basis Function type neural networks is superior to the other classifier Multilayer Perceptron Neural Networks. The promising results obtained through this design show that this new neural networks approach can compete with the traditional speech recognition approaches. Promising results are obtained both in the training and testing phases due to the exploitation of discriminative information with neural networks. It is found that RBF trains and tests faster than MLP.**

Keywords: Radial Basis Function Neural Network, Multi-Layer Perceptron, Linear predictive cepstral coefficient, Classifiers, Performace.

I. INTRODUCTION

Automatic speech recognition has been an active research topic for more than five decades. With the advent of digital computing and signal processing, the problem of speech recognition was clearly posed and thoroughly studied. These developments were complemented with an increased awareness of the advantages of conversational systems. The range of the possible applications is wide and includes: voice-controlled appliances, fully featured speech-to-text software, automation of operator-assisted services, and voice recognition aids for the handicapped.

Different approaches in speech recognition have been adopted. They can be divided mainly into four trends: The acoustic-phonetic approach, The pattern recognition approach, The artificial intelligence approach and Neural networks for ASR. Hidden Markov Model (HMM) and Gaussian Mixture Models.

Speech recognition has a big potential in becoming an important factor of interaction between human and computer in the near future. A successful speech recognition system has to determine features not only present in the input pattern at one point in time, but also features of the input pattern changing over time [6, 5]. The classic methods based on multilayer perceptron use the Time Delay Neural Network, it is the first model used by Weibel in the speech recognition domain [6]. But the problem was the hard time processing and the adjustment of parameters that become a laborious stain for the new applications. In the opposite, the RBF networks don't require a special adjustment and the training time becomes shorter with regard to the Time Delay Neural Network. But the problem of RBF is the shift invariant in time [6].

The NN approach for SR can be divided into two main categories: conventional neural networks and recurrent neural networks. RBF neural network is becoming an increasingly popular neural network with diverse applications and is probably the main rival to the multi-layered perceptron. Much of the inspiration for RBF networks has come from traditional statistical pattern classification techniques. The unique feature of the RBF network is the process performed in the hidden layer. The idea is that the patterns in the input space form clusters. If the centres of these clusters are known, then the distance from the cluster centre can be measured. Furthermore, this distance measure is made non-linear, so that if a pattern is in an area that is close to a cluster centre it gives a value close to 1. Statistical feed-forward networks such as the RBF network are serious rivals to the MLP. Learning mechanisms in statistical neural networks are not biologically plausible –

978-1-4244-8679-3/11/$26.00 ©2011 IEEE

so have not been taken up by those researchers who insist on biological analogies.

This is becoming an increasingly popular neural network with diverse applications and is probably the main rival to the multi-layered perceptron .Much of the inspiration for RBF networks has come from traditional statistical pattern classification techniques.

II. SYSTEM CONCEPT

2.1. Dataset

The vocabulary set is composed of six words: “passion”, ”galaxy”, “marvellous”, “manifestation”, “almighty”, “pardon”. 6 different speakers (2 Male and 4 Female) are allowed to utter the above words, for uttering each word six times and the speech databases were recorded in wave files. So there are 216 wave files. Each of these wave files are trained and tested.

2.2 Preprocessing

The speech signals are recorded in a low noise environment with good quality recording equipment. The signals are samples at 11 kHz. Reasonable results can be achieved in isolated digit recognition when the input data is surrounded by silence.

2.3 Sampling Rate

150 samples are chosen with sampling rate 11 kHz, which is adequate to represent all speech sounds.

2.4 Windowing

In order to avoid discontinuities at the end of speech segments the signal should be tapered to zero or near zero and hence reduce the mismatch.

III. FEATURE EXTRACTION

The goal of feature extraction is to represent speech signal by a finite number of measures of the signal. This is because the entirety of the information in the acoustic signal is too much to process, and not all of the information is relevant for specific tasks. In present Speech Recognition systems, the approach of feature extraction has generally been to find a representation that is relatively stable for different examples of the same speech sound, despite differences in the speaker or various environmental characteristics, while keeping the part that represents the message in the speech signal relatively intact. Linear predictive coding (LPC) is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive mode. It is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters. LPC analyzes the speech signal by estimating the formats, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered modeled signal is called the reside. The number which describe the intensity and frequency of the buzz, the formants, and the reside signal, can be stored or transmitted somewhere else. LPC synthesizes the speech signal by reversing the process: use the buzz parameters and the residue to create a source signal. Use the formants to create a filter (which represents the tube), and run the sources through the filter, resulting in speech. Because speech signals vary with time, this process is done on short chunks of the speech signal, which are called frames; generally 30 to 50 frames per second give intelligible speech with good compression.

