JES
Journal of Engineering Sciences

Vol 11, Issue 1, Jan /2020
ISSN NO:0377-9254

ANALYSIS OF SPEECH RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK

Mohit Bansal[1] and Dr. T. K. Thivakaran[2]

[1]*Student, Department of CSE, Presidency University, India*

[2]*Professor, Department of CSE, Presidency University, India*

***Abstract*— Nowadays in the current world, speech recognition has gained prominence and use with the rise of AI and intelligent assistants, such as Amazon Alexa, Apple Siri, Microsoft Cortana, Google assistant. Speech recognition is the ability of a machine or a program to identify words and phrases in spoken language and convert them to a machine-readable format. Speech recognition has many applications such as voice dialing, call routing, search keywords, simple data entry. The research aims to compare the performance to identify the best model between the two neural networks i.e., CNN (convolution neural network) and basic Neural Network for speech recognition. To find the best model among CNN and basic NN the audio data is collected from the internet and noise elimination on the audio data and data should be cleaned for further process, then the cleaned audio data is fed into CNN or basic NN architecture and trained with different layers. Finally, the trained model is checked for accuracy, validation accuracy and then the trained data is tested with test data to check test accuracy of a given Model.**

***Keywords*— CNN, Basic NN, spectrogram, accuracy.**

1. INTRODUCTION

Today in the current generation speech recognition is playing a major role in most of the fields such as smartphones, Tv, voice call routing, voice dialing, search keywords, simple data entry. We all know that whenever we call to any customer care service there will be a virtual assistant to assist us

before we reach to main person to whom we want to talk, the technique used here was Call routing which refers to the procedure of sending voice calls to a specific queue based on predetermined criteria. A call routing system is also known as an automatic call distributor (ACD). Since traditional models of customer service were based on phone support or call support as one of the primary methods of contact between customers and companies for business purposes, the procedure of sending calls to the right agent became very much important. Today, modern agents interact with customers through a variety of channels.

In earlier days most of the people use to call by typing the number in the phone but nowadays voice-enabled calling is also available where people use to call anyone through their voice without typing any number in the phone, this has made easier to the people but the problem is if anyone is in such a place where there is more noise and more disturbance then voice-enabled calling may not work correctly as more than two or more voices mixes, where it will be difficult for a system to recognize our voice in such a worst environment. To overcome this best noise elimination technique should be used where it can eliminate noise up to a level. Voice-enabled calling is also known as voice dialing which uses speech recognition software.

In the present world, it is possible to entry some of the data in excel or word documents using our voice which enables you to do hands-free data entry

by dictating the text or numbers that we want to be entered in the current cell and to issue voice commands that allow you to choose menu items, dialog box options, or even toolbar buttons by simply saying their names. This saves our time and work can be done faster compare to typing work. Even for voice data entry speech recognition software have been used. When using Speech Recognition to dictate data entries, we need to keep the microphone close to our mouth and in the same position as you dictate. Depending upon the microphone quality we need to speak normally and in a low but not monotone voice, pausing only when you come to the end of a thought or the data entry for that cell and it takes time for our computer to process our speech, and therefore, depending upon the speed of your processor, it may take some time before your words appear on the Formula bar and in the current cell. This can be improved by training more and more audio data using deep learning or machine learning algorithms.

As we know Google Assistant and Microsoft Cortana are widely used nowadays such as searching for information on the internet or it may search for information on the computer such as files, folders, documents, and many other things. All these are done through our voice, whatever ever we speak, or we tell Google Assistant and Microsoft Cortana it will search and gives us a piece of information about what we required. Therefore, Speech recognition is playing a major role in Google assistant, Microsoft Cortana, and Apple Siri. Day by Day the accuracy of converting from speech to text in Google Assistant, Microsoft Cortana, and Apple Siri is increasing.

Think of a situation where you want to share your feelings to someone or you want someone to entertain you in your sad times then Amazon Alexa or amazon echo can be used, whenever we speak it will listen to us and give some information like if we tell "Alexa sing a song" it will sing some song for us or if we tell "Alexa tell us some news about today" so Alexa will tell us the news about today, Amazon Alexa is speech recognition device which recognizes our speech and depending upon that it gives some output.

2. LITERATURE REVIEW

Speaker Recognition is a technique used to automatically recognize a speaker from a recording of their voice or speech utterance. In [1] They have presented automatic speaker recognition of Sepedi home language speakers using four classifiers namely Support vector machines (SVM), K-Nearest Neighbors, Multilayer Perceptron (MLP) and Random Forest (RF) which are trained using WEKA data mining tool. The performance of each Model was evaluated using 10-fold cross validation. The Best accuracy was achieved by two classifier Multilayer Perceptron (MLP) with 97% of accuracy and Random Forest with 99.9% of accuracy. Finally, RF model was implemented on a graphical user interface for development testing.

