Automatic Spoken Digit Recognition Using Artificial Neural Network

P. Sarma, S. Sarmah, M.P. Bhuyan, K. Hore, P.P. Das

Abstract: Speech Processing is a vast domain for research work where Speech Recognition is a small part of it. This research work is an attempt to recognize ten spoken English digits starting from zero to nine by using Artificial Neural Network. The system will be able to recognize digits spoken in English. Although the spoken language is English, the utterance may vary from speaker to speaker. We have built the system by training the voice samples of the people of the Northeastern region of India. After recognizing the utterance it will display the recognized digit as output. We have programmed and simulated the design in MATLAB. The design of our research is based on using the Linear Prediction Coefficient (LPC) and Principal Component Analysis (PCA) for signal analysis. We were able to get an overall accuracy of 82%.

Index Terms: Automatic Speech recognition, artificial neural network, deep learning, LPC, PCA.

1. INTRODUCTION

SPEECH is the easiest and powerful way of communication among people. Mostly human body expresses their thoughts through speech in any language. Speech recognition by a machine means the machine can understand the speech and act accordingly. In recent years, research in the field of speech technology is getting more important for dialect and emotion recognition, stress and accent identification, etc. [1]. Artificial intelligence has already occupied almost every area of modern digital technology which also includes different speech processing technology. Nowadays, we are having a voice control mechanism for most of the electronics and computer devices. They work through speech recognition methodology. The speech recognition system is incorporated for the aid of blind people and in medical fields. For speech recognition, the system has to be highly sophisticated to recognize the appropriate person. For Automatic speech recognition (ASR) the annotated speech corpus has to be built along with the language model and grammar of that language. In the case of continuous speech recognition of word boundary ambiguity should be addressed properly [2]. Input to any speech recognition system is a voice but the output may be in some other format, perhaps an activity. For the computer to act correctly it is necessary to understand the command or information given to it. The use of machine learning, mainly deep learning improves the recognition rate of speech recognition to a remarkable degree. The ASR system built using deep learning can be trained in the end to end manner. This model is simplified and maps speech to text directly.

A speech recognizer can assist in the following ways.

1. Illiterate or very little literates who are unable to type on the keyboard or not having the keyboard facility.
2. Dyslexic people having problems in word typing or text manipulation. Patients and old people unable to type text.
3. People with health problems such as unable to read (blind people), disability in typing, etc.

- *P. Sarma is currently working as an Assistant Professor in Information Technology department in Gauhati University, India. Email: parismita.sarma@gmail.com*
- *S. Sarmah is currently working as an Assistant Professor in Information Technology department in Gauhati University, India. Email: satyajitnov2@gmail.com*
- *M. P. Bhuyan is currently pursuing PhD degree in Information Technology in Gauhati University, India, E-mail: mpratim250@gmail.com*
- *K. Hore was a BTech student of GUIST, Gauhati University, India.*
- *P. P. Das was a BTech student of GUIST, Gauhati University, India.*

1.1 Speech Recognition

There are different types of speech recognition systems. Different speech recognition categories are constructed depending on the recognizing capability of different speech units, words, and collection of words they pose. Some of these speech recognizers are as follows:

a. Isolated Speech: This is the simplest of all speech recognition systems. In this case, isolated words are recognized individually. For this a pose between each word of the utterance is necessary. Endpoints or word boundary becomes easy to recognize when using this recognition system.

b. Connected Speech: The connected speech recognition system is analogous to the isolated speech recognition system. This method can be used in a different continuous speech application. Very small pause is allowed in between individual unit, decoding of the whole sentence is done by concatenation of the individual model for each word.

c. Continuous Speech: This synthesizer uses the natural stream of speech units to build the recognizer. It can even recognize speech in the absence of pauses or other delimiters. Continuous speech recognizers can easily recognize a huge number of utterances compared to other methods.

d. Spontaneous speech: This recognizer is a result of the effective development of speech and information processing. This is a challenge for the Automatic Speech Recognition process as it has to deal with bogus starting, repetitions, extra pauses, etc. As a whole spontaneous speech processing system has to work with an impulsive real-time speech that was not rehearsed earlier.

