IJCSC International Journal of Computer Science and Communication Vol. 2, No. 1, January-June 2011, pp. 53–57

# NEURO BASED APPROACH FOR SPEECH RECOGNITION BY USING MEL-FREQUENCY CEPSTRAL COEFFICIENTS

R.L.K. Venkateswarlu[1] and R. Vasanthakumari[2]

[1] Department of Information Technology, Sasi Institute of Technology and Engineering, Tadepalligudem, India, *E-mail: rlk.maths@gmail.com*.

[2] Perunthalaivar Kamarajar Arts College, Puducherry-605107, India, *E-mail: vasunthara1@gmail.com.*

ABSTRACT

This paper presents continuous speech recognition system based on neural network concept. Features are extracted and the data is compressed using Mel-frequency Cepstral coefficients method. These Mel-frequency Cepstral coefficients are used as inputs to train neural networks. Neural networks are useful to solve complex problems which do not require accurate solution. The backpropagation algorithm is used in multilayer perceptron. The solution found in this approach is convergent. This research work is aimed at speech recognition using multilayer perceptron neural networks. A small vocabulary of 11 words were established first, these words are upload, search, browse, import, export, send, remove, attach, help, format, install. These chosen words involved with executing some computer functions such as export a file or an image; find a file or a folder or a image; to view the data; download some properties; transfer out of a database or document in a format; to move a file; to delete a file; to add a file; some assistance; to delete existing content; to add new software. Are introduced to the computer and then subjected to feature extraction process using Mel-frequency cepstral coefficients. These features are used as input to an artificial neural network in speaker dependent mode. Half of the words are used for training the artificial neural network and the other half are used for testing the system. The system components consist of three parts, speech processing, feature extraction, training and testing by using neural networks and information retrieval. The retrieve process proved to be 81.44% – 93.18% successful, which is quite acceptable, considering the variation to surroundings, state of the person, and the microphone type.

*Keywords:* Feature Extraction, Mel-frequency Cepstral Coefficients, Multilayer Perceptron, Neural Network, Speech Recognition, Frames.

## 1. INTRODUCTION

Speech Recognition has been a goal of research for more than five decades. A variety of knowledge sources need to be established in the Artificial Intelligence approach to speech recognition. Therefore, two key concepts of artificial intelligence are automatic knowledge acquisition (learning) and adaptation. One way in which these concepts have been implemented is via the neural network approach. Speech processing is the transmission and reception of the information which was conveyed by speech. Speech conveys three different kinds of information simultaneously. The most important of these is the linguistic information, which means the semantics of the utterance. With the growth of the digitalization, speech can be used as a input acquisition for front-end analysis, feature extraction, local matching, global decodes and language model. Speech recognition is the act of converting acoustic signal captured by microphone or a telephone to a set of words. These recognized words are used for controlling the digital computer with commands to enter the data into digital computer to prepare a document or to retrieve the data. Ron cole and victor zue [1998] have studied about the aforesaid. J.C. Siemon [1980] has studied speech recognition systems by pattern matching approach. Neural networks have been considered for speaker recognition applications for the past ten years. For speaker verification, the neural network based modeling approaches include multilayer perceptrons [9, 5], radial basis functions [15], predictive neural networks [10], neural tree networks [8], and recurrent neural networks [11].

## 2. PRESENT WORK

The objective of speech recognition is to determine the sequence of sound units from the speech signal so that the linguistic message in the form of text can be decoded from the speech signal. The steps used in the present speech recognition system are discussed below.

### 2.1 Input Acquisation

After capturing the speech by using microphone the speech data is saved in.wav files. The speech data is converted to analog signal by using Praat object software tool. The signal is then converted into mono speech signal with 11 kHz.

### 2.2 Front–End Analysis

The acoustic speech signal exists as pressure variations in the air. The micro phone converts these pressure variations into an electric current that is related to the pressure. The ear converts these pressure variations into a series of nerve impulses that are transmitted to the brain. Selection of features is very important for speech recognition task. Good features are required to achieve good result for recognition. Basic problem with speech recognition is identification of proper features for speech recognition task, and a strategy to extract these features from speech signal.