Automatic Speech Recognition (ASR) and spoken language understanding are one of the most important part of applied machine intelligence. In [2] they have focused on isolated voice command recognition for autonomous man-machine and intelligent robotic systems, they have created grammar model for small testing of command set with self-loops for each state to return blank symbols for noise and out of vocabulary words. They have compared recognition accuracy and average decision-making time of our approach with the state of the art continuous speech recognition engines based on language models and it has been experimentally proved that their approach was achieved 60% higher accuracy than conventional offline speech recognition methods based on language models.

With the advance in the Deep learning, the performance of automatic speech recognition (ASR) has improved

rapidly [2]. In [3] They have presented a method of enhancing automatic speech recognition dataset with an immature pre-trained model and script. They have compared the chunks obtained from the pre-trained model with ground truth script and produced the pair of an audio and its script. In each pair, the audio has beginning and end of an utterance, and the script is since they have used human-written script. The method which they have used extracted automatic speech recognition dataset in exact and effective manner.

The voice is most important and one of the natural forms of communication among from livelihood. In [4] they have done survey on speech recognition approaches and techniques where from longer time most of the people are working in a area of voice recognition and communication. According to [4] most of the researchers also contributed in the field of speaker dependent and speaker independent voice recognition. They have discussed various approaches available for developing an ASR system with advantages and disadvantages and it is shown that ASR system performance depends upon two factors which is feature extraction techniques and speech recognition approach for language.

In [5] They have proposed a neural vocoder-based test-to-speech (TTS) system that effectively utilizes a source-filter modeling framework. They have used two neural vocoder algorithms such as SampleRNN and WaveNet which are well known to generate high-quality speech, but its generation speed was very slow to be used for real-world applications. They have trained two models separately i.e. training the spectrum or acoustic parameters with a long short-term memory model and the excitation component with a SampleRNN-based generative Model. Finally, the results confirm that the superiority of the proposed system over a glottal modeling-based parametric and original SampleRNN-based speech synthesis systems.

3. EXPERIMENT

Literature review was done on 13 research papers and founded that convolutional neural network provided better performance on speech recognition compare to other neural networks when audio data is converted into image and fed into CNN.

3.1. IMPLEMENTATION

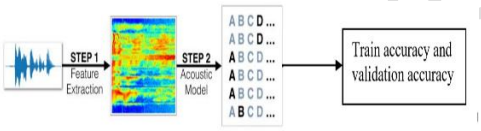

Fig. 1 Block diagram for implementing CNN

The TensorFlow speech command dataset was collected from online which consist of 30 classes out of which 21 classes where commands namely down, eight, five, four, go, happy, left, nine, no, off, on, one, right, seven, six, stop, three, two, up, yes, zero and others were background noises. The data which was collected are analyzed and converted into spectrogram only for yes, nine, no, up, two speech commands to extract and represent some features from it and then the data has been trained using Convolutional neural network and classification result was obtained and model was created which was later used to predict.

3.2. DATASET

There are not so many publicly available datasets that can be used for simple audio recognition problems. Luckily, Google's TensorFlow and AIY teams have created freely available Speech Commands Dataset. This contains around 65000 one-second sound files with commands like Go, Yes or Stop. In the Experiment only 5 speech commands to train convolutional neural network model namely yes, nine, no, up, two.

3.3. FEATURE EXTRACTION

To extract feature from dataset one audio file was taken from each of 5 speech command and plotted the graph between amplitude and time to see the difference between all the 5 commands. The below figure 3.1 shows amplitude vs time graph for one audio of each 5-speech command and graph show that how amplitude is varying with time for different speech commands, amplitude vs time graph was used to check how our audio signal looks.

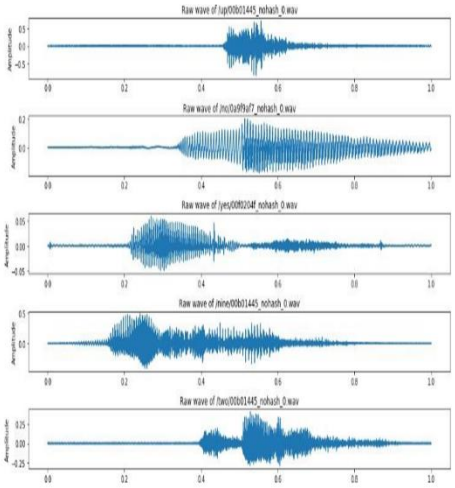

Fig. 2 Amplitude vs time graph for up, no, yes, nine, two speech commands.

