Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

Automatic Speech Recognition using Recurrent Neural Network

Sruthi Vandhana T
Undergraduate, Department of CSE
SVCE
Chennai, Tamil Nadu, India

Srivibhushanaa S
Undergraduate, Department of CSE
SVCE
Chennai, Tamil Nadu, India

Sidharth K
Undergraduate, Department of CSE
SVCE
Chennai, Tamil Nadu, India

Sanoj C S
Assistant Professor
Department of CSE
SVCE
Chennai, Tamil Nadu, India

***Abstract*— Speech recognition is one of the major developing field in computer science, more people now focusing on speech recognition as like others. Automatic Speech recognition or speech to text conversion is nothing but converting the given input speech of the user into the text or query format depending upon the usage of where the speech action takes place. Nowadays more over every people is using this technology instead of typing or using buttons to give a specific command. Speech to text is an very intriguing task as the sounds of different word hear similar and same word has different sounds depends on the people. So its hard to do the speech to any form of conversion. We have proposed a system which is a simple query processing system for the railways where the input is speech and an output is a text being displayed. The process is converting the input speech into the query for processing the railway system queries.**

Keywords—Speech Recognition, Railways, Neural Network

1. INTRODUCTION

Speech is human's most efficient way of communication nowadays. There is always some problem occurs with the communication with the computers and systems, yet speech recognition is one of the way to resolve this problem. But this is always been a most challenging tasks to achieve. Speech recognition is an interdisciplinary subfield of computer science and Natural language Processing that develops methodologies and technologies to enable the recognition and translation of spoken language into text by computers.With the help of speech recognition technology, it easy for people to control devices from phone to car and access applications by speaking. It can also be useful in recording the user's ID, name and reason of call. It also delivers a great experience of self-service system rate.

Our project is based on an Interactive Voice Response system for Railway ticket reservation and related queries.This system comes with 2 phases of development. Phase-1 includes speech to spectogram conversion phase 2 which is converting them into a text format and providing the response. In this project we are implementing the Phase-1 of the Interactive Voice Response System. The conversion of speech to text is a challenging task. Even though various technologies have been developed, the level of accuracy achieved is low. That's the reason, we need a better training model to achieve the same. This training model can be achieved through deep learning using the Recurrent neural network which is used for sequential data analysis. Training and testing by this model will help us attain a best accuracy.

1.1 Scope

Despite the complexity of Speech recognition, it always plays an inevitable role in many fields. It helps many people easily access to any contents they desire. Speech recognition is a thriving domain with many important applications. It's easy to predict that speech recognition research will continue as well as important practical applications will be created. Even though Speech recognition is major thriving field, getting the accuracy is the major issue in this field. Research and development is still in processing to get the most accurate machine possible. And it's not about AI because it's obvious that most of the speech recognition issues are not caused by the lack of understanding but rather a lack of good algorithms.

Noises, accents and so on are just purely technical problems which will be eventually solved. Research finds that noisy environment a major trouble in the speech recognition with a practical goal to build an application that works. At the same time our knowledge about speech fundamentally improves from day to day and the goals are improving more.

2. LITERATURE SURVEY

2.1 Phoneme Segmentation of Tamil Speech Signals Using Spectral Transition Measure

Geetha, K. and Vadivel, R. 'Phoneme Segmentation of Tamil Speech Signals Using Spectral Transition Measure ' the authors have termed that the process of identifying the end points of the acoustic units of the speech signal is called speech segmentation. Speech recognition systems can be designed using sub-word unit like phoneme. A Phoneme is the smallest unit of the language. It is context dependent and tedious to seek out the boundary. The methodology used by them are pre processing and Feature Extraction, Spectral Transition Measure (STM), Phoneme Boundary Detection. But some disadvantages were found to occur because of the low accuracy obtained and large phenome words could not be recognised.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

