Review on Different Speech Recognition Models

Dr. R. Jayalakshmi[1], Yetukuri Chandrakala[2], Talabattula Manikanta[3], Voleti Nagendra Kumar[4], Tanguturi Venkata Sai Koundinya[5], Tapala Manoj[6], Veluru Venkat Reddy[7]
[1]*Assistant Professor, Department of Electronics and Communication Engineering, SCSVMV University, Kanchipuram, India*
[2,3,4,5,6,7]*UG Students, Department of Electronics and Communication Engineering, SCSVMV University, Kanchipuram, India*

***Abstract*— Language plays a significant role in human life as it allows us to communicate, express ourselves and understand others. It is the primary means of communication. Many research activities are conducted on automatic speech recognition. The Major drawback of the ASR systems is their Efficiency. The overall performance of the ASR system Depends mostly on the Acoustic model and also affected by the environment. For ASR systems Mostly we use Deep learning techniques like Recurrent neural networks used for speech recognition, voice recognition, time series prediction and natural language processing and Convolutional Neural Network which uses different modules for speech emotion recognition and classifiers are used to Differentiate emotions such as happiness, surprise, Anger, neural state, sadness etc.**

Keywords— acoustic model, language model, MFCC, HMM, Word error rate, Deep Neural Networks, Recurrent Neural Networks, Convolutional Neural Networks.

I.INTRODUCTION

In recent years Automatic Speech Recognition is one of the most advancing Field for research Automatic Speech Recognition (ASR) is a rapidly due to its growing applications in various domains. ASR technology enables machines to recognize and transcribe spoken audio into written text, enabling efficient communication between humans and machines. With the advancement of deep learning techniques, several automatic speech recognition models have been proposed in recent years, which improves the efficiency of the system and also for improving the accuracy of the system. In this paper, we present a comprehensive review of different automatic speech recognition models, including their architecture, training methodologies, different speech recognition Models and applications. We analyse the strengths and weaknesses of various ASR models and highlight recent advancements and research trends. Our review aims to give detailed information about the speech recognition models and their potentials which can serve as a valuable resource for researchers, practitioners, and industry professionals in the field of speech recognition.

Currently ASR systems are used in many applications such as Live captioning and transcript, method of typing on computer or phone without typing, Virtual assistants and chatbots, Voice commands and dictation.

The performance of the ASR systems are affected by the following factors:

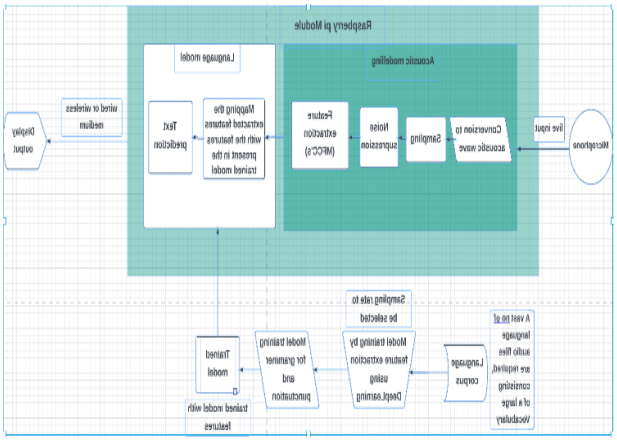
1. Speech Variability: Speech can vary in terms of accent, dialect, phoneme pronunciation, and speech rate, among other factors, making it difficult for the automatic speech recognition system to accurately recognize and transcribe the spoken words.
2. Noise: The Noise can interfere with the input audio signal which results in reduction in the efficiency of the output which can be shown in text format.
3. Speaker Variability: The voice characteristics of different speakers vary, and this variation can impact the performance of an automatic speech recognition system.
4.Vocabulary Size: The size of the vocabulary that an automatic speech recognition system can handle influences the system's accuracy.
5.Training Corpus: Large language audio files and text files are given to train the model. The accuracy will be more if the data given to train the model is high since the chances of predicting the words spoken will be high.
6. Domain Specificity: The ASR systems trained for a particular domain may not work accurately in different domains or topics.
7. Speech style: The speech recognition accuracy could be affected based on the tone and style of the speech.

