WORD BASED RECOGNITION OF ASSAMESE LANGUAGE USING ARTIFICIAL NEURAL NETWORK

Dr. Mousmita Devi,
Department of Computer Science, Handique Girls' College

ABSTRACT
This paper discusses the speech recognition process using the classifier Artificial Neural Network. In this work, MLP structure with **error back propagation algorithm** is used in which the errors are propagated backwards from the output nodes to the input nodes. Usually, one hidden layer is enough for efficient speech recognition or classification. In my research study, only one hidden layer is considered. The number of nodes in the hidden layer is adjusted empirically for the better performance of the system. The main problem here is to classify the speech sample feature vectors into several speech classes. To reduce the volume of data which is to be feed in the input layer clustering technique is used in this study. The clustering of data decides how the related data can be categorized into different classes. The **K-means** clustering, one of the clustering techniques is used in the present study. Speech recognition process by both ways i.e., by Speaker dependent as well as Speaker independent ways techniques are used in this work. Comparative performance evaluation is done for both mode of Speech Recognition system.

KEYWORDS: Artificial Neural Network, Speech Recognition, MLP, Speaker dependent, Speaker independent

INTRODUCTION
ANN is a mathematical computational model which is designed to mimic the human brain and is presently used as a very popular and efficient tool in pattern recognition and prediction problems [2]. A Neural Network (NN) consists of a number of interconnected processors which are known as neurons. There are weights associated with the neurons and these are multiplied with the signal value passing through it [3]. Due to its characteristics which are given below ANN is considered to have a remarkable ability in recognizing patterns:

a) ADAPTIVE LEARNING-Neural networks store their information (data) in the strengths of their interconnections. But in case of a computer, information (data) is stored in the memory which is addressed by its location.

b) PARALLEL ORGANIZATION- The real power of brain comes from its capability of parallel processing. Parallel processing is one where simultaneously more than one instruction or job can be executed. In a conventional computer, generally programs have large number of instructions or lines of coding, and they operate in a sequential mode or we can say one instruction after another.

c) FAULT TOLERANCE-Neural networks exhibit the property fault tolerance too. On the other hand, computer system does not exhibit the fault tolerance. The information once corrupted in the memory, it cannot be retrieved again.

d) ROBUSTNESS-Neural network is consisting of a large number of computing elements. And this computing is not restricted to within neurons only. The number of neurons in a brain

is estimated to be about $\mathbf{10^{10}}$and the total number of interconnections to be around $\mathbf{10^{14}}$ [4]**.**Thus, robustness of the size and complexity of connections make the brain to perform the complex pattern recognition tasks which are not possible to realize in today's computer.
While training the feature vectors using ANN as classifier, the first step is to initialize a set of parameters that affect and also influence the network learning process. This step includes the network topology adopted as well as the learning parameters chosen such as learning rate and momentum. Network topology provides the structure of the network which deals with the suitable number of hidden layers and the number of neurons chosen in the hidden layers which in turn improves the mapping between the input and output nodes [5]. The number of neurons present in the hidden layer has direct impact on the performance of the ANN as well as we can say which in turn affects the overall recognition rate. Because, if the number of neurons is more, then it may have the over fitting problems and if the number of neurons is too low, it may have the under fitting problems. In this work, the **Multi-Layer Perceptron (MLP)** structure of the ANN which is a feed forward network consisting of multiple layers with one **input layer** which accepts the N inputs through N parallel input connections, one or more **hidden layers** which accept the weighted sum of the output from the input units and an **output layer** which accepts the weighted sum of the output from the hidden units which finally forms the output is used.

The **backpropagation** algorithm is the most widely used and popular supervised training algorithm for neural network. During training of a network using backpropagation involves **three stages**: the **feedforward** of the input training pattern, the **calculation** and **backpropagation** of the **associated error**, and the **adjustment of the weights** [5].