2.2 Speech Recognition Using Neural Networks

Dhanashri, D. and Dhonde, S.B. 'Speech Recognition Using Neural Networks, the authors briefed about the types of neural networks and their introduction. Also the hybrid design of HMM and NN is additionally studied. Deep neural networks square measure largely used for ASR systems. They give better performance as compare to traditional GMM. When employed in hybrid design with HMM, deep neural networks give better performance as compared to HMM-GMM system. The methods used are Feed forward neural networks: It is the one way connection without back loop is used. It has only connections forward in time, Perceptrons and multi-layer perceptrons: It is one of the type of feed forward neural network. A perceptron is a simple neuron model that consists of set of inputs, weights regarded each input and the activation functions, Recurrent Neural Network: In this sort of neural network output of vegetative cell is increased by a weight and fed back to the neuron including delay.

Recurrent neural networks (RNNs) square measure a strong model for sequential data, Hidden Markov model: In hybrid NN-HMM model each output unit of NN is trained to estimate the posterior probability of a continuous density HMMs state given the acoustic observations. Advantages include hybrid architecture of HMM and NN works well for the acoustic model of speech recognition.

2.3 LSTM-based Language Models for Spontaneous Speech Recognition

Medennikov, I. and Bulusheva,'LSTM-based Language Models for Spontaneous Speech Recognition' the authors stated that the language models (LMs) used in speech recognition to predict the next word (given the context) often rely on too short context, which leads to recognition errors. In theory, using recurrent neural networks (RNN) should solve this problem, but in practice the RNNs do not fully utilize the potential of the long context. The RNN-based language models with long short-term memory (LSTM) units take better advantage of the long context and demonstrate good results in terms of perplexity for many datasets.The method used is regularization with Dropout with an advantage that this technique takes into account a longer context to predict the next word as compared with the n-gram and even with the RNN-based models.

2.4 Speech Recognition using Deep Learning

Halageri A, Bidappa A, Arjun C, Sarathy M and sultana 'Speech Recognition using Deep Learning', the authors stated that speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format. Many speech recognition applications, such as voice dialing, simple data entry and speech-to-text are in existence today. The main purpose of the paper is to review the pattern matching abilities of neural networks on speech signal. The methods used are neural networks can be taught to map an input space to any kind of output space. They are simple and intuitive,hence they are commonly used.They are naturally discriminative.

2.5 Speech Recognition using Neural Networks

Tebelskis.J 'Speech Recognition using Neural Networks',the author has stated how examines how artificial neural networks can benefit a large vocabulary, speaker independent, continuous speech recognition system.

Currently, most speech recognition systems are based on hidden Markov models (HMMs), a statistical framework that supports both acoustic and temporal modeling. Neural networks avoid many assumptions, while they can also learn complex functions, generalize effectively, tolerate noise, and support parallelism. Dimensions taken are vocabulary size and confusability, speaker dependence vs. independence, isolated, discontinuous, or continuous speech and Read vs. spontaneous speech. The properties of neural networks are trainability, generalisation, nonlinearity, robustness, uniformity, parallelism.

2.6 Speaker based Language Independent Isolated Speech Recognition System

Therese S and Lingam C 'Speaker based Language Independent Isolated Speech Recognition System',the authors have stated that speech has evolved as a primary form of communication between humans. The advent of digital technology,gave us highly versatile digital processors with high speed, low cost and high power which enable researchers to transform the analog speech signals in to digital speech signals that can be scientifically studied. Achieving higher recognition accuracy, low word error rate and addressing the issues of sources of variability are the major considerations for developing an efficient Automatic Speech Recognition system. In speech recognition, feature extraction requires much attention because recognition performance depends heavily on this phase.

After the survey of these papers, our idea of using Long Short Term Memory and Connectionist Temporal Classification has been proposed which overcomes the disadvantages by other existing models. It includes the intention of remembering long sentences and having a prolonged memory for better prediction of content and using a shared decoder and encoder for mapping characters.