II.ARCHITECTURE OF SPEECH RECOGNITION SYSTEM

Acoustic Modelling: Acoustic modelling is a process of audio signal processing where the signal undergo sampling to find the letter or word in a sample. It converts the input Analog audio signal into digital signal. The primary models used for acoustic modelling are Gaussian Mixture models and Hidden Markov Models. In acoustic model we take raw audio recordings of speech and compile them into statistical representation of sounds that make words. Every spoken word is made of basic elements of a specific language which are known as phonemes. Then HMM of speech and unit of speech is depicted by the acoustic model.

Model training: a large number of language audio files required which consists of large vocabulary to train the model. The model will be trained with feature extraction using deep learning technique. The model also trains for grammar and punctuations.

Language model: language model takes the input from the acoustic model and trained model. It maps the extracted features along with features present in trained model with this the system predicts the text.



1.CONVOLUTION NEURAL NETWORKS FOR AUTOMATIC SPEECH RECOGNITION

Convolution Neural Network is regarded as variant of standard Neural Networks. In convolution neural networks they use a special network structure which consists of alternating convolution and pooling layers.

A. ORGANIZATION OF THE INPUT DATA TO CNN

In pattern recognition using CNN the input data need to be organized as number of feature maps to fed into the CNN.

CNNs utilize a small window known as filter that can scan the input image at both training and testing time. This allows the network to learn from various features of the input data regardless of the position of the image within the networks. CNNs use weight sharing, full weight sharing is used which refers to the use of same weights at every position of the window during convolution operation.

CNNs are often also referred as local because computations of individual units of the window depend on features of the local region of the image. In this, speech feature vectors must be arranged into feature maps appropriate for CNN processing.

B. Convolution Ply

In a convolutional layer of a CNN, each input feature map (assuming a total number of feature maps) is connected to multiple feature maps (assuming a total number) through local weight matrices (assuming a total of weight matrices). Mapping is one of the well-known convolution operation in signal processing.

There are two important aspects in which a convolution ply is different from standard fully connected hidden layer. First, each convolution unit receives input only from a local area of input which means that it represents features of a local region. The convolution ply units are arranged into multiple feature maps, where units within the same feature map share identical weights but receive input from diverse locations in lower layer.

C. POOLING PLY

In a CNN, a pooling operation is performed on the convolutional layer (convolution ply) to generate a corresponding pooling layer (pooling ply). The pooling ply also consists of feature maps, with the same number of maps as the convolution ply, but each map is smaller in size. The purpose of the pooling ply is to reduce the resolution of the feature maps, resulting in generalized representations of the features from the lower convolution layer.

D. Learning Weights in the CNN

The weights in the convolutional layer can be learned using the error back-propagation algorithm, which is

commonly used in neural networks for updating weights based on the prediction errors. However, some special modifications are required to account for the sparse connections and weight sharing properties of convolutional layers.

The handling of shared weights in the convolutional layer is distinct from that in fully-connected deep neural networks, as weight sharing is a unique characteristic of convolutional layers. In fully-connected networks, there is no weight sharing, and each neuron has its own set of weights.

E. TRAINING OF CNN LAYERS

RBM-based pretraining improves the performance of the deep neural networks when the size of the training set is small.

F. ENERGY FEATURES

In Automatic Speech Recognition (ASR), the calculation of log-energy is done on the basis of per frame energy along with other spectral features. It is not suitable to treat energy like other filter bank energies because it is the sum of energies in all frequency bands and it is not depending on the frequencies.

G. BENIFITS OF CNNs

They have key features like pooling, weight sharing and locality. The locality allows more robustness and also reduces the number of weights to learn. Weight sharing reduces the over fitting and also improves model robustness. Suitable pool size needed to be chosen for state labels localization.

2.RECURRENT NEURAL NETWORKS FOR AUTOMATIC SPEECH RECOGNITION

Recurrent Neural Networks can be used for sequential data. RNNs possess feedback connections and utilize the internal states that have memory to address the temporal relationships of inputs. Recurrent Neural Networks are used for predictions. For predicting Recurrent Neural Network will use past outputs. RNN networks have one or more feedback connections. A feedback connection is used to send the output of a neuron in one layer to the previous layers.