STRUCTURE OF NEURAL NETWORK (TOPOLOGY)
An Artificial Neural Network (ANN) consists of a pool of simple processing units i.e. neurons, which communicate by sending signals to each other over a large number of weighted connections. Each unit performs some relatively basic task. First it receives input from neighbors or some external sources and uses that information to compute an output signal that is again propagated to other units in the ANN. Weights has been adjusted next to the actual information processing task. An ANN is inherently parallel in the sense that many units can perform computation cycles simultaneously [65]. ANN's architecture has basically consists of three types of units. (1) The **InputUnits** that receive data from external source feed into the net. (2) The **OutputUnits** that act as the ANN output or endpoints. (3) The **HiddenUnits** where the input as well as output signals remain within the ANN framework.
The different units of Artificial Neural Network (ANN) are depicted as bellow:

INPUT UNIT:
The feature vector from **Linear Predictive Coding (LPC)** and **Mel Feature Cepstral Coefficient** (**MFCC)** are compressed (clustered) individually by **K-means** clustering algorithm which is described in the previous Chapter are given as the input to the neural network. This results a fixed number of inputs (feature vectors) to the neural net. In this research work in order to get a better performance of the recognizer a different number of clusters have been tested. For better performance evaluation the in this present work I have considered the values of K are **10** and **16** respectively.

OUTPUT UNITS
In the present research work to recognize five different words from our database, the number of nodes in the output layer is considered as five (5). If the output value is equal to one (1) or

is very close to one (1), then it is considered that the word is detected. The output result from the logsig() function will again return into the matrix which is squashed into [0,1]. In this study, I have considered output layer limited to only five, such that the neural net computation can give a better performance with a greater accuracy. During speech recognition phase, for each utterance a combined feature vector is generated. Then in training phase a word class is computed from the feature values of each utterances of our speech database. If the computed feature vector of a particular utterance has the maximum **probability** of having that word class, then that utterance is said to belong to that word class only.

HIDDEN LAYERS

It isvery difficult to choose the number of hidden layers and number of neurons present in an ANN. Usually, for most applications, one hidden layer is enough. Technically, due to attentuation problems, models such as the back propagation-trained multilayer perceptron have issues with too many layers [6]. It is proven that MLPs with only one hidden layer are universal function approximators [7]. Therefore, one hidden layer is enough for efficient speech recognition or classification. Hence, in my research study, only one hidden layer is considered.

In this present study I have begin with an MLP having a hidden layer comprises of a small number of nodes, then increase the number of nodes in the hidden layer, until the generalization error begins to increase due to over fitting problem. Therefore in this present study the numbers of nodes considered in the hidden layer are **40, 60, 80** and **100.**

NEURAL NETWORK DATA SET ANALYSIS

During the operation or working phase of ANN, the input data are divided into three subsets. They are:

TRAINING PHASE

Once a neural network is structured for a particular system, that network is ready to be trained. The initial weights are chosen randomly to start this process. In this way, the training, or learning, begins. The data is generally part of the whole database, during the training phase. After using all the training data once, then it is called a **learn cycle or one epoch** [8]. The training data will be given in to the neural network continuously until the weight values are determined.

VALIDATION PHASE

In order to avoid **over-fitting or over-training** problems during the validation phase the main task is to check the performance of the network and to determine the epoch at which training should be stopped. While training the neural network, normally the best validation performance and validation checks are calculated. During the phase of the learning or working of the neural network, the best validation performance normally **decreases** while the epoch **increases**. When the neural network is still leaning or working correctly, the validation checks is normally 0 (zero). But will **increase** if the neural network is **over-fitting** or **over-training**. The neural network stops, while the threshold is reached. If the validation data set indicates the network is over-trained, then using a different number of parameters values the network should be retrained.

TESTING PHASE

After the neural network is trained well, then a testing data set is passed to the neural network to evaluate the neural network performance [5]. Normally the neural network uses all the data

including old data, which has been used before and new data, which has not been used before, as a testing data set in the testing procedure. With a different input data set, it will have different errors or error rate.

In this research work, the architecture of the MLP network has been considered, with an input layer, one or more hidden layers, and an output layer. The algorithm which is used is the back propagation training algorithm. In this type of network, the input is presented to the network and moves through the weights and nonlinear activation functions towards the output layer, and the error is corrected in a backward direction using the well-known error back propagation correction algorithm [3]. After having extensive training to the database, we have considered for training phase, the network eventually establishes an input-output relationship through the updated weights on the network. The training parameters set in my study are listed in the **Table 1**-

Table 1: Training parameters set

PARAMETERS	**VALUES**
Learning rate	0.05
Epochs	2000
Goal	0.005

PERFORMANCE EVALUATION

There are different methods which are generally used for evaluating the performance of a speech recognition system. The major metrics which are mostly used for the better performance evaluation are **Recognition Accuracy** and **Confusion Matrix**.