3. PROPOSED WORK

The way we proposed our system includes the Phoneme being extracted from the speech data using Mel Frequency Cepstral Coefficient MFCC which is then fed into the training model of Long Short Term Memory LSTM and Connectionist Temporal Classification CTC. This is followed by validation by calculating the CTC loss. The resultant obtained from this system is the resultant recognized text for the given speech input. And we are trying to provide the accuracy of the result.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

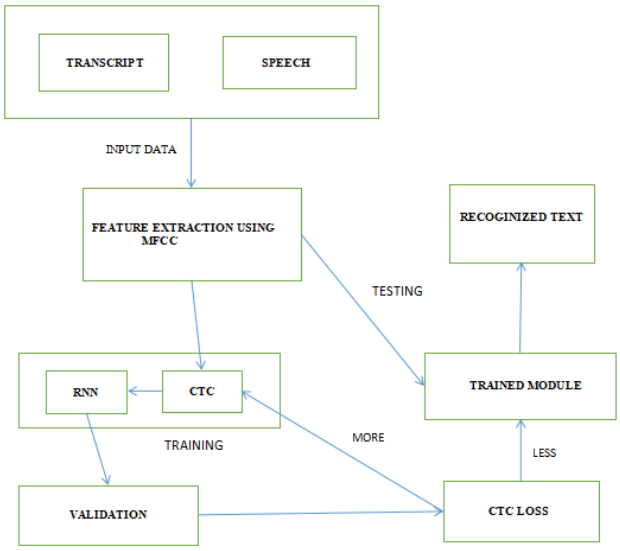

Fig no: 3.1 Architecture diagram of speech recognition

3.1 Speech Contents

The Speech corpus for our railway system has been collected from six different cities like Chennai, Mysore, Salem, Bangalore and KaniyaKumari. There are 500 data samples being stored and each sample being carefully obtained by containing the following formats.

For example- "Where does Mysore Express depart and arrive?" can be modulated in several ways such as:

- Where does Mysore Express depart?
- Where does Mysore express arrive?
- At which station does Mysore Express depart?
- At which station does Mysore Express arrive?
- At which point does Mysore Express depart?
- At which point does Mysore Express arrive?
- What is the departing station of Mysore express?
- What is the departing station of Mysore express?
- What is the arriving junction of Mysore express?

Similarly for a single type of input 9 or more basic sentence format has been recorded by using 3 different person inorder to give a complete accuracy for the model.

3.2 Mel Frequency Cepstral Coefficient

The first step of speech recognition is to convert a speech signal into streams of acoustic feature vectors, referred to as speech feature vectors. The speech data undergoes the process of feature extraction using Mel Frequency Cepstral Coefficient (MFCC). In this technique helps us to identify the linguistic content, discarding background noises. Initially the signal spectrums are pre-emphasized to spectrally flatten the signal in order to make it less susceptible to find precision effects during signal processing. Followed by framing where the speech signal is normally divided into small duration blocks, called frames. After framing, each frame is multiplied by window function prior to reduce the effect of discontinuity introduced by framing process by attenuating the values of the samples . Next spectral estimation is computed for each frame by applying FFT to produce spectral coefficients.

Spectral coefficients are computed for each frame by applying Fast Fourier Transform (FFT) to produce spectral coefficients. This process is of Mel filtering is being used. The natural logarithm approximates the relationship between the human's perception of the loudness and the sound intensity. The cepstral coefficients are obtained by applying the DCT on the log Mel filterbank coefficients. The higher order results represent the excitation information of the input, or the periodicity in the waveform, while the lower order cepstral coefficients results represent the vocal tract shape or smooth spectral shape.

3.3 Connectionist Temporal Classification

Connectionist temporal classification aims at satisfying the disadvantages that comes out of RNN. CTC (connectionist temporal classification) is a sequence-to-sequence classifier.