### 2.3 The Speech Utterance (Data Collection)

The source of data is a database consists of 11 words spoken 8 times by 4 speakers; those are 3 males and 1 female of various ages. The data, which is speaker dependent, will be used for training and testing phases. In speaker dependent form, the first four utterances of each of the 11 words spoken by every speaker are used to train the network and the remaining utterances are used to test the network. Therefore, the speech database contains 176 utterances, which can be used for training the network, and 176 utterances, which are available for testing. These words are recorded by: 1. Using Praat Object Software with sampling rate 11 kHz, 8-bit and mono is used to record the utterance. 2. In a closed room, the same microphone is used to record the spoken words. 3. The files are saved in a .wav format.

### 2.4 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 word recognition when the input data is surrounded by silence.

### 2.5 Sampling Rate

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

### 2.6 Windowing

Inorder to avoid discontinuities at the end of speech segments the signal should be tapered to zero or near zero and hence reduce the mismatch. To the given 12 Mel frequency coefficients, and for time 0.005 seconds, a window length of 0.015 is selected by the Praat Object software tool.

### 2.7 Feature Extraction

Feature extraction consists of computing representations of the speech signal that are robust to acoustic variation but sensitive to linguistic content. The Mel-filter is used to find band filtering in the frequency domain with a bank of filters. The filter functions used are triangular in shape on a curvear frequency scale. The filter function depends on three parameters: the lower frequency, the central frequency and higher frequency. On a Mel scale the distances between the lower and the central frequencies and that of the higher and the central frequencies are equal. The filter functions are:

$$H(f) = 0 \; for \; f \le f_l \; and \; f \ge f_h$$

$$H(f) = (f - f_l)/(f_c - f_l) \; for \; f_l \le f \le f_c$$

$$H(f) = (f_h - f)/(f_h - f_c) \; for \; f_c \le f \le f_h$$

Mel frequency cepstral coefficients are found from the Discrete Cosine Transform of the Filter bank spectrum by using the formula given by Davis and Mermelstein [1980].

$$c_i = \sum_{j=1}^{N} P_j \cos(i\pi / N(j - 0.5))),$$

$P_j$ denotes the power in dB in the $j$th filter and $N$ denotes number of samples.

12 Mel frequency coefficients are considered for windowing. Mel-Frequency analysis of speech is based on human perception experiments. Sample the signal with 11 kHz, apply the sample speech data to the mel-filter and the filtered signal is trained. Number of frames are obtained for each utterance from frequency coefficients by using Praat object software tool.

### 2.8 Neural Networks

Neural networks model some aspects of the human brains, where thinking process is achieved in synaptic connections between neurons. The structure of the network is layered and capable of high parallelism. Neural networks are useful in classification, function approximation and generally in complex problems, which do not require accurate solution. Neural networks must be taught before they can be used, which correspond to how humans learn. A Neural network consists of units that are interconnected with several other such units; they function independently on the input they are given and their local data. Usually all of the units are homogenous, but also heterogeneous networks exists.

Neural networks use a set of processing elements loosely analogous to neurons in the brain. These nodes are interconnected in a network that can then identify patterns in data as it is exposed to the data. In a sense, the network learns from experience just as people do. This distinguishes neural networks from traditional computing programs that simply follow instructions in a fixed sequential order. The structure of a neural network is given below:

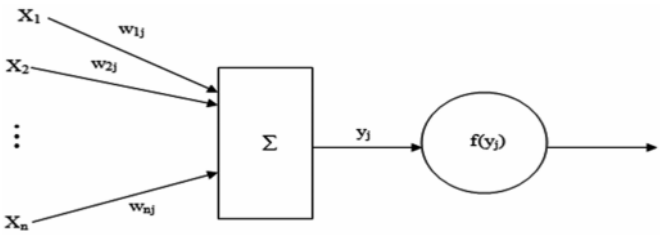


Fig. 1: General Structure of Neural Network

A set of inputs is applied to each node representing the inputs from outside world or, alternatively, they may be outputs from other nodes. Each input is multiplied by a weight associated with the node input to which it is connected and the weighted inputs are then summed together. A threshold value local for each node is added to the weighted summation and the resulting sum is them passed through a hard limiting activation function. The sigmoid function is used as a transfer function.

