2011 2nd International Conference on Networking and Information Technology
IPCSIT vol.17 (2011) © (2011) IACSIT Press, Singapore

A Novel Isolated Speech Recognition Method based on Neural Network

Fu Guojiang

Information and Control Engineering Institute, Shenyang Jianzhu University, Shenyang,Liaoning, China ,110168,

fuguojiang@guigu.org

Abstract. 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 performances 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 performances is the best performance in testing MLP classifier and speaker 1 average performance is the best performance in testing RBF classifier.

IndexTerms—*Neural Network, Speech Recognition,Multi- Layer* Perceptron

1. Introduction

Automatic speech recognition has been an active research topic for 50 years. Along with the digital computing and signal processing, speech recognition problem, puts forward the comparison system of further research, clear. Possible application range, including:-controlled electrical appliances, on fully featured have to-text software,-automatic operator-assisted service, and voice recognition AIDS for the handicapped.

They mainly divided into four trends: acoustic-phonetic method, pattern recognition method, artificial intelligence method and neural network to realize.

Speech recognition has a great potential, to become an important factor of interaction between the human and computer in the near future. The success of a speech recognition system not only must determine the characteristics exist in input mode in a point in time, but also has input mode changing over time [6, article 5]. On the contrary, RBF networks without the need for special adjustment and the training time becomes short about delay neural network.

2. System Concept

2.1. Dataset

The vocabulary set is composed of six words: “passion”, “galaxy”, “marvellous”, “manifestation”, “almighty”, and “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.

3. Feature Extraction

The goal is for feature extraction of speech signal through the finite number of measures signal. This is because of the whole information acoustic signal is the process of too many, but not all of the information is about a specific task. In the present speech recognition system, feature extraction method are usually find a relatively stable said different example, although such a speech sound differences or different environment characteristics and, at the same time, the spokesman for part of the relatively complete information on behalf of speech signal. Linear forecast coding (LPC) is a tool is mainly used for in the audio signal processing and speech processing, the spectrum of the envelope on behalf of a digital signal compressed form of speech, by using the linear forecasting information model. This is one of the most powerful voice analysis techniques, and one of the most powerful methods for quality good speech coding in low bit rate and gave a very accurate estimate parameters of the lecture. Analyzed the LPC speech signal through the forecast, and eliminate the effects of the format, from the speech signal intensity and frequency estimation of the rest of the buzz. The process of removing the filter is called the resonant; the rest of the filtered signal after subtraction analog signal is called the living. Description of the digital frequency and intensity buzz and the resonant signals can live and stored or transmitted to other places. LPC comprehensive speech signal process: reversing with buzz parameters and residual create a source signal. Use the resonant to create a filter (tube), represents the source and run through the filter, lead to the speech. Because speech signal, in the process, with time on doing a lot of short speech signal, called framework; General 30 to 50 frames per second to understand speech has better compression.

The spread directly filter coefficients is not recommended, because they are very sensitive mistakes. In other words, a very small error may distort the whole spectrum, or worse, a small mistake may make the prediction filter does not stable.

Usually used for voice LPC is analysis and secondary synthesis. As a kind of speech compress telephone companies, such as in the GSM standards. It can also be used to secure wireless, where the voice must be digital, encryption and transmit speech narrow 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 $\{a_1, a_2, \cdots, a_p\}$ 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 is obtained for each utterance from LPC coefficients.

4. Recognition Methodology

In the present case as model, each classifier trying to determine the set characteristic vector and input from current signal, belong to a specific type of digital or incomplete, which class. As a sample, not as a specific professional class is a random choice.

5. 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 structure, due to the use of the original Rumelhart today and McClelland, (1986). Each a biased unit weighted and their input and through this transfer function through the activation level and the production, this unit are arranged in a layered topology. Fed has a simple network, so understanding for a kind of input-output model, with weight and threshold (bias) free parameter model. This kind of network model function can be almost any complexity, number of layers, and the number of conveying unit, to determine the function of each layer of complexity. Simple design issues including specifications of the number of hidden layers and the number of units in these layer.