1.2 Use of Artificial Neural Network in Speech Processing

In general, a human can easily understand the voice of a person or recognize a speech when he or she is already heard it. But how a machine can recognize a voice is a complicated process. To accomplish this work the machine has to go through several stages. The artificial neural network helps much to recognize human speech by using different optimization algorithms. In general digitization of speech, the signal is the most important first step. Speech is an analog signal which should be converted to digital form. Sampling and quantization are done to convert the continuous signals into discrete form. In the next phase signal processing is done to separate speech signals from background noise. Several phonetics and phonological processing are done on the

filtered speech signal to deal with the variability of the signal. Next semantic and pragmatic analysis is done to understand the meaning of the utterance.

1.3 Linear Prediction filter Coefficients (LPC)

Linear prediction coefficients (LPC) are widely used in many areas of speech processing. With the LPC interpolation technique, higher data compression rates of the voice codec could be achieved. Thus it helps to recover the corrupted segments of the input speech at the receiving end. Most of the modern voice/speech communication systems use LPC coding to encode and decode speech signals to reduce data throughput and thus save bandwidth. Parameters needed to represent a signal can be reduced using the LPC interpolation technique. It is a very good feature extraction method for speech recognition.

1.4 Principal Component Analysis (PCA)

The principal component analysis is a feasible method used basically for variable reduction. We have used this technique in this work and mention it in the proper place. The method uses orthogonal transformation to transform a set of correlated values to linearly uncorrelated values, which are termed as principal components. In PCA, the initial principal component has leading potential variance and succeeding components generally have the highest possible variance under the limitation that is orthogonal to previous components. When there is a probability of redundant data in a system PCA can be used to remove these. The principal components are able to account for most of the variance of other variables. The remaining sections of the paper are organized as follows: Section II has its focused on the related works, Section III describes the proposed work, and Section IV shows the experimental results and discussions on the results at last Section V concludes the present work with future scopes.

2 RELATED WORKS

We have studied some journals and conference papers to get an insight into our work, a few of them are mentioned below which we found useful. In paper [3], by Heidi Horstmann Koester we studied a method to recognize discrete speech using Artificial Neural Network (ANN) training and testing. We have got an idea of the working of the discrete speech recognition system from this paper. Another paper by Sherry Doggett et al. [4] was on the use of speech recognition in the Electronic Health Record system. They have used front end speech recognition as well as back end speech recognition technologies that are good for preparing legible and comprehensive products. Paper [5] by Kamble et al. has an elaborative discussion on ASR using different ANN techniques. They showed how efficient is a neural network for speech signal classification. They concluded in spite of complex algorithms like Recurrent Neural Network (RNN) that artificial neural networks can classify speech signals more accurately than Multi-Layer Perception (MLP). Another paper [6] by Katz, M. et al. discuss different speech recognition methods for several South African dialects. They mainly carried out their research work for the recognition of different accents of different languages of that area. The role of accent in speech recognition is analyzed in their research paper. We have studied another recent work [7] by Nassif et al. and got ideas of using deep neural networks for automatic speech recognition. The paper has done statistical analysis on almost 174 different research papers published between 2006 and 2018. The authors of this paper reveal that most of the papers used the word error rate (WER) to measure their respective system's efficiency. They have shown that still today most of the researchers use MFCC, LPC to extract speech features while using deep learning models in the field of ASR uses.

3 PROPOSED METHOD

In this research work, we are showing a speech recognition model using a neural network. It recognizes ten spoken English digits (from zero to ten) in a speech. The processing blocks to recognize those words utterance is as follows: Data Preparation, Feature Extraction, Training/Analysis, and Identification/Testing. Fig. 1 shows the block diagram for the system model.