LPC is frequently used for transmitting spectral envelope information, and as such it has to be tolerant of transmission errors. Transmission of the filter coefficients directly is undesirable, since they are very sensitive to errors. In other words, a very small error can distort the whole spectrum, or worse, a small error might make the prediction filter unstable.

LPC is generally used for speech analysis and resynthesis. It is used as a form of voice compression by phone companies, for example in the GSM standard. It is also used for secure wireless, where voice must be digitized, encrypted and sent over a narrow voice channel.

In the LPC analysis one tries to predict x_n on the basis of the p previous samples,

$$x_n^{'} = \sum a_k \, x_{n-k}$$

Then $\{a_1, a_2, \ldots\ldots\ldots, a_p\}$ can be chosen to minimize the prediction power Q_p where

$$Q_p = E\left[\left| x_n - x_n^{'} \right|^2 \right]$$

Linear Predictive Coding is used to extract the LPCC coefficients from the speech tokens. The LPCC coefficients are then converted to cepstral coefficients. The cepstral coefficients are normalized in between -1 and 1. The speech is blocked into overlapping frames of 20ms every 10ms using Hamming window. LPCC was implemented using the autocorrelation method. A drawback of LPCC estimates is their high sensitivity to quantization noise. Convert LPCC coefficients into cepstral coefficients where the cepstral order is the LPCC order and to decrease the sensitivity of high and low-order cepstral coefficients to noise, the obtained cepstral coefficients are then weighted.

16 Linear Predictive Cepstral Coefficients are considered for windowing. Linear Predictive Coding analysis of speech is based on human perception experiments. Sample the signal with 11 kHz. Number of frames are obtained for each utterance from LPC coefficients.

IV. RECOGNITION METHODOLOGY

In multi-class mode such as the present case, each classifier tries to identify whether the set of input feature vectors, derived from the current signal, belongs to a specific class of numbers or not, and to which class exactly. For samples that can not be realized as a specific class a random class is selected.

V. CLASSIFIERS

Several classifiers are tested for mentioned dataset. The structures of successful classifiers in recognition are described in following subsections.

5.1. Multi-Layer Perceptron

This is perhaps the most popular network architecture in use today, due originally to Rumelhart and McClelland (1986). The units each performed a biased weighted sum of their inputs and pass this activation level through a transfer function to produce their output, and the units are arranged in a layered feedforward topology. The network thus has a simple interpretation as a form of input-output model, with the weights and thresholds (biases) the free parameters of the model. Such networks can model functions of almost arbitrary complexity, with the number of layers, and the number of units in each layer, determining the function complexity. Important issues in MLP design include specification of the number of hidden layers and the number of units in these layers.

The number of input and output units is defined by the problem (there may be some uncertainty about precisely which inputs to use, a point to which we will return later. However, for the moment we will assume that the input variables are intuitively selected and are all meaningful). The number of hidden units to use is far from clear. As good a starting point as any is to use one hidden layer, with the number of units equal to half the sum of the number of input and output units. Again, we will discuss how to choose a sensible number later.

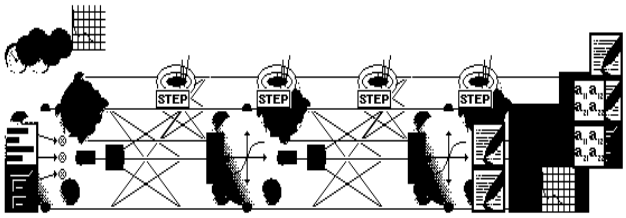

Figure 1: MLP Network architecture with step learning rule.

This network has an input layer (on the left) with three neurons, one hidden layer (in the middle) with three neurons and an output layer (on the right) with three neurons.

There is one neuron in the input layer for each predictor variable. In the case of categorical variables, N-1 neurons are used to represent the N categories of the variable.

Input Layer — A vector of predictor variable values (x1...xp) is presented to the input layer. The input layer (or processing before the input layer) standardizes these values so that the range of each variable is -1 to 1. The input layer distributes the values to each of the neurons in the hidden layer. In addition to the predictor variables, there is a constant input of 1.0, called the bias that is fed to each of the hidden layers; the bias is multiplied by a weight and added to the sum going into the neuron.

Hidden Layer — Arriving at a neuron in the hidden layer, the value from each input neuron is multiplied by a weight (wji), and the resulting weighted values are added together producing a combined value uj. The weighted sum (uj) is fed into a transfer function, σ, which outputs a value hj. The outputs from the hidden layer are distributed to the output layer.

