TECHNICAL JOURNAL Vol 1, No.1, July 2019
Nepal Engineers' Association, Gandaki Province
ISSN : 2676-1416 (Print)
Pp.: 87-95

Performance Analysis and Recognition of Speech using Recurrent Neural Network

Er. Bishon Lamichanne
Er. Hari K.C.
Department of Electronics and Computer Engineering
Paschimanchal Campus, Institute of Engineering, Tribhuvan University
harikc@wrc.edu.np

Abstract
Speech is one of the most natural ways to communicate between people. It plays an important role in our daily lives. To make machines able to talk with people is a challenging but very useful task. A crucial step is to enable machines to recognize and understand what people are saying. Hence, speech recognition becomes a key technique providing an interface for communication between machines and humans. There has been a long research history on speech recognition. Neural network is known as a technique that has ability to classify nonlinear problem. Today, lots of research are going in the field of speech recognition with the help of the Neural Network. Even though positive results have been obtained from continuous study, research on minimizing the error rate is still gaining lots attention. The English language offers a number of challenges for speech recognition. This paper implements the RNN to analyze and recognize speech from the set of spoken words.

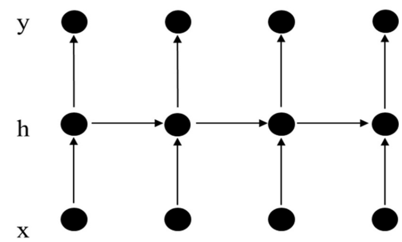
Keywords: *Recurrent Neural Network, speech recognition, spoken words, Artificial Intelligence, Machine, multi-layer perceptron*

Introduction
The first speech recognition system, a digit recognizer, was invented in 1952 in Bell lab. Since then, research on speech recognition has been carried out in both academia and industry and gained vast attention. Hidden Markov models (HMMs), were introduced into speech recognition in the 1970s, and became the cornerstone in the area of speech recognition. The standard HMM has been refined in a number of ways such as state clustering, adaptation and discriminative training in subsequent years. Significant progress has been achieved in speech recognition over the last several decades. In the early years, speech recognition was studied on isolated word recognition, with small vocabulary size (e.g. several hundreds). Native speakers spoke under clean environment, such as reading speech. Nowadays, large vocabulary (hundreds of thousands of words), continues speech recognition becomes the main research interest, such as voice search and

conversational telephone speech recognition. The speech is spontaneous under diverse acoustic environments, which is more similar to how people behave in their daily lives. Advance in computer hardware (multi-core CPU and GPU) and parallel algorithms also facilitate the use of dramatically increasing amount of training data. Nowadays, thousands of hours of speech and billions of text data can be used to train recognition systems. Various adaptation techniques are also developed to address the acoustic mismatch caused by speaker, noise, channel and so on. The improvement in performance can be also obtained from multi-microphone by overcoming reverberation and reducing noise. Recently, deep learning has attracted extensive research interests and presented significant improvement in performance over a range of tasks. With the big advance in speech recognition techniques, many companies have integrated speech recognition into products, such as Siri from Apple, Google watch from Google and speech translation in Skype from Microsoft. It is clear that the speech recognition techniques are entering our daily lives and gradually changing the way of life.

Recurrent Neural Network

RNN have feedback connections and address the temporal relationship of inputs by maintaining internal states that have memory. RNN are networks which have one or more feedback connection. A feedback connection passes the output of a neuron in a certain layer to the previous layer(s). RNN is different from MLP as RNN have feedforward connection for all neurons so RNN shows dynamic behavior is more suitable for speech recognition than MLP because it allows variability in input length.



x represents an input label.

y represents the output label.

h is the memory, computed from the past memory and current input.

Figure 1: Recurrent Neural Network

Problem Definition

Speech recognition is a computationally heavy task. To make this possible a fully connected network is essential. However, a fully connected network for speech is computationally difficult task. Generally, speech consists of noise and external factors that affects the quality of the speech. The degraded quality of speech causes the problem in the output. Training the system for correct recognition of speech and testing it to determine the accuracy and consistency of recognized character can make speech recognition a computationally heavy task.