### 2.9 Learning Methods

Learning is necessary when the information about inputs/outputs is unknown or incomplete. Learning is the method of setting the appropriate weight values. There are two types of training namely supervised and unsupervised. The supervised learning method is used to train the neural network in this paper. Supervised learning requires the network to have an external teacher. The algorithm adjusts weights using input-output data to match the input-output characteristics of a network to the desired characteristics.

In the learning without supervision, the desired response in not known and in supervised learning at each instant of time when the input is applied, the desired response of the system provided by the teacher is assumed. The distance between the actual and desired response serves as an error measure and is used to correct network parameters externally.

### 2.10 Training

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.

### 2.11 Neural Networks for Speech Recognition

Multilayer perceptrons are layered feed forward networks typically trained with static backpropagation. These networks have found their way into countless applications requiring static pattern classification. Their main advantage is that they are easy to use, and that they can approximate any input/output map. MLP can learn arbitrary mappings for classifications. Further, the inputs and outputs can have real values. Typical Back Propagation network architecture is shown below.

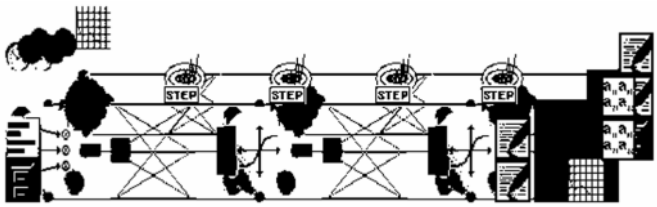


Fig. 2: MLP Network Architecture with Step Learning Rule

The hidden layer learns to recode the inputs. More than one hidden layer can be used. The architecture is more powerful than single-layer networks: it can be shown that any mapping can be learned, given two hidden layers (of units). The step learning rule is used. Step size for gradient this is the multiplying factor for the error correction during backpropagation; it is roughly equivalent to the learning rate for the neural network. A low value produces slow but study learning, a high value produces rapid but erratic learning value for the step size typically ranging from 0.12 to 0.9.

### 2.12 Cross Validation

This is a method for estimating generalization error based on resampling. Cross validation markedly superior for small data sets; the cross validation set is used to determine the level of generalization produced by the training set. Cross validation is executed in concurrence with the training of the network.

## 3. TRAINING PHASE

The multilayer back propagation algorithm is used to train the neural network for spoken words for each speaker. Four speakers are trained using the multilayer perceptron with 176 input nodes, 100 hidden nodes and 11 output nodes each for one word, with the noncurvear activation function sigmoid. The learning rate is taken as 0.1, momentum rate is taken as 0.5.Weights are initialized to random values between +0.1 and –0.1 and accepted error is chosen as 0.009. Frames are obtained for each mel-frequency coefficients of the utterance for each speaker.

### 3.1 Performance Evaluation

The cross validation performance and Active performance are obtained for each speaker and presented in Table 1.

Table 1
The Performance Evaluation for Each Speaker

| *Speaker* | *Active performance* | *CV performance* |
|---|---|---|
| 1 (Male) | 86.66 | 88.86 |
| 2 (Female) | 87.46 | 88.15 |
| 3 (Male) | 86.70 | 82.87 |
| 4 (Male) | 89.01 | 86.76 |

The cost cross validation (cost CV) and the cost time (cost T) are obtained for each speaker [Figs 3–6] by taking epochs on *X*-axis and elapsed time on *Y*-axis.

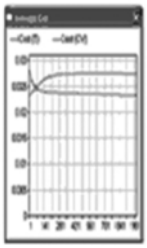
**Fig. 3**

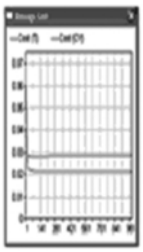
**Fig. 4**

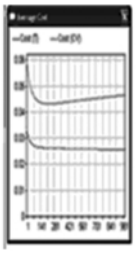
**Fig. 5**

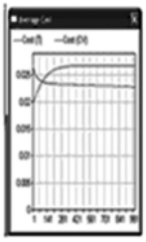
**Fig. 6**

## 4. TESTING PHASE

The same multilayer backpropagation algorithm is used to test the network of the spoken words for the four speakers. Each speaker has to test the network by 11 words repeated four times. Each speaker, tests the word four times and the node with the higher number in the output will be the winner node. The correct answer will be indicated by comparing this node with the input word to the network.