Output Layer — Arriving at a neuron in the output layer, the value from each hidden layer neuron is multiplied by a weight (wkj), and the resulting weighted values are added together producing a combined value vj. The weighted sum (vj) is fed into a transfer function, σ, which outputs a value yk. The y values are the outputs of the network.

If a regression analysis is being performed with a continuous target variable, then there is a single neuron in the output layer, and it generates a single y value. For classification problems with categorical target variables, there are N neurons in the output layer producing N values, one for each of the N categories of the target variable.

5.2 Radial Basis Neural Networks

The core of a speech recognition system is the recognition engine. The one chosen in the paper is the Radial Basis Function Neural Network (RBF). This is a static two neuron layers feed forward network with the first layer L1, called the hidden layer and the second layer, L2, called the output layer. L1 consists of kernel nodes that compute a localized and radially symmetric basis functions.

The pattern recognition approach avoids explicit segmentation and labeling of speech. Instead, the recognizer used the patterns directly. It is based on comparing a given speech pattern with previously stored ones. The way speech patterns are formulated in the reference database affects the performance of the recognizer. In general, there are two common representations,

The output y of an input vector x to a (RBF) neural network with H nodes in the hidden layer is governed by:

$$y = \sum_{h=0}^{H-1} w_h \, \phi_h (x)$$

Where w_h are linear weights ϕ_h are the radial symmetric basis functions. Each one of these functions is characterized by its center c_h and by its spread or width σ_h. The range of each of these functions is [0, 1].

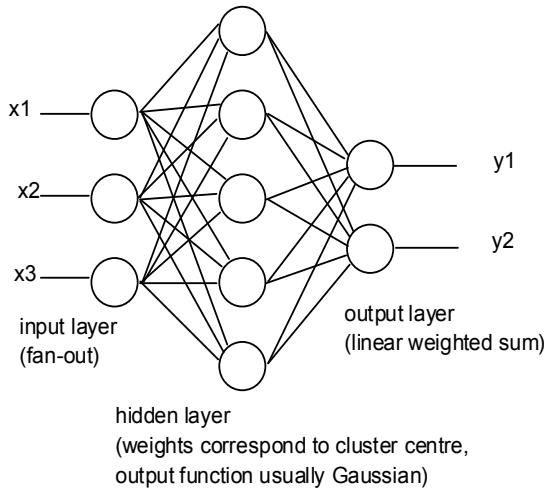

Figure: 2 Radial Basis Function Neural Network Architecture

Once the input vector x is presented to the network, each neuron in the layer L1 will output a values according to how close the input vector is to its weight vector. The more similar the input is to the neuron's weight vector, the closer to 1 is the neuron's output and vice versa. If a neuron has an output 1, then its output weights in the second layer L2 pass their values to the neurons of L2 . The similarity between the input and the weights is usually measured by a basis function in the hidden nodes. One popular such function is the Gaussian function that uses the Euclidean norm. It measures the distance between the input vector x and the node center c_h . It is defined as:

$$\phi_h = \exp\left(\| x - c_h \| / 2\sigma_h^2\right)$$

VI. TRAINING PHASE

The networks are usually trained to perform tasks such as pattern recognition, decision-making, and motory control. The original idea was to teach them to process speech or vision, similarly to the tasks of the human brain. Nowadays tasks such as optimization and function approximation are common. Training of the units is accomplished by adjusting the weight and threshold to achieve a classification. The adjustment is handled with a learning rule from which a training algorithm for a specific task can be derived. The Multilayer Perceptron and Radial Basis Function Neural Networks are trained for spoken words for 6 speakers. The learning rate is taken as 0.01, momentum rate is taken as 0.3. Number of epochs are taken as 100. The Random Gaussian Method is chosen for initialization.

6.1 Performance Evaluation

The performance for MLP classifier and RBF classifier for each speaker have been computed and presented in Tables 1 and 2 respectively. The overall performance average for both classifiers MLP and RBF have been computed and presented in Table 3.