Objectives

- To analyze and recognize the speech using Recurrent Neural network**.**
- To implement the speech recognition system.

Literature Review

Speech can vary from one person to another, one gender to another at different situations. This can cause the affect in the quality of the speech. Creating a network for speech recognition is a complex task. Speech recognition system can aid computers in Natural language processing in the different aspects of the computer applications. The main purpose of this project is to recognize speech from English speakers and to analyze the accuracy of the implemented algorithm. In this project the algorithm that is going to be implement is RNN.

A recurrent neural network (RNN) is a type of advanced artificial neural network (ANN) that involves directed cycles in memory. One aspect of recurrent neural networks is the ability to build on earlier types of networks with fixed-size input vectors and output vectors.

Abdel-Hamid, O. et.al. (2014) [1] described the way to implement CNN in speech recognition where they have proposed limited weight sharing scheme that can handle speech features in a better way. If CNN is used for image processing, values can be viewed as 2-D feature maps. But when CNN is used for speech recognition, the input "image" is considered as speech "spectrogram.

Sak, H. et.al. (2014, September) [2] proposed that for speech recognition LSTM RNN architecture is efficient. They proposed an alternative to standard architecture called Long Short-Term Memory Projected (LSTMP).

Graves, A. et.al. (2013, December) [3] compared the performance of DBLSTM-HMM hybrid with sequence training and measured the possibility of DBLSTM-HMM hybrids for large vocabulary speech recognition. Another advantage of using DBLSTM is that it is able to store the past and future context internally. So the data is presented in a single frame at a time.

Recurrent Neural Network using LSTM architecture has shown state-of-art performance on speech recognition. Zhang et.al (2016) [6] replaced Deep Neural Support Vector Machine for speech recognition (Zhag et.al (2015) [5]) by Recurrent Neural Network.

Methodology

Speech recognition system works in stages i.e training and testing phase. Training phase consists of Feature extraction followed by Generate/Update Reference whereas testing phase consists of Feature extraction, reference data, similarity measure followed by decision logic.

The block diagram of training and testing phase is shown in the figure below:

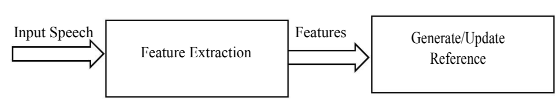


Figure 2: The Block Diagram of Training Phase

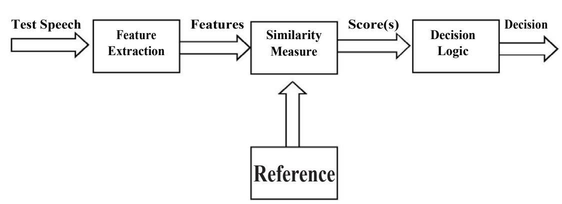


Figure 3: The Block Diagram of Testing Phase

Feature extraction

Audio features represent specific properties of audio signals. Feature extraction is the process that captures audio properties such as the fundamental frequency and the loudness of a signal. The outputs produced from feature extraction process are numerical descriptions of signals. The amount of raw data of the audio is too big, so the audio retrieval system cannot process the raw data of audio directly. Feature extraction process aims to reduce data by extracting the most meaningful information from signal which lead to reduces the dimensionality of the input vectors while maintaining the discriminating power of the signals.

Mel Frequency Cepstral Coefficients (MFFCs)

The main purpose of the MFCC is to copy the behavior of human ears. The derivation of MFCCs is done by the following steps:

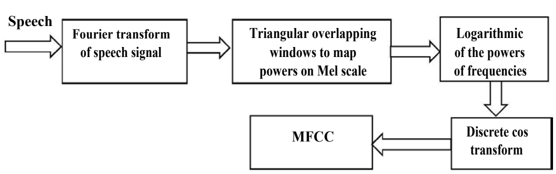


Figure 4: MFCCs Derivation

We start with a speech signal, the datasets used here are sampled at 8kHz.