RECOGNITION ACCURACY: The major criterion for evaluating the performance of a Automatic Speech Recognition (ASR) system is the **RecognitionAccuracy**, which is a practical value and an important measure for all speech recognition applications.**RecognitionAccuracy** is the number of utterances recognized correctly out of the total number of utterances spoken. It is expressed in a percentage value. The **recognition rate** or **accuracy** of each of the utterance has been computed with the help of the following **Equation 1**

$$\mathrm{Re}cognitionRate = \frac{Number\ of\ \ samples\ correctly\ recognised}{Total\ \ No\ of\ \ Tested\ \ Speech\ Samples} * 100\% \quad \textbf{…. (1)}$$

CONFUSION MATRIX: A confusion matrix as the name says is a performance analysis tool which is used to represent the results in a matrix format. The diagonal elements of a confusion matrix contain the instances that are correctly classified [8]. Recognition accuracy provides the percentage of correctly and wrongly classified utterances only. But on the other hand confusion matrix provides more information about where the classifier (recognizer) failed in recognizing the features and also gives the detailed class conditional error rates. If there are m classes, a confusion matrix table will be of at least size m by m. For a recognizer or a classifier to have a good accuracy the rows along the diagonal of the confusion matrix should contain maximum values and the rest of the entries should have zero or a very small values. A confusion matrix can be represented using the **Equation 2**

$$Eji = \Pr\{decision\ j \mid class\ i\} \quad \textbf{…. (2)}$$

Which denotes the matrix of counts where the true class i is classified as j [9].

PERFORMANCE EVALUATION OF SPEECH RECOGNITION SYSTEM

Automatic Speech Recognition (ASR) can also be defined as the speech to text conversion process. Because, Speech recognition or which is more commonly known as automatic speech recognition is the process of converting an acoustic waveform into the text form which is known to the user [8]. Commonly Speech Recognition can be divided into two types. One is called speaker–dependent and the other is speaker–independent. Speaker–dependent software is commonly used for dictation software, while speaker–independent software is more commonly found in telephone applications [10].

SPEAKER DEPENDENT SPEECH RECOGNITION

In Speaker dependent Speech Recognition system, only one speaker is selected from the rest of **20 (twenty)** speakers. During network training one word is uttered by **10** times is used to train the network. In other words,**50** speech samples i.e. **10*5=50 (10 repetitions*5 words)** of the speech sample uttered by a single speaker is used to train the neural net. The five output nodes represent five different speech or word. For testing mode, all the **1000(20 speaker*5 words*10 repetitions)** tokens were used in recognition phase (testing mode). This implies that the training data set is a subset of the testing data set. The selected word list is depicted in the **Table 2** below.

Table 2: Word List Training Performance

WORD LIST					
Words in Assamese	Words in English	IPA Format	English meaning	Syllable Structure	Number of Syllables
সমাজ	Xomaj	/xɔmaz/	Society	CV/CVC	2
নাক	Nak	/nak/	nose	CVC	1
জনাজাত	Jonajat	/zɔnazat/	Popular	CV/CV/CVC	3
গছ	Gos	/gɔs/	tree	CVC	1
এক	Ek	/ek/	one	VC	1

The training parameters are-

PARAMETERS	**VALUES**
Learning rate	0.05
Epochs	2000
Goal	0.005

The following **Figure 1** gives the training performance results of isolated word speech recognition system. From the plot it can be depicted that a satisfactory result is present in this study which classify the speech or words of Assamese language. From the performance plot i.e. from **Figure 1**, it can be seen that the goal was meet at **957** epochs. In this Figure the x-axis represents the number of epochs, where the y-axis represents the performance.