Then the audio data from each speech command out of 5 was converted into log spectrogram to extract some features like frequency and time and to feed those features into our model. The spectrogram is a representation of audio file in a frequency domain (instead of a temporal domain as it was for a raw data). In order to convert raw data to spectrograms I have applied Short-time Fourier Transform (STFT).

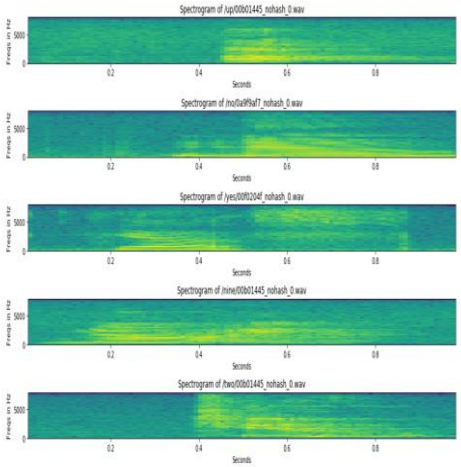

Fig. 3 Log spectrogram of yes, nine, two speech command

3.4. TRAINING MODEL

The Basic Neural Network consist of one input layer, one flattens input layer, 3 dense layer and one output layer. The Spectrogram which was generated from the audio data was fed into Basic neural network as input with three dense layers.

CNN use additional layers at the beginning of the neural network to reduce the size and preprocess an image. The basic architecture of CNN includes: Convolutional Layer– uses convolutional layer to filter required input signal and extract some more additional image features from the audio data, Activation Function – applies non-linear function to the given data such as rectifier,relu to the outputs of convolutional layer, Pooling Layer – It performs a down sampling operation reducing the size of an input with max() or sum() operation Fully-Connected Layer – each neuron in the previous layer is connected to each neuron on the next layer with last such layer producing outputs of neural network. In other words, convolutional and pooling layers represent high-level features of the input image. The pooling layer reduces the size of an image to control overfitting. Moreover, convolutional and pooling layers are still valid to use during backpropagation algorithm so that the neural network can be still trained using

gradient descent approaches. In the experiment it has been implemented three convolutional layers. In each case the layer has 32 filters, 3×3 window with stride=1 and same padding. The output generated was 4D array of size (32,177,98,1). For each batch CNN accepts 3D array, therefore it has been artificially extended the last dimension of a log- spectrogram.

4. RESULT ANALYSIS

Trained Speech command dataset for 5 speech command and obtained better performance and accuracy on train, test and validation data using convolutional Neural Network compare to Basic Neural Network. From fig 3.5 and fig 3.6 it has been seen that the model which was trained using convolutional Neural Network has achieved 99 % of accuracy on train dataset ,79 % accuracy on validation and 91% accuracy on test dataset and Model which was trained using basic Neural Network as achieved 20% of accuracy on train dataset, 19% of accuracy on validation and 19% of accuracy on test dataset. An early stopping has been implemented while training the model so that the model do not overfit and it stops the execution if model try to overfit, for CNN model it has been executed for all 15 epoch but for basic Neural Network it has been executed only for 5 epoch.

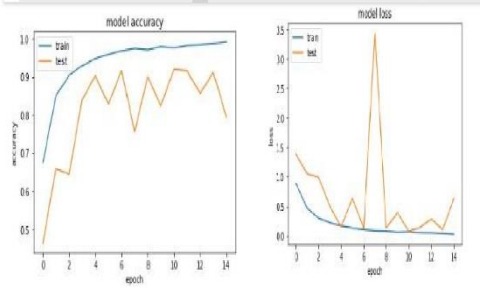

Fig. 4 Model accuracy and Model loss for CNN

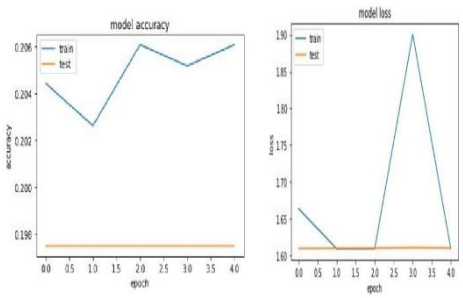

Fig. 5 Model accuracy and Model loss for Basic NN

Fig. 4 and Fig. 5 shows the plot between accuracy vs epoch and loss vs epoch. Validation loss is very much poor in Basic NN and it just the straight line which means the Basic NN model is under fitting the model. On the other hand, it has been seen that CNN model perform well on used dataset and it give better accuracy on train dataset compare to Basic NN. Even we can see that there is huge loss in Basic NN model on given dataset and on CNN model the loss is less compare to Basic CNN model.