1. Framing the signal into 20-40 ms frames. 25ms is standard. This means the frame length for an 8kHz signal is 0.025*8000 = 200 samples. Frame step is usually something like 10ms (80 samples), which allows some overlap to the frames. The first 200 sample frame starts at sample 0, the next 200 sample frame starts at sample 80 etc. until the end of the speech file is reached. If the speech file does not divide into an even number of frames, padding it with zeros so that it does.

The next steps are applied to every single frame, one set of 12 MFCC coefficients is extracted for each frame. A short aside on notation: we call our time domain signal S(n). Once it is framed we have Si(n)where n ranges over 1-200 (if our frames are 200 samples) and 'i' ranges over the number of frames. When we calculate the complex DFT, we get Si(k) where the 'i' denotes the frame number corresponding to the time-domain frame. Pi(k) is then the power spectrum of frame 'i'.

2. To take the Discrete Fourier Transform of the frame, we perform the following:

$$S_i(\mathrm{k}) = \sum_{n=1}^{N} s_i(n)h(n)e - j2\pi\mathrm{kn}/N \qquad 1 \le \mathrm{k} \le K$$

$$P_i(k) = \frac{1}{N} | S_i(k) |^2$$

This is called the Periodogram estimate of the power spectrum. We take the absolute value of the complex fourier transform, and square the result. We would generally perform a 512 point FFT and keep only the first 257 coefficents.

3. We compute the Mel-spaced filterbank. This is a set of 20-40 (26 is standard) triangular filters that we apply to the periodogram power spectral estimate from step 2. Our filterbank comes in the form of 26 vectors of length 257 (assuming the FFT settings fom step 2). Each vector is mostly zeros, but is non-zero for a certain section of the spectrum. To calculate filterbank energies we multiply each filterbank with the power spectrum, then add up the coefficents. Once this is performed we are left with 26 numbers that give us an indication of

how much energy was in each filterbank. Detailed explanation of how to calculate the
filterbanks is given below in figure 5.

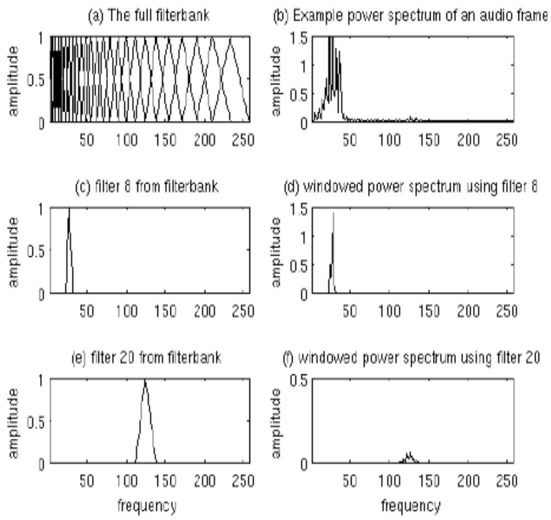


Figure 5: Plot of Mel Filterbank and windowed power spectrum

4. We take the log of each of the 26 energies from step 3. This leaves us with 26 log filterbank energies.
5. We take the Discrete Cosine Transform (DCT) of the 26 log filterbank energies to give 26 cepstral coefficents. For ASR, only the lower 12-13 of the 26 coefficients are kept.

The resulting features (12 numbers for each frame) are called Mel Frequency Cepstral Coefficients.

Generate/Update Reference

The MFCC features are extracted from audio files and feed for the training purpose and the reference is generated and updated with new values when neurons weights are updated.

Updated reference is used for testing purpose in testing phase. The trained model is saved in the same directory.

Similarity Measure

The similarity between the reference data and test data are calculated by comparing the MFCC features of reference data and test data. The calculated numerical values are known as scores and which are then fed to the decision logic.

Decision Logic

The aim of decision logic is to output a word based on the score obtained from the similarity measure. Here similarity of inputs with all the digits is calculated and the digit with highest similarity value is predicted.