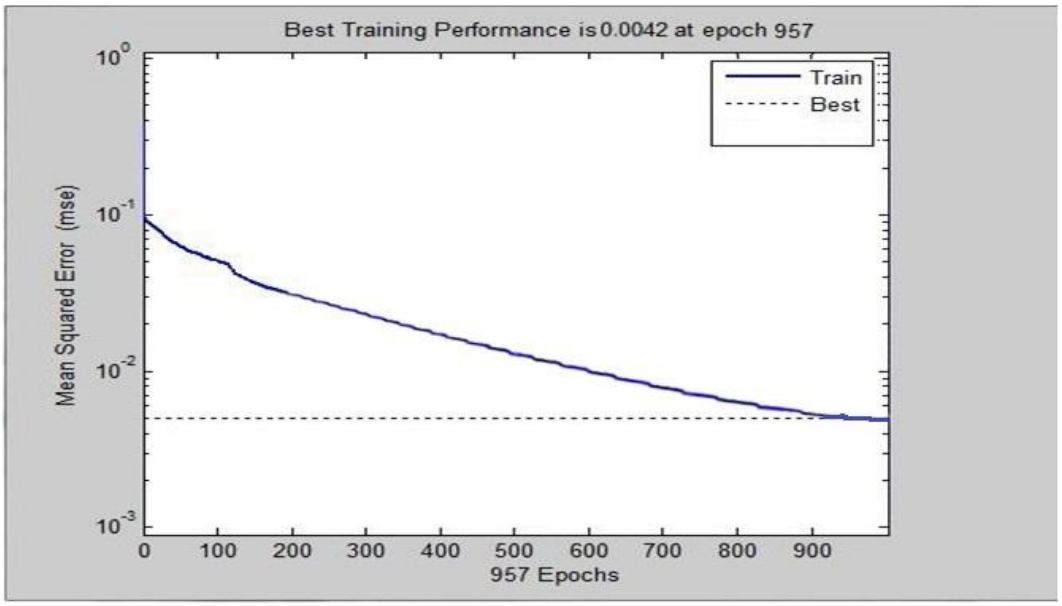

Figure 1:- Performance evaluation of Speaker dependent Speech Recognition System

The recognition rate is shown in the **Table 3** for each cluster size and number of hidden neurons present in the resultant matrix. If we study the behavior of the neural network, it is observed that the recognition rate is increased with the number of nodes in the hidden layer for both the clusters. It has produced about **99%** correct speech recognition at cluster 10 which is satisfactory in a recognition system to classify speeches.

Table 3: Recognition rate for Speaker dependent Speech recognition system

RECOGNITION RATE		
No. of nodes ⬇	CLUSTER 10	CLUSTER 16
40	91%	89%
60	92%	91%
80	97%	96%
100	**99%**	97%

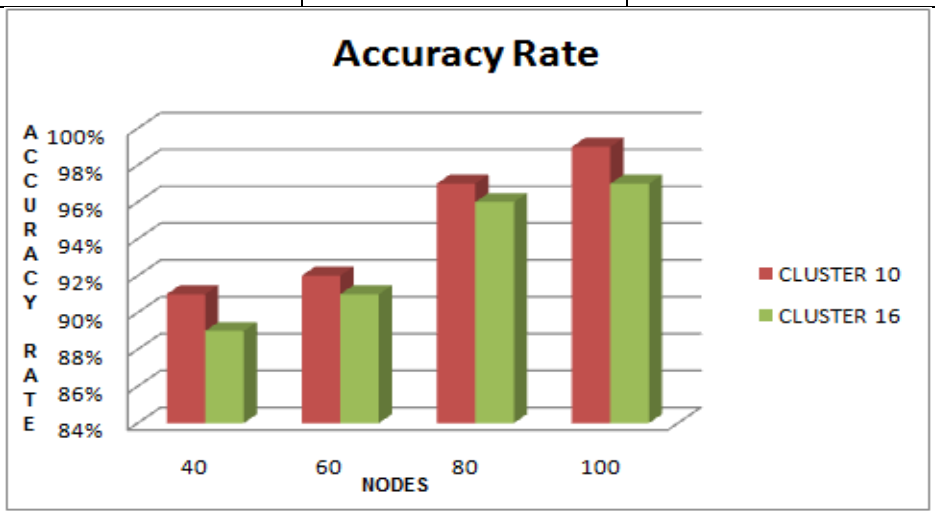

Figure 2:- Accuracy Rate of Speaker dependent Speech Recognition System

So, from the **Table 3** and **Figure 2** we found that the recognition rate is higher at cluster 10 with number of nodes/neurons in the hidden layer is 100. From this we can say that the

architecture for the best recognition rate is found at **400-100-5** i.e. 400 input nodes for 400-element feature vector, 100 nodes in the hidden layer and 5 output nodes for discriminating between 5 words. Further analysis is done based on this architecture only.

SPEAKER INDEPENDENT SPEECH RECOGNITION

In speaker-independent Speech Recognition mode, on the other hand, ten speakers were used for the training phase purpose. The total speech samples which were dedicated for this phase was **500 (10 speakers × 10 repetitions × 5 words).** These 10 speakers with their voices are known to the system. But the voice of rest of the remaining 10 speakers of our database is unknown to the system. In testing phase the speech sample of 10 speaker's speech samples which are unknown to the system are fed into the neural net for recognition. This data setting was applied for the ANN based systems.