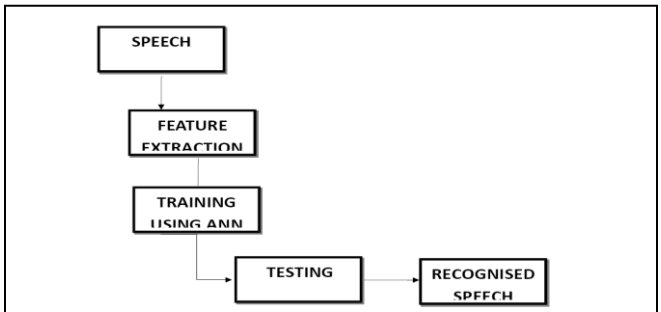

***Fig. 1**. System Model Diagram*

The whole process is broken down into several sub-processes to decrease the complexity of the long original process. The following sub-processes of the original one are performed:

1) Specific dataset creation (0 - 9)
2) Data recording.
3) Feature Extraction.
4) Training of the model
5) Real-time data evaluation and testing.
6) Result analysis.

A brief description of the whole system building is given below.

3.1 Dataset Preparation

A data set is built containing the numerical words (0-9) of English language. We had selected the numerical words as an initial experiment which could be extended for any word recognition in future.

3.2 Recording and Preprocessing

Recording and pre-processing of data is the second phase of our work. For the recording of the numerals from zero to nine by different persons were done by a headphone connected to our working system. We had recorded the voice using the in-built Matlab function "Wave record", which recorded sound using a PC-based audio input device. The recording for each sample is done for 2 sec. each with a standard sampling rate of 44100 Hz which is also approved by the International Phonetic Association as a Standard sample rate for Speech processing. We had given the numerical set to around 50 speakers (30 male and 20 female) to speak out at different expressions and environment. After recording, we have selected all the voices and convert the vector into a matrix form. As given as vector input, it creates a matrix one column at a time; 'vec2mat' function places extra entries in the output matrix if necessary.

TABLE 1

VOICE SAMPLES OF SIZE 10

Voice samples (10)	Recognition Rate (%)	Un-Recognized Rate (%)
Zero	70	30
One	70	30
Two	65	35
Three	85	15
Four	70	30
Five	72	28
Six	80	20
Seven	77	23
Eight	86	14
Nine	76	24

3.3 Training/Analysis

Our next phase of processing is the recorded samples to perform Linear Prediction filter Coefficients (LPC) and then apply Principal Component Analysis (princomp) on the data matrix which returns the principal component coefficients. These two operations are done to identify and extract the utterance features from the samples. After both the operations, the Data Matrix object is converted to a double-precision array for each of the sample stored in the database. After obtaining the various features of the voice samples, we had calculated the average value of the different voice samples. Various training algorithms that are available in the Neural Network Toolbox are 'trainlm', 'trainbr', 'trainbfg', 'trainrp', 'trainscg', 'traincgb', etc. among them we have used 'trainsig' for taking the input and single 'logsig' as output layer for the Eleman Network (newlm). The other hidden layers commonly have 'tansig' transfer function. We also had used two network training functions 'trainscg' and 'trainrp' which updates weights and bias values according to the scaled

TABLE 2

VOICE SAMPLES OF SIZE 20

Voice samples (10)	Recognition Rate (%)	Un-Recognized Rate (%)
Zero	75	25
One	80	20
Two	72	28
Three	80	20
Four	70.5	29.5
Five	80	35
Six	75	25
Seven	78	22
Eight	82	18
Nine	81	19

conjugate gradient method and according to the resilient back propagation algorithm for observing the different variation among the graph between the two different functions. Fig. 2 shows the graph of 'zero' speech using 'trainrp' algorithm

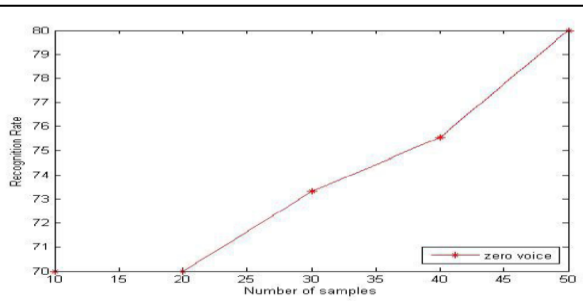

***Fig.2**. Graph for zero voice using 'trainrp' algorithm*

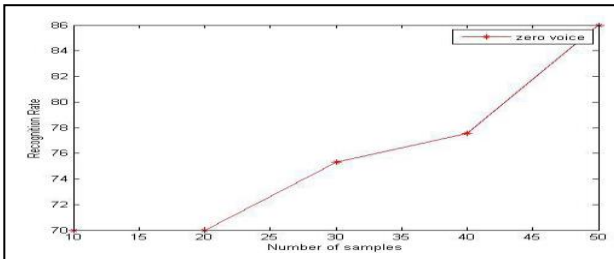

***Fig.3**. Graph for zero voice using 'trainscg' algorithm*

Fig. 3 shows the graph of 'zero' voice using 'trainscg' algorithm.