Result and Analysis

Output

RNN is trained with 2400 spoken words collected with batch size 64, learning rate 0.0001,

test fraction 0.1 and variable number of epochs. The result obtained is discussed below.

The overall accuracy for implemented system is calculated using:

$$\text{Accuracy (\%)} = \frac{\text{Total of successful attempts}}{\text{Total Number of all attempts}} \times 100$$

Feature Extraction

The following figure 6 shows the MFCC features of an audio file.

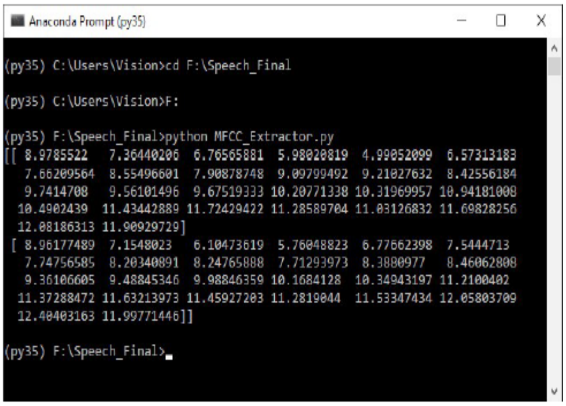


Figure 6 : MFCC feature of an audio file

Training and Testing RNN with 1 Epoch

RNN is trained with batch size 64, learning rate 0.0001 and number of epoch 1. The output obtained is shown in figure 7,8 and 9.

RNN was tested with 10% of test data set and gave the accuracy of 11.33%.

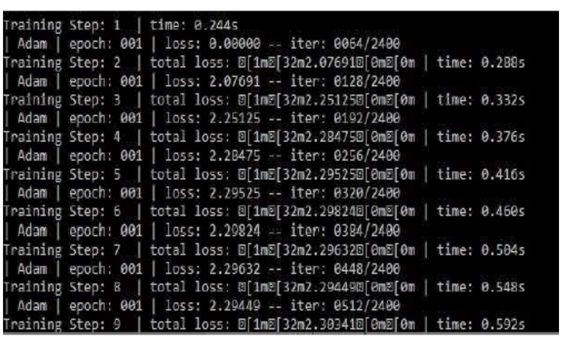


Figure 7 :Training RNN with 1 epoch.

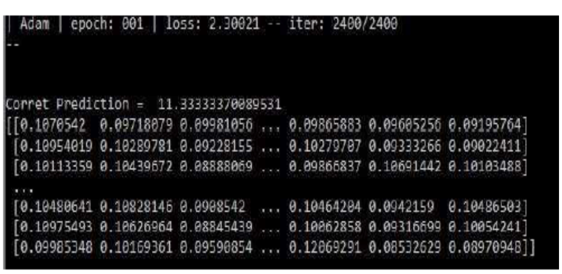


Figure 8: Testing RNN with 1 epoch

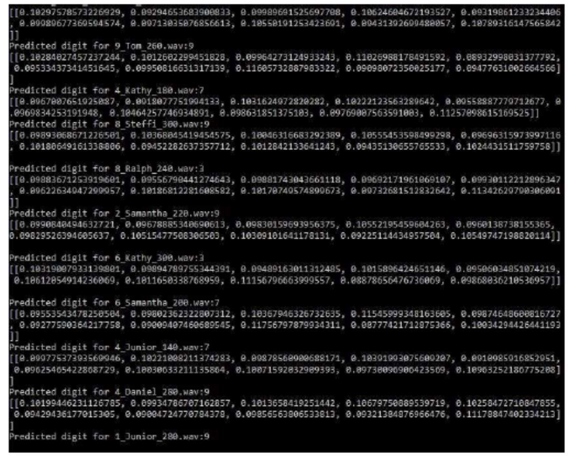


Figure 9 :Prediction of Spoken digits with trained model with 1 epoch.