The number of input and output are defined as the problem of unit there may be some don't determine the precise production elements of the use, a point, we will put the later. However, at present, we assume that the input variable selection and are meaningful intuitive). Some hidden units use unclear. A good starting point, what is the use of a hidden layer of network node number of units, equal to half the amount of money of the input and output unit. Again, we will discuss how to choose a wise after the number.

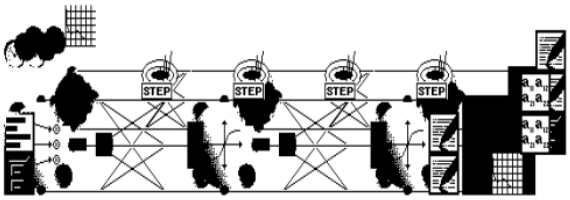

Figure 1. MLP Network architecture with step learning rule.

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, 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 radically 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].

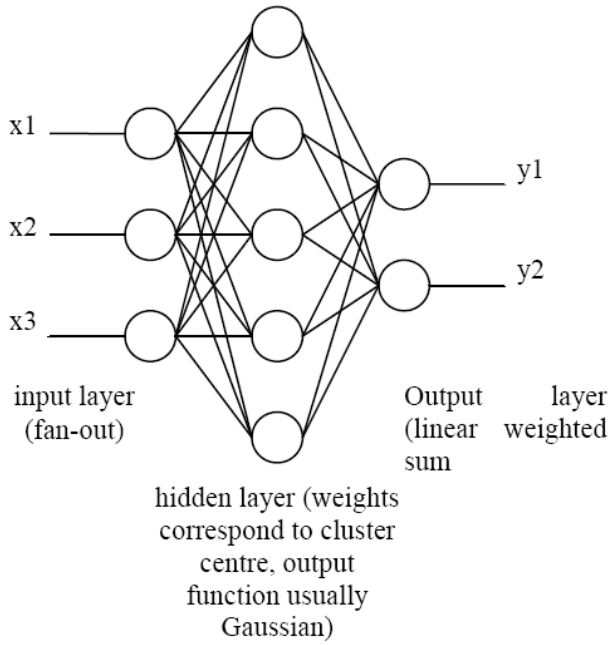

Figure 2. Radial Basis Function Neural Network Architecture

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:

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

6. 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 is 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 I. RESULTS FOR TRAINING MLP (%)

	Passion	Galaxy	Marvellous	Manifest-aion	Almighty	Pardon
Speaker1	95%	96%	98%	88%	95%	97%
Speaker2	97%	96%	97%	92%	97%	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 II. RESULTS FOR TRAINING RBF (%)

	Passion	Galaxy	Marvellous	Manifest-aion	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 III. OVERALL PERFORMANCE AVERAGE

Classifier	Overall Performance average
MLP	95.47%
RBF	98.21%

7. 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 IV. RESULTS FOR TESTING MLP (%)

	Passion	Galaxy	Marvellous	Manifest-aion	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 V. RESULTS FOR TESTING RBF (%)

	Passion	Galaxy	Marvellous	Manifest-aion	Almighty	Pardon
Speaker1	100%	99%	99%	99%	100%	100%
Speaker2	99%	100%	98%	99%	99%	100%
Speaker3	98%	99%	99%	100%	99%	99%
Speaker4	99%	99%	99%	99%	98%	99%
Speaker5	99%	99%	98%	99%	97%	96%
Speaker6	97%	98%	98%	98%	98%	97%

TABLE VI. OVERALL PERFORMANCE AVERAGE.

Classifier	Overall Performance average
MLP	96%
RBF	98.69%

8. Conclusion

RBF neural network to become an increasingly popular neural network and the different application, is likely to be the primary competitors, multi-layer perceptron. Most of the inspiration comes from traditional RBF network the statistical pattern recognition technology. The unique feature is a process of radial basis function neural network of hidden layer. The idea is that the input space patterns. If these clusters of cluster center are known, so cluster centre distance can be measured. In addition, the distance measure is nonlinear, so that if a pattern is a close to cluster centre it provides a value close to 1. In statistical neural network learning mechanism are not biologically plausible--not occupied the researchers insist on biological analogy.

This is to become an increasingly popular neural network and the different application, is likely to be the main rival multi-layer perceptron RBF network.

9. 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.