3.4 Identification/Testing

To measure the performance of a system it should be tested with a test data set. So we have tested our system with some other samples prepared by us. In the training phase, we had trained the neuron with 500 different samples of speech containing both male and female in different environments and expressions. During the testing phase, our program was tested with 10 male and 10 female voice samples. We have found the samples to be recognized. Fig. 9 is the comparisons graphs of our system. It shows that with the increase of several samples, the accuracy rate of the system increases.

4 RESULTS AND DISCUSSIONS

After taking the voice sample i.e. numeric (0-9) from both males and females, we have calculated the recognition percentage of the samples in different sets. We have tried to get the recognition rate of the samples by taking 10, 20, 30, 40 and 50 voices per spoken numeric digit. The results found and their corresponding graphs are shown in this chapter. TABLE 1 to TABLE 5 shows the recognized and unrecognized values of the digits with 10, 20, 30, 40 and 50 number of samples in tabular format. Fig. 4,5,6,7 and 8 show the graphs of input voices for 10, 20, 30, 40 and 50 samples of every numeral utterance. Fig. 9 is showing variation /comparison graphs of all samples starting from 10 to 50.

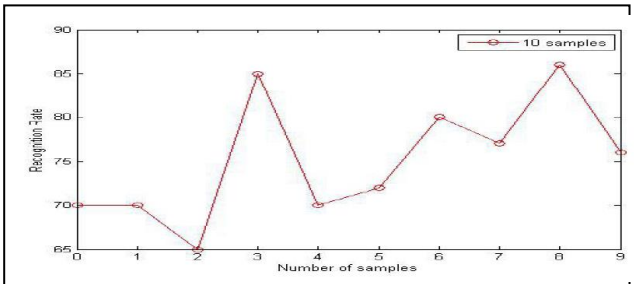

***Fig. 4**. Recognition graph for 10 voice samples*

TABLE 4

VOICE SAMPLES OF SIZE 40

Voice samples (10)	Recognition Rate (%)	Un-Recognized Rate (%)
Zero	79	21
One	87.5	12.5
Two	82.5	17.5
Three	82.5	17.5
Four	75	25
Five	72.5	27.5
Six	82.5	17.5
Seven	81.3	18.7
Eight	78	22
Nine	80	20

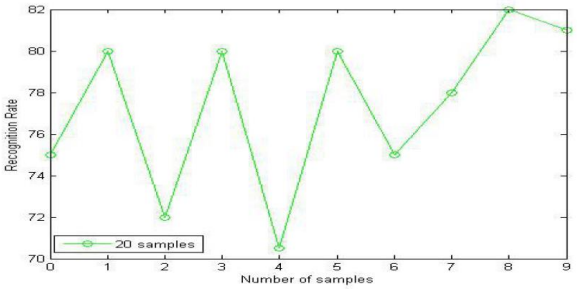

***Fig. 5**. Recognition graph for 20 voice samples*

TABLE 5

VOICE SAMPLES OF SIZE 50

Voice samples (10)	Recognition Rate (%)	Un-Recognized Rate (%)
Zero	82	18
One	88	12
Two	80	20
Three	88	12
Four	82	18
Five	80.8	19.2
Six	82.6	17.4
Seven	83	17
Eight	86	14
Nine	86	14