TRAINING PERFORMANCE

The training parameters are-

PARAMETERS	VALUES
Learning rate	0.05
Epochs	2000
Goal	0.005

The following **Figure 4** gives the training performance results of isolated word speech recognition system. From the plot it can be depicted that a satisfactory result is present in this study which classify the speech or words of Assamese language. From the performance plot i.e. from **Figure 4**, it can be seen that the goal was meet at **1088** epochs. In this Figure the x-axis represents the number of epochs, where the y-axis represents the performance.

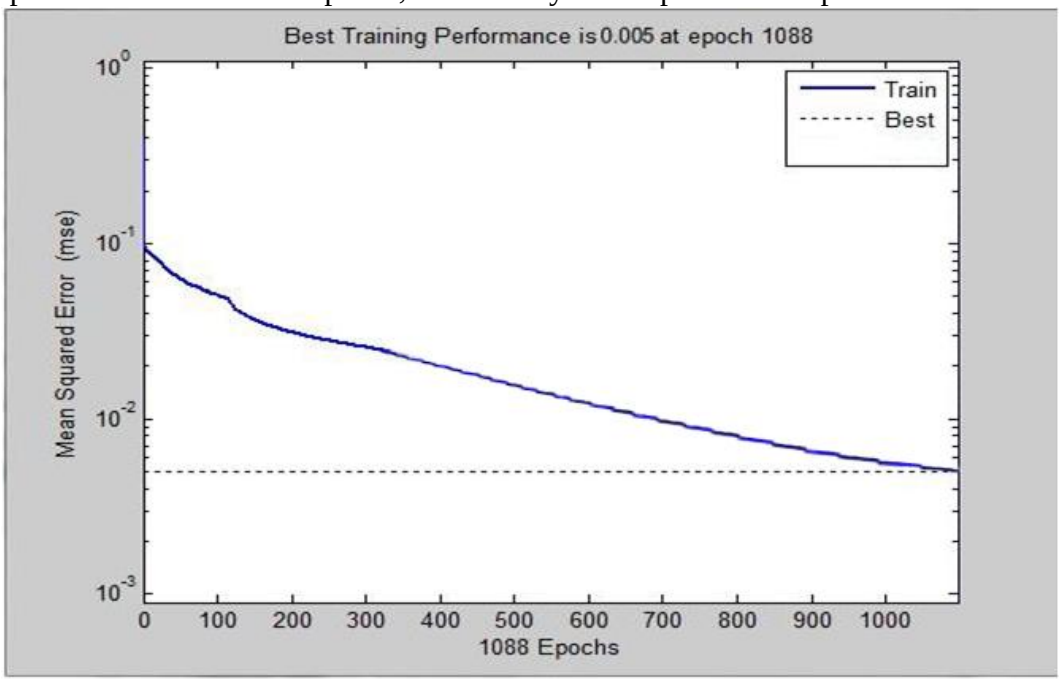

Figure4 :- Performance evaluation of Speaker independent Speech Recognition System

The recognition rate is shown in the **Table 4** for each scenario cluster size and number of hidden neurons present in the resultant matrix.

Table 4: Recognition rate for Speaker independent Speech recognition system

RECOGNITION RATE		
No. of nodes ⬇	CLUSTER 10	CLUSTER 16
40	85%	84%
60	89%	87%
80	94%	89%
100	**96%**	91%