Table 1: RESULTS for Training MLP (%)

	Passion	Galaxy	Marvellous	Manifestation	Almighty	Pardon
Speaker1	95%	96%	98%	88%	95%	97%
Speaker2	97%	96%	97%	92%	94%	98%
Speaker3	96%	98%	96%	97%	95%	96%
Speaker4	97%	95%	97%	87%	94%	97%
Speaker5	95%	95%	97%	98%	96%	89%
Speaker6	95%	96%	98%	97%	97%	96%

Table 2: Results for Training RBF (%)

	Passion	Galaxy	Marvellous	Manifestation	Almighty	Pardon
Speaker1	98%	98%	99%	97%	99%	99%
Speaker2	99%	99%	99%	99%	99%	99%
Speaker3	99%	98%	99%	99%	98%	99%
Speaker4	99%	99%	99%	99%	97%	99%
Speaker5	98%	99%	98%	97%	98%	95%
Speaker6	97%	98%	98%	97%	98%	96%

Table 3: Overall performance average

Classifier	Overall Performance average
MLP	95.47%
RBF	98.21%

VII. TESTING PHASE

The same Multilayer Perceptron and Radial Basis Function Neural Networks are trained for spoken digits for 6 speakers. The learning rate, momentum rate and the number of epochs chosen are same as in the training phase. The initialization chosen is also same as that of training phase.

7.1 Performance Evaluation

The performance for MLP classifier and RBF classifier for each speaker have been computed and presented in Tables 4 and 5 respectively. The overall performance average for both classifiers MLP and RBF have been computed and presented in Table 6.

Table 4: Results for Testing MLP (%)

	Passion	Galaxy	Marvellous	Manifestaion	Almighty	Pardon
Speaker1	98%	97%	97%	98%	88%	98%
Speaker2	97%	97%	97%	97%	96%	88%
Speaker3	96%	96%	98%	96%	97%	89%
Speaker4	98%	99%	96%	97%	98%	97%
Speaker5	97%	95%	97%	97%	97%	89%
Speaker6	96%	95%	98%	98%	95%	98%

Table 5: Results for Testing RBF (%)

	Passion	Galaxy	Marvellous	Manifestaion	Almighty	Pardon
Speaker 1	100 %	99%	99%	99%	100%	100%
Speaker 2	99%	100%	98%	99%	99%	100%
Speaker 3	98%	99%	99%	100%	99%	99%
Speaker 4	99%	99%	99%	99%	98%	99%
Speaker 5	99%	99%	98%	99%	97%	96%
Speaker 6	97%	98%	98%	98%	98%	97%

Table 6: Overall performance average.

Classifier	Overall Performance average
MLP	96 %
RBF	98.69%

VIII. CONCLUSION

The Radial Basis Function Neural Network architecture has been shown to be suitable for the recognition of isolated words. Recognition of the words is carried out in speaker dependent mode. In this mode the tested data presented to the network are same as the trained data. The 16 Linear Predictive Cepstral Coefficients with 16 parameters from each frame improves a good feature extraction method for the spoken words, since the first 16 in the cepstrum represent most of the formant information. It is found that the performance of RBF classifier is superior to MLP classifier. It is found that speaker 6 average performance is the best performance in training MLP classifier and speaker 2 average performance is the best performance in training RBF classifier. It is found that average speaker 4 performance is the best performance in testing MLP classifier and speaker 1 average performance is the best performance in testing RBF classifier.

REFERENCES

[1] Al-Alaoui, M.A., Mouci, R., Mansour M.M., Ferzli, R., A Cloning Approach to Classifier Training, IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans, vol.32, no.6, pp.746- 752, (2002)

[2] Picton, P. Neural Networks, Palgrave, NY (2000)

[3] Tan Lee, P. C. Ching, L.W. Chan, Isolated Word Recognition Using Modular Recurrent Neural Networks, Pattern Recognition, vol. 31, no. 6, pp. 751- 760 (1998)

[4] Gurney, K., An Introduction to Neural Networks, UCL Press, University of Sheffield (1997).

[5] Benyettou, A., Acoustic Phonetic Recognition in the Arabex System. Int. Work Shop on Robot and Human Communication, ATIP95.44, Japan. 1995.

[6] Berthold, M.R., A Time Delay Radial Basis Function for Phoneme Recognition. Proc. Int. Conf. on Neural Network, Orlando, USA. 1994

[7] Rabiner, L. and Juang, B. -H., Fundamentals of Speech Recognition, PTR Prentice Hall, San Francisco, NJ (1993).

[8] N Kandil, V K Sood, K Khorasani and R V Patel, Fault identification in an AC–DC transmission system using neural networks, IEEE Transaction on Power System, 7(2):812–9, 1992.

[9] Morgan, D. and Scolfield, C., Neural Networks and Speech Processing, Kluwer Academic Publishers (1991).

[10] D C Park, M A El-Sharakawi and Ri Marks II, Electric load forecasting using artificial neural networks, IEEE Trans Power System, 6(2), pp 442–449, 1991.