Training and Testing RNN with 10 Epochs

RNN is trained with batch size 64, learning rate 0.0001 and number of epochs 10. The output obtained is shown in figure 10 and 11.

RNN was tested with 10% of test data set and gave the accuracy of 27.71%.

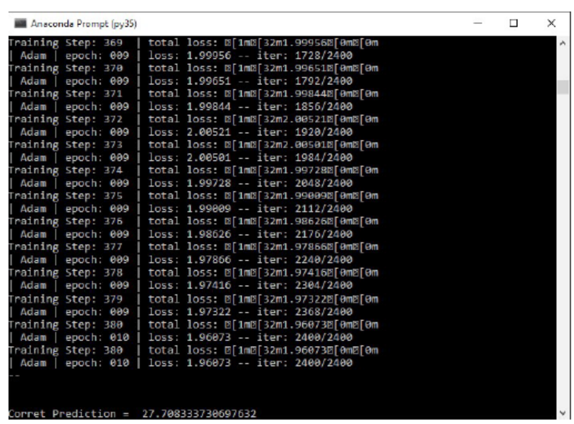


Figure 10 :Training and Testing RNN with 10 epochs.

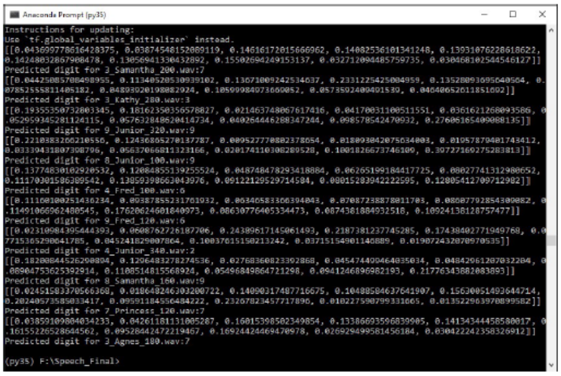


Figure 11: Prediction of Spoken digits with trained model with 10 epochs.

Training and Testing RNN with 100 Epochs

RNN is trained with batch size 64, learning rate 0.0001 and number of epochs 100.The output obtained is shown in figure 12 and 13.

RNN was tested with 10% of test data set and gave the accuracy of 50.75%.

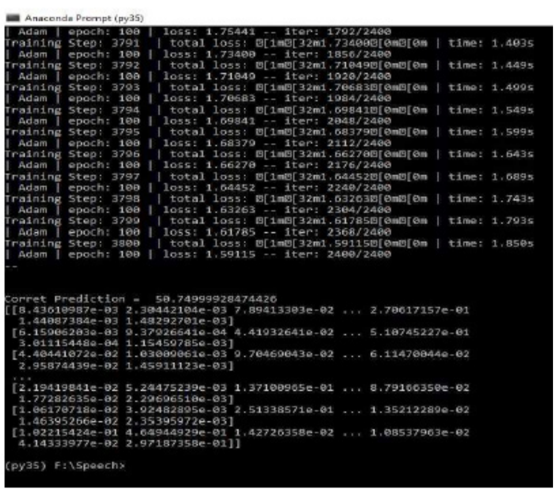


Figure 12: Training and Testing RNN with 100 epochs.

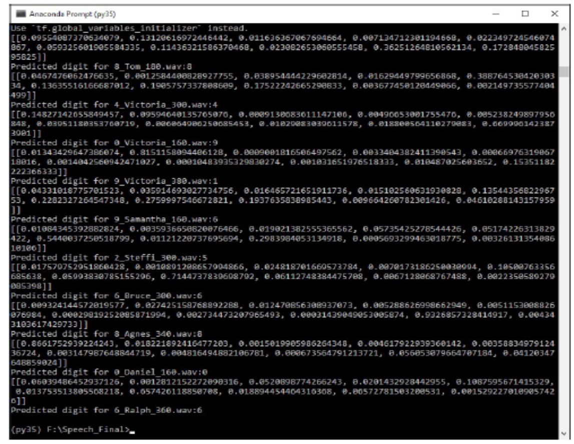


Figure 13: Prediction of Spoken digits with trained model with 100 epochs.