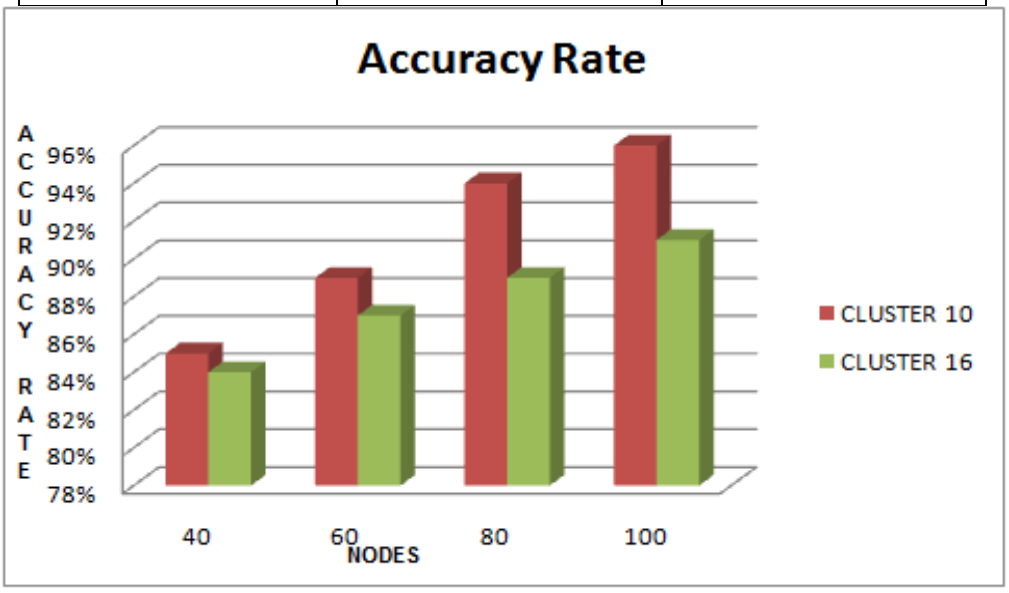

Figure 5:- Accuracy Rate of Speaker independent Speech Recognition System

Hence, from the **Table 4** and **Figure 5**, we can say that the best recognition rate is found when cluster size is 10 and number of nodes/neurons present in the hidden layer is 100. So, the best architecture to find out the best recognition rate is **400-100-5**, where 400 is the number of input nodes, 100 is the number of neurons present in the hidden layer and 5 is the output node to recognize 5 different words. Further analysis part has been done adopting this architecture only.

COMPARISON OF SPEAKER DEPENDENT AND SPEAKER INDEPENDENT SPEECH RECOGNITION

Further, we used a confusion matrix which is a simple matching matrix commonly used to predict the classification results and to visualize the performance of the method. The confusion matrix is designed by labeling the desired classification on the rows whereas on the column side they have the predicted classifications. Since we want the predicted classification to be the same as the desired classification, the ideal situation is to have the values of the matrix diagonal equal to the total number of test data for each word [11]. For speaker dependent mode values on matrix diagonal must be equal to **200** for acquiring 100% recognition rate. Which is the total number of test data of each word (**20 speaker * 10 repetitions)**. Whereas, in case of speaker independent mode values on matrix diagonal must be equal to **100 (10 speaker* 10 repetitions)** for acquiring 100% recognition rate. The speaker dependent mode and speaker-independent mode which were used in configuring the system and their performances in confusion matrix is shown in **Figure 6** and **Figure 7**respectively. It has seen in the confusion matrix that the overall system accuracy is 92% for Speaker dependent Speech recognition system and 88.2% for Speaker independent System.

The worst performance was found in the case of word **/zɔnazat/** with accuracy equal to 75% in case of Speaker Independent case; and the best performance was encountered in the case of word **/nak/** and **/ek/**with accuracy equal to 100% in case of Speaker dependent system. Hence from the confusion matrix we can depicted that the recognition rate was best in case of **monosyllabic** words, whereas recognition rate gradually decreases as the number of syllables increases. Also the recognition rate of speaker dependent mode has found maximum as compared to that of speaker independent recognition mode.

	/xɔmaz/	**/nak/**	**/zɔnazat/**	**/gɔs/**	**/ek/**	**ACC(%)**
/xɔmaz/	**165**	**0**	**5**	**30**	**0**	**82.5%**
/nak/	**0**	**200**	**0**	**0**	**0**	**100%**
/zɔnazat/	**34**	**0**	**156**	**10**	**0**	**78%**
/gɔs/	**0**	**0**	**0**	**199**	**1**	**99.5%**
/ek/	**0**	**0**	**0**	**0**	**200**	**100%**
Average						**92%**

Figure 6: -Confusion matrix for Speaker dependent mode for Isolated Words database using LPC+MFCC+MLP combination

	/xɔmaz/	**/nak/**	**/zɔnazat/**	**/gɔs/**	**/ek/**	**ACC(%)**
/xɔmaz/	**78**	**0**	**8**	**12**	**2**	**78%**
/nak/	**0**	**98**	**0**	**2**	**0**	**98%**
/zɔnazat/	**15**	**0**	**75**	**6**	**4**	**75%**
/gɔs/	**1**	**6**	**0**	**93**	**0**	**93%**
/ek/	**0**	**3**	**0**	**0**	**97**	**97%**
Average						**88.2%**

Figure 7: -Confusion matrix for Speaker independent mode for Isolated Words database using LPC+MFCC+MLP combination

Table 5 gives the words that were picked in case of miss-recognition for both mode (Speaker dependent and Speaker Independent)

Table 5: ANN words that were picked in case of miss-recognition for both modes.