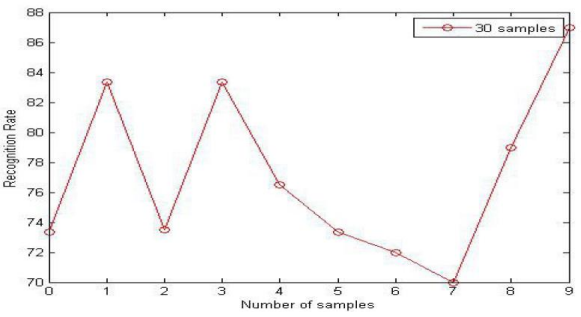

***Fig. 6**. Recognition graph for 30 voice samples*

TABLE 3

VOICE SAMPLES OF SIZE 30

Voice samples (10)	Recognition Rate (%)	Un-Recognized Rate (%)
Zero	73.33	26.67
One	83.33	16.67
Two	73.5	26.5
Three	83.33	16.67
Four	76.6	23.4
Five	73.33	26.67
Six	72	28
Seven	70	30
Eight	79	21
Nine	87	13

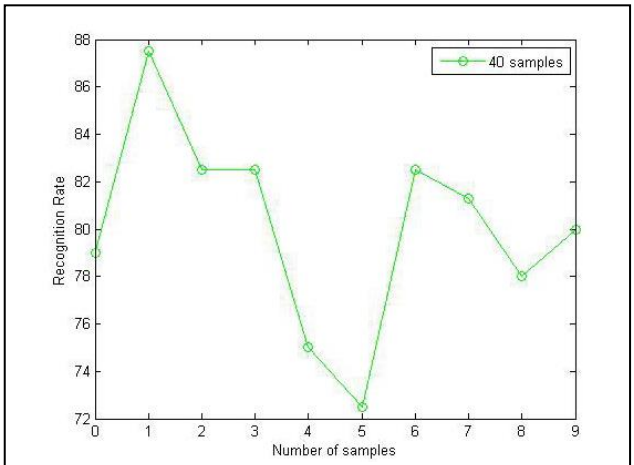

***Fig. 7**. Recognition graph for 40 voice samples*

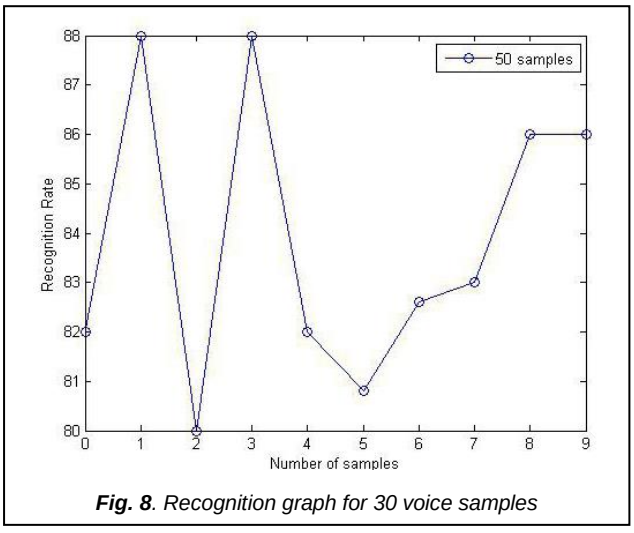

***Fig. 8**. Recognition graph for 30 voice samples*

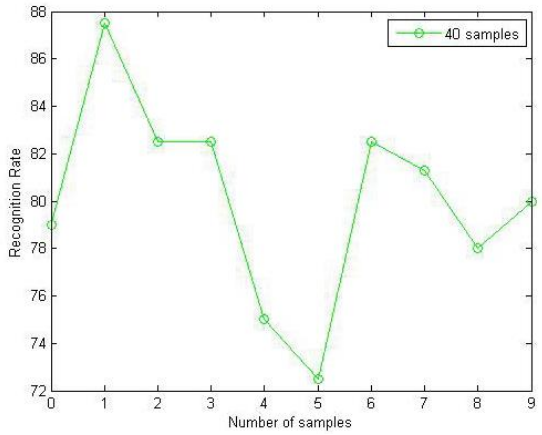

***Fig. 9**. Graph of variation of Recognition Rates for all number of samples.*

TABLE 6 shows the average recognition rates for different number of samples. Recognition rates here are calculated from individual voice samples respectively from TABLE 1, 2, 3, 4 and 5. At last we have found that with the increase of number of samples the accuracy of the recognition rate increases. The resulting output of each step of this research work is shown in snapshot of Fig. 10.