	CNN Model		Basic NN Model	
No. of Epochs	Accuracy	Validation Accuracy	Accuracy	Validation Accuracy
1	67.56 %	46.48 %	20.44%	19.75%
2	84.93 %	65.86 %	20.26%	19.75%
3	90.35 %	64.36 %	20.61%	19.75%
4	92.86 %	83.99 %	20.52%	19.75%
5	94.72 %	90.25 %	20.61%	19.75%
6	95.82 %	82.72 %	-	-
7	96.73 %	91.57 %	-	-
8	97.40 %	75.56 %	-	-
9	97.02 %	89.89 %	-	-
10	97.86 %	82.30 %	-	-
11	97.49 %	91.87 %	-	-
12	98.18 %	91.57 %	-	-
13	98.31 %	85.55 %	-	-
14	98.63 %	91.03 %	-	-
15	99.13 %	79.41 %	-	-

Table I. Accuracy obtained from CNN and Basic NN for speech command dataset.

5. CONCLUSION AND FUTURE WORK

The speech command dataset which was collected from online had 30 classes from which 5 classes was selected namely up, no, yes, nine, two for extracting features, training model and testing the model. The data was split into 3 parts i.e. train, test and validation and was converted into log spectrogram which have been fed into convolutional Neural network, from result analysis it has been proved that Convolutional Neural network performed well on 5 speech command which gave train accuracy of 99% and validation accuracy of 79% and the model was tested on test data which gave 91% of accuracy compare to basic Neural Network which gave train accuracy of 20% and validation accuracy of 19% and the model was tested on test data which gave 19%, from this it has been concluded that Convolutional Neural Network performed well on images and gave better accuracy. After Implementation of Basic Neural Network and Convolutional Neural Network for speech command dataset the future work is towards implementing Recurrent Neural Network for same speech command dataset and hybrid model of CNN and RNN on same speech command dataset.

REFERENCES

[1] Tumisho Billson Mokgonyane, Tshephisho Joseph Sefara, Thipe Isaiah Modipa et.al, "Automatic Speaker Recognition System based on Machine Learning Algorithms" SAUPEC/ RobMech/PRASA Conference, 2019

[2] Artem Sokolov, Andrey V. Savchenko, "Voice command recognition in intelligent systems using deep neural networks" IEEE 17th World SAMI,2019

[3] Minsu Kwon, Ho-in Choi, "Automatic Speech Recognition Dataset Augmentation with Pre-Trained Model and Script" IEEE BigComp ,2019

[4] Atma Prakash Singh, Ravindra Nath, Santhosh Kumar, "A Survey: Speech Recognition Approaches and Techniques",5th IEEE UPCON,2018

[5] Kyungguen Byun, Eunwoo Song, Jinseob Kim, Jae-Min Kim and Hong-Goo Kang , "Excitation by Sample RNN model for Text-to-Speech", 10.1109/ITC-CSCC.2019.8793459,2019

[6] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh and Khaled Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review" IEEE Access vol 7,2019

[7] Saeed Mian Qaisar, Raviha Khan, Noofa Hammad, "Scene to Text Conversion and Pronunciation for Visually Impaired People", IEEE ASET,2019

[8] Bagus Tris Atmaja, Masato Akagi, "Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model", IEEE International conference (ICSigSys), 2019

[9] Tara N. Sainath, Brian Kingsbury, Abdel-Rahman Mohamed, George E Dahl et.al, "Improvements to Deep Convolutional Neural Networks For LVCSR" arXiv:1309.1501v3, 2013.

[10] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional Neural Networks for Speech Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.22,2010.

[11] Xuejiao Li, Zixuan Zhou, "Speech Command Recognition with Convolutional Neural Network", CS229 Stanford education, 2017.

[12] Ying Zhang, Mohammad Pezeshki, Philemon Brakel, Saizheng Zhang,

Cesar Laurent Yoshua Bengio, Aaron Courville, "Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks", arXiv:1701.02720v1,2017.

[13] D. Nagajyothi, P. Siddaiah, "Speech Recognition Using Convolutional Neural Networks", IJET 133-137, 2018.