Words	**Confuse with words**	
	Speaker dependent mode	**Speaker independent mode**
/xɔmaz/	/zɔnazat/,/gɔs/	/ zɔnazat/,/gɔs/,/ek/
/nak/	-	/gɔs/
/zɔnazat/	/xɔmaz/,/gɔs/	/xɔmaz/,/gɔs/,/ek/
/gɔs/	/ek/	/xɔmaz/,/nak/
/ek/	-	/nak/,

CONCLUSION

This chapter describes the speech recognition systems using two spectral analysis techniques, LPC and MFCC respectively. Classification algorithms Artificial Neural Network (ANN) were considered to recognize the Assamese words. The schemes developed were tested on the database in Assamese language. From the results obtained, it is observed that the

recognition rates which are obtained using speaker dependent mode is slightly better than that of speaker independent in recognizing the speech samples for Assamese language. The speech recognition experiments using MLP are repeated by changing a number of parameters by trial and error experiment. The learning parameters which were used for both Speaker dependent and Speaker independent speech recognition are given in **Table 6.**

Table 6:- Learning parameters used by speech recognition system

Parameters	Values
Number of hidden layer	1
Number of neurons in the hidden layer	[40 60 80 100]
Learning rate	0.05
Epoch	2000

The inferences were obtained from the experiments performed which are as follows:

a) Between Speaker Dependent and Speaker Independent Speech recognition system. Speaker Dependent produced better results on the dataset I considered. This result can be depicted from the **Figure 8**

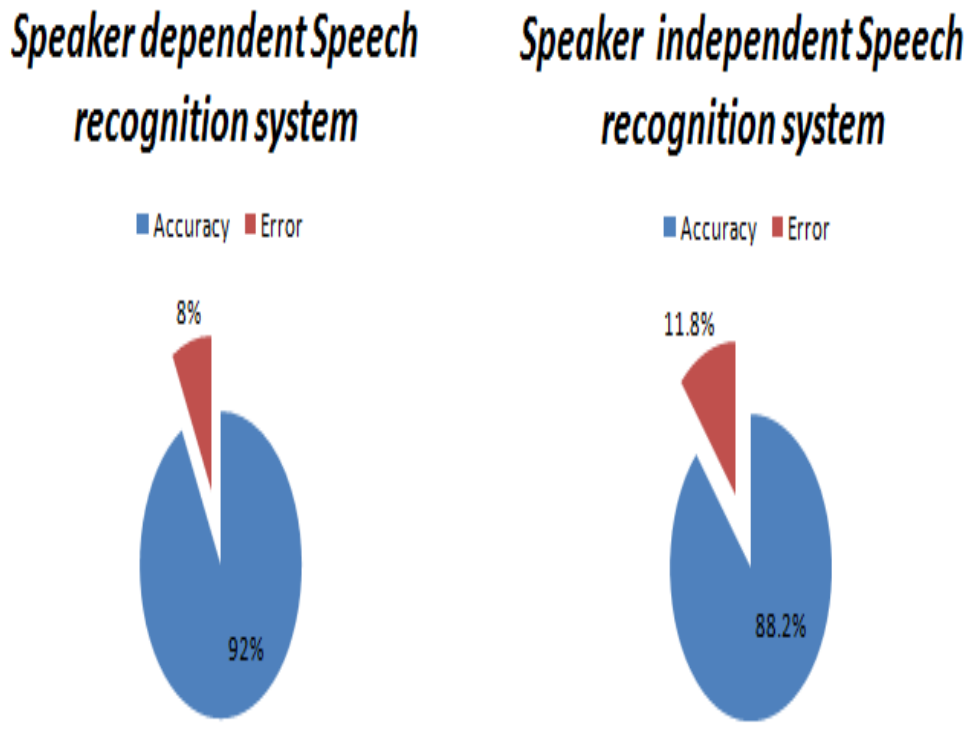

Figure 8: - Results obtained using Speaker dependent and Speaker independent system.