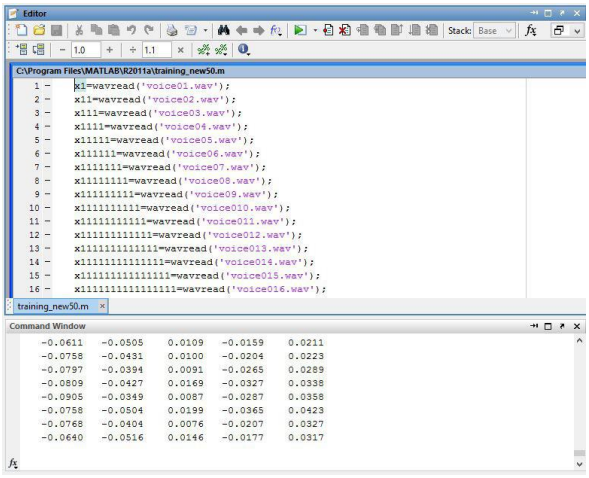

***Fig. 10**. Snapshot of Training phase*

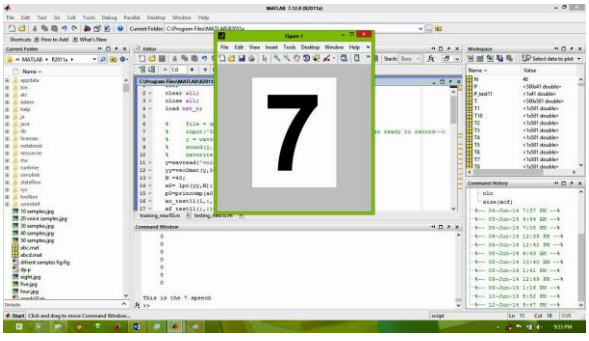

***Fig. 11.** Screen shot of final output*

5 CONCLUSION AND FUTURE WORK

This research work aims to identify English numerals from zero to nine using neural networks. Our dataset is limited to only Zero to Nine (0-9) English digit utterances. We have collected voice samples from different people both male and female. The system is tested against the voice signal of around 50 persons and the system gives approximately 82% accuracy. We have tried for some other numbers also with this model. After all these experiments, we can conclude that with this model English numerals can be recognized with an increased recognition rate which is quite satisfactory. In the future with the help of this we can develop recognition for the Vowels, words then sentences. We can also prepare the systems for various types of inputs. Also with the use of various tools and functions in MATLAB, we can design a system for continuous speech. This number of recognition can be applied to other non-English languages.

TABLE 6

AVERAGE RECOGNITION RATE

No. of Voice samples	Recognition Rate (%)
10	75.1
20	75.35
30	77.15
40	80.08
50	83.84

REFERENCES

[1] Siddika Imani, Parismita Sarma, and K. Samudravijaya. "AUTOMATIC IDENTIFICATION OF NATIVE LANGUAGE FROM SPOKEN ENGLISH." Proceedings in FRSM 2019, Kanpur, India, July 6-7, 2019

[2] H. Petkar, “A Review of Challenges in Automatic Speech Recognition,” International Journal of Computer Application(IJCA), Vol. 151(3), October 2016, pp:23-26.

[3] Koester, Heidi Horstmann, and Simon P. Levine. "User performance with continuous speech recognition systems." Proc. RESNA’00 (2000): 549-554.

[4] W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[5] Kamble, Bhushan C. "Speech Recognition Using Artificial Neural Network–A Review." Int. J. Comput. Commun. Instrum. Eng 3.1 (2016): 61-64.

[6] Katz, M. I. C. H. E. L. L. E., and A. Mbogho. "Speech Recognition Across South African Accents." Computer Science Department, University of Cape Town (2009).

[7] Nassif, Ali Bou, et al. "Speech recognition using deep neural networks: A systematic review." IEEE Access 7 (2019): 19143-19165.