Training and Testing RNN with 1000 Epochs

RNN is trained with batch size 64, learning rate 0.0001 and number of epochs 1000.The

output obtained is shown in figure 14 and 15.

RNN was tested with 10% of test data set and gave the accuracy of 95.13%.

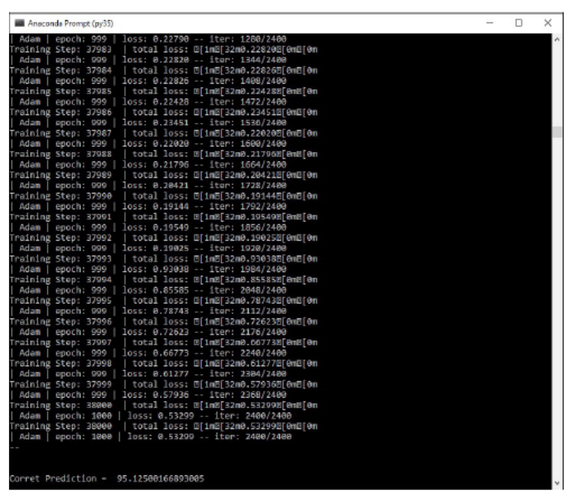


Figure 14: Training and Testing RNN with 1000 epochs.

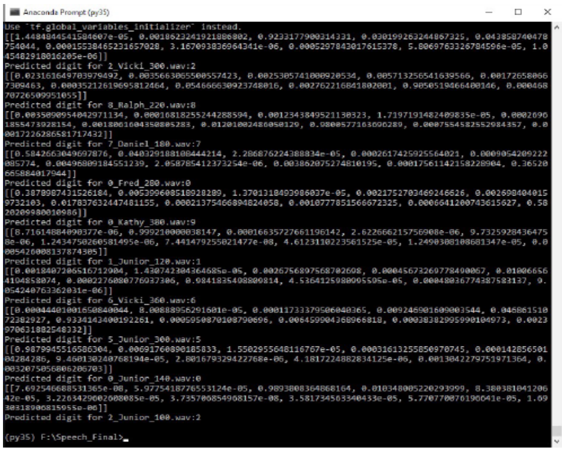


Figure 15: Prediction of Spoken digits with trained model with 1000 epoch.

Training and Testing RNN with 2000 Epochs

RNN is trained with batch size 64, learning rate 0.0001 and number of epochs 2000.The

output obtained is shown in figure 16 and 17.

RNN was tested with 10% of test data set and gave the accuracy of 98.25%.

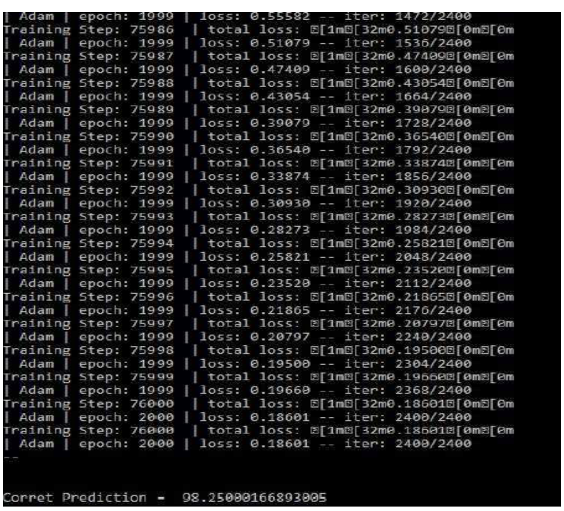


Figure 16: Training and Testing RNN with 2000 epochs.