b) As compared to more syllabled structured words having recognition rate smaller than that of monosyllabic or disyllabic. In this study **/zɔnazat/** of trisyllabic having the lowest recognition rate **75%** as compared to other monosyllabic words like **/nak/ ,/ek/** and **/gɔs/** having recognition rate **100%** or **near to 100%.**

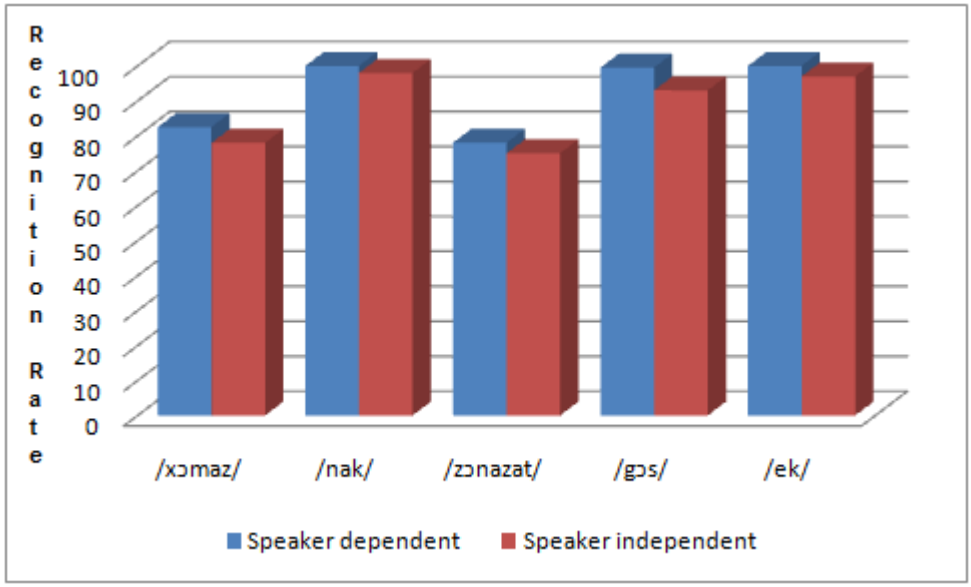

Figure 9: - Performance evaluation of Speaker dependent and Speaker independent system.

REFERENCES

[1]. Mousmita Devi "Spectral Analysis of Assamese Words and Their Recognition Using ANN", PhD Thesis submitted to Gauhati University, 2016.

[2] Christopher M. Bishop, Neural Networks for Pattern Recognition, 1st ed. Oxford University Press, 1996.

[3] S. N. Sivanadam, S. Sumathi, and S. N. Deepa, Introduction to Neural Networks using Matlab 6.0, New Delhi, India: Tata McGraw-Hill, 2006.

[4] Brian D. Ripley, Pattern Recognition and Neural Networks, 1st ed.Cambridge, New York: Cambridge University Press, 2008.

[5] Md Salam, Dzulkifli Mohamad, and Sheikh Salleh, "Malay Isolated Speech Recognition Using Neural Network: A Work in Finding Number of Hidden Nodes and Learning Parameters," The International Arab Journal of Information Technology, vol. 8, no. 4, pp. 364-371, Oct. 2011.

[6] http://stackoverflow.com/questions/10002782/artificial intelligence - How to choose number of hidden layers and nodes in neural network - Stack Overflow.mht

[7] http://stackoverflow.com/questions/10002782/machine learning - multi-layer perceptron (MLP) architecture criteria for choosing number of hidden layers and size of the hidden layer - Stack Overflow.mht

[8] Laurene Fausett, Fundamentals of Neural Networks Architectures, Algorithms and Applications, 1st ed. Prentice Hall, 1994.

[9] Ajith Abraham, "Artificial Neural Networks," in Handbook of Measuring System Design, vol. 1, Wiley, 2005, ch. 129, pp. 901-908.

[10] CiniKurian, and KannanBalakrishnan, "Malayalam Isolated Digit Recognition Using HMM and PLP Cepstral Coefficient," International Journal of Advanced Information Technology (IJAIT), vol. 1, no. 5, pp.31-38, Oct. 2011.

[11] Amer M. Elkourd "Arabic Isolated word speaker dependent Recognition System", MS Thesis submitted to The Islamic University, 2014.