The built-in memory in RNN helps them to consider the information from previous predictions so that the RNN systems accuracy increases in future predictions.

The RNN model accuracy if we train the model. it will give better accuracies than the previous results.

Recurrent Neural Networks have feed forward connections for all neurons and the connections allow the network to show dynamic behaviour.

If we consider for example, if we have said “ANDHRA” then it is obvious others will say “PRADESH” next to complete it is less likely for someone to pronounce other words than “PRADESH”. so having the knowledge of previous predictions improve the accuracy of the model. RNN allows variability in input length. RNN networks can confer better performance and can learn in shorter duration compared to conventional feed forward networks. RNNs have ability to process short-term spectral features but can also respond to long-term temporal events. we are using this advantage of RNNs in speech recognition. If the duration of utterance increases the system also improves.

RNN ARCHITECTURE

RNN is using Backpropagation through time as learning algorithm. This architecture is better than MLP in phoneme recognition. It terms of accuracy by using back propagation algorithm.

The back propagation through time (BPTT) algorithm is based on converting the network from A feedback to purely forward system by folding the network overtime. If the network needs to process a signal with a length of multiple time steps, the network can be replicated and the feedback connections can be adjusted to act as shared weights, the entire network can be trained as a single large feed forward network.

MULTI-LAYER PERCEPTRON

It is the most popular and in use network architecture nowadays, due originally to Rumelhart and McClelland(1986). Each unit in the network computes a weighted sum of its inputs, incorporating biases, and passes this sum through a transfer function to generate its output. The units are organized in a layered feedforward structure, allowing for a straightforward interpretation of the network as an input-output model. The input, hidden and output layers are the three layers.

TRAINING PHASE

The multilayer backpropagation algorithm is a common method used to train neural networks for recognizing spoken characters, with separate training performed for each speaker. It compares the predicted output with the actual output for a given speaker and adjusts it to

minimize the dissimilarities. By training the network captures unique speech pattern and characteristics.

TESTING PHASE

We need to test the accuracies of the model for more speakers and then can see the increase in percentage of accuracy increases after training the model.

III.CONCLUSION

In this paper, we have discussed about different speech recognition models. Mainly we discussed about Convolution Neural Networks and Recurrent Neural Networks. We have seen performance improvement in these compared to the standard Deep Neural Networks. In CNNs key features like pooling, weight sharing and locality are playing important role in system efficiency. The RNN model can increase the accuracy by training the system. The performance in RNN is depend on the quality of pre-processing. The results obtained in this study demonstrate that CNN and RNN are best suitable methods for speech recognition.

REFERENCE

1. Yogesh kumar, Navdeep singh, “A Comprehensive view of Automatic speech Recognition system – A systematic literature Review”, 2019 international conference on Automation, Computational and Technology Management (ICATM) Amity University.
2. SADEEN ALHARBI, MUNA ALRAZGAN ALRASHED, TURKIAYH ALNOMASI, RAGHAD ALMOJEL, RIMAH ALHARBI, SAJA ALHARBI, SAHAR ALTURKI, FATIMAH ALSHEHRI, AND MAHA ALMOIL, “Automatic speech Recognition: Systematic Literature Review”, Received August 8, 202, accepted September 2, 2021, date of publication on September 14, 2021, date of current version October 1, 2021.
3.Ossana 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, NO. 10, OCTOBER 2014.
4.Aditya Amberkar, Gaurav Deshmukh, Parikshit Awasarmol, Piyush Darse, “Speech Recognition using Recurrent Neural Networks”, proceeding of 2018 IEEE International Conference on current trends toward Converging Technologies, Coimbatore, India.
5. Sruthi Vandhana T, Srivibhushanaa s, Sidharth K, Sanoj C S, “Automatic Speech Recognition Using Recurrent Neural Network”, International Journal of Engineering Research and Technology (IJERT) ISSN:2278-0181, vol.9 Issue 08, August-2020.
6. Dr.R.L.K. Venkateswarlu, Dr.R.Vasantha Kumari, G.Vani Jayasri, “Speech Recognition by using Recurrent Neural Networks”, international journal of scientific and engineering RESEARCH VOLUME 2, Issue 6, June-2011, ISSN 2229-5515.