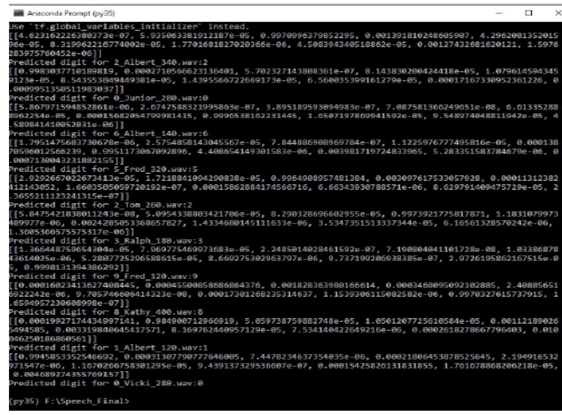


Figure 17: Prediction of Spoken digits with trained model with 2000 epochs.

The loss during the training started to reduce as the number of epochs increased.

The accuracy obtained for different number of epochs is plotted in the graph as shown in

figure 18 below.

From graph we can conclude that the accuracy has exponentially improved till the number of epochs reached 1000 and then the accuracy started to grow slowly after the number of epochs reached 2000.

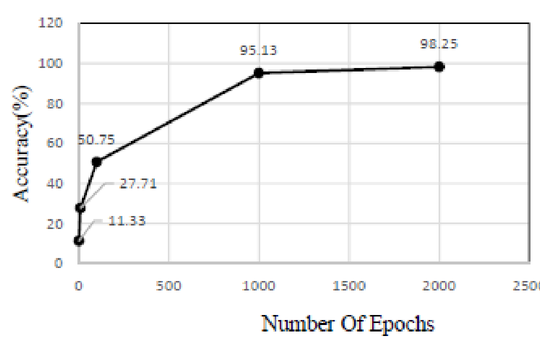


Figure 18: Graph showing variations of accuracy along with number of epochs

Conclusion

The algorithm is implemented over limited dataset (2400 altogether). The accuracy produced by RNN over given dataset depend upon the learning rate, number of epochs. The best accuracy obtained in this research is 98.25% with the learning rate 0.0001. The accuracy of the algorithm does not depend only upon the limited factors rather it is affected by the various factors like the quality of the training dataset, testing environment, number of datasets and so on. If we can gather huge amount of dataset then the accuracy can be increased to higher level and can be trained in different environments to obtain different results. The loss reduced as the number of epochs increased from 1 to 1000.When the loss is decreased then it increases the accuracy of the algorithm. The accuracy was plotted against number of epochs and visualized. From graph we can conclude that the accuracy is increasing as the number of epochs are increasing. Here the research is limited only with nine spoken words, but it can be extended to higher number of spoken words with higher computing power. Nowadays end to end speech recognition systems are more common which implement RNN as their fundamental algorithm. From the completed research, we can conclude that the RNN can be implemented in the environment where present output depends upon previous states as in voice recognition. RNN seems to be more efficient as all the neurons are connected together. The speech recognition systems can be implemented in the areas where voice commands are common such as assisting the blinds. Similarly, speech recognition increased the interest of people in regional language which helps to prevents languages from being extinct.

Reference

[1] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing, 22(10), 1533-1545.

[2] Sak, H., Senior, A. W., & Beaufays, F. (2014, September). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In INTERSPEECH (pp. 338-342).

[3] Graves, A., Jaitly, N., & Mohamed, A. R. (2013, December). Hybrid speech recognition with deep bidirectional LSTM. In Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on (pp. 273-278). IEEE.

[4] Song,W., & Cai, J. (2015) End-to-End Deep Neural Network for Automatic Speech Recognition.

[5] Zhang, S. X., Liu, C., Yao, K., & Gong, Y. (2015, April). Deep neural support vector machines for speech recognition. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4275-4279). IEEE.

[6] Zhang, S. X., Zhao, R., Liu, C., Li, J., & Gong, Y. (2016, March). Recurrent support vector machines for speech recognition. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5885-5889). IEEE.

[7] Mel Frequency Cepstral Coefficient (MFCC) Retrieved from: http://practicalcryptography.com/miscellaneous/machinelearning/guide-mel-frequency-cepstral-coefficients-mfccs/ Retrieved date: 17th May, 2018.