So by testing the words said by each speaker the performance can be found by the equation.

Performance = Total succeeded number of testing words/Total number of words * 100%.

Table 2
The Performance for the Test Phase for Each Speaker

| *Speaker* | *Performance* |
|---|---|
| 1 (Male) | 93.18% |
| 2 (Female) | 90.90% |
| 3 (Male) | 86.36% |
| 4 (Male) | 81.44% |

## 5. CONCLUSION

The neural network architecture has been shown to be suitable for the recognition of isolated words for small vocabularies. The architecture needs iterations to reach acceptable error of 0.01. Recognition of the words is carried out in speaker dependent mode. In this mode the tested data presented to the network are different from the trained data. The Mel-frequency cepstral coefficients with 12 parameters from each frame improves a good feature extraction method for the spoken words since the first 12 in the cepstrum represent most of the formant information. In all speech recognition systems, the data is recorded using a noise-canceling microphone. Since this type of microphone is not available, the data was recorded using a normal microphone, but recorded in a closed room without any type of noise. It is found that the cross validation performance exceeds the active performance for speakers 1 and 2. It is also found that the active performance exceeds cross validation performance for speakers 3 and 4. The performance for the test phase for speakers 1-4 is computed with an error of 14.72%, 9.10%, 3.64% and 12.56% respectively. The promising results are obtained in the test phase due to the exploitation of discriminative information with neural networks.

## REFERENCES


[1] Srinivas Desai, E. Veera Raghavendra, B. Yagnanarayana, Alan W. Black, Kishore Prahallad, "Voice Conversation using Artificial Neural Networks", *International Conference on Accoustics*, 2009.

[2] Ben Gold and Nelson Morgan, "Speech and Audio Signal Processing", Wiley India Edition, New Delhi, 2007.

[3] B. Yegnanarayana, "Artificial Neural Networks", Prentice-Hall of India, New Delhi, 2006.

[4] R.O. Duda, P.E. Hart, and D.G. Strok, "Pattern Classification", 2nd edn, John Wiley, New York, 2001.

[5] D.P. Delacretaz and J. Hennebert, "Text-Promoted Speaker Verification Experiments with Phoneme Specific MLPs, Proc. ICASSP, pp. 777–780, 1998.

[6] E. Gose, R. Johnsonbaugh, and S. Jost, "Pattern Recognition and Image Analysis", Prentice Hall, Englewood Cliffs, New Jersey, 1997.

[7] Kevin Gurney, "An Introduction to Neural Networks", Ucl, Press Limited Taylor & Francis Group London, 1997.

[8] K.R. Farrell, R.J. Mammone, and K.T. Assaleh, "Speaker Recognition using Neural Networks and Conventional Classifiers", *IEEE Trans. Speech and Audio Processing*, **2 (1)**, Part II, 1994.

[9] J.M. Naik and D.M. Lubensky, "A Hybrid HMM-MLP Speaker Verification Algorithm for Telephone Speech", *Proc. ICASSP*, pp. 153–156, 1994.

[10] H. Hattori, "Text-Independent Speaker Verification using Neural Networks", *Proc. ESCA Workshop Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland, 1994.

[11] A.C. Tsoi, Shrimpton, B. Watson and A. Black, "Application of Artificial Neural Network Techniques to Speaker Verification", *Proc. ESCA Workshop Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland, 1994.

[12] M. Nadler and E.P. Smith, "Pattern Recognition Engineering", John Wiley, New York, 1993.

[13] L. Rabiner, Biing-Hwang Juang, B. Yeganarayana, "Fundamentals of Speech Recognition", Pearson Education, 1993.

[14] R.J. Schalkoff, "Pattern Recognition: Statistical, Structural, and Neural Approaches", John Wiley, New York, 1992.

[15] J. Oglesby and J.S. Mason, "Radial Basis Function Networks for Speaker Recognition", *Proc. ICASSP*, pp. 393–396, 1991.

[16] C.H. Chen, "Signal Processing Handbook", 1985.

[17] Frank Fallside/William A. Woods, "Computer Speech Processing", Printic-hall, 1985.

[18] Lawrence R. Rabiner and Stephene E. Levinson, "Isolated and Connected Word Recognition Theory and Selected Applications", *IEEE Transaction on Communications*, **Com-29 (5)**, May 1981.