This maps an input sequence to a target sequence. In speech recognition, it predicts a sequence of labels (can be a phonemes or characters) from speech frames. CTC refers to the outputs and scoring, and is independent of the underlying neural network structure. The CTC algorithm can assign a probability for any Y given an X. The key to computing this probability is how CTC works between inputs and outputs alignment. Start by looking at these alignments and then by showing how to use those to compute the loss function and perform inference. Feature extraction, frame classification, sequence model, lexicon model, language model are the steps involved in CTC. Feature extraction outputs the values of the spectrogram which is used for further classification into frames to be fed into the RNN. From this model, sequence states are produced from the CTC which are then encoded and decoded to generate CTC loss. The encoder of CTC transforms the input sequence into high level features and the decoder generates the letter sequence. Here, CTC which targets an output at each input timestep and performs character mapping such that each character follows its subsequent character to form the word using probabilistic functions. CTC loss compares the decoded text with the transcript and directs the output to the trained model if the loss is found to be less and it redirects the output to the training module if the loss is found to be high.

3.4 Recurrent Neural Network and Long Short Term Memory

Recurrent Neural Networks are the best algorithm for sequential data. It process and gives the predictive results due to it's internal memory. At the same time LSTMs have an edge over conventional feed-forward neural networks. Hence using both together will always works in a good way.

LSTM layer acts like encoder-decoder, encoding the sequence of CNN features and emitting characters as outputs. The core concept of LSTMs are the cell state.

The cell state act as a transport highway that transfers relative information all the way down the sequence chain. It acts as the memory of the network. The cell state, in theory, can carry relevant information throughout the processing of the sequence. So even information from the earlier time steps can make its way to later time steps, reducing the effects of short-term memory. As the cell state goes on its journey, information gets added or removed to the cell state via gates.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

The gates are different neural networks that decide which information is allowed on the initial cell state. The gates will learn what information is relevant to keep or forget during training. The input to this neural network is the output of Mel Frequency Cepstral Coefficient which produces a spectrogram based on certain extracted features from the speech. Stacking of many RNNs sums up to a LSTM. The information flows through a mechanism known as cell states. The information at a particular cell state has three different dependencies. LSTMs can selectively remember or forget things. The input to the cell flows through the input squashing function a non-linear squashing function g(x)[Also known as 'S' or sigmoid function maps the range of z over [0,1]] and the result is then multiplied by the output of the input gating unit.

4. WORK SETUP

Major steps involved in our proposed work are below:

- The MFCC features extracted from the sample speech recordings are passed to the training module where the character mapping and file mapping is done using CTC.
- The output from the previous are fed into the cells of RNN and the stacking up of the cells of it inorder to forms the LSTM module.
- CTC loss is being calculated from the encoding and decoding technique
- From the result of the CTC loss, the value is checked if it is less, it is sent to the trained module where testing is done and the output test is given out or it is more, then it is sent back to the training module until the loss meet out the requirement
- Performance and accuracy of the system is observed by using different metrics namely Word Error Rate(WER), Label Error Rate(LER) and Accuracy.

4.1 MFCC Extraction

- Using a tool named Librosa in python for speech data processing.
- The features are extracted from the speech files and stored in pickle files to save storage space which is generated in a separate folder to load the speech data files from.
- The number of features are 13 and the window size when calculating MFCC is 10-30 msec for 10000 Hz.
- The path for storage of these speech data files are set in a json file along with the sampling rate.
- The features extracted are dumped as pickle files in the given path location specified in the json file.
- The text file from the transcript is mapped to its corresponding to the wave file to help in training and testing.

4.2 CTC Format conversion

- The mfcc features extracted are transformed into 3D array and the length of the sequence is found.
- The text from the transcript are read and preprocessing of the text are made inorder to match the conversion text.
- The characters are mapped to a particular index from 0 for space and 1 which corresponds to alphabet the 'a' till all 26 alphabets.
- A sparse representation is created to feed the placeholder where the background noises and silent frequencies are being truncated.
- The filtered wave files are stored in the placeholders to achieve the following.

4.3 Testing and Training

- The session is being declared for the graph and the global variables are initialised.
- For every epoch, the input, target and sequence length are being fed into the model
- The batch cost, train cost and label error rate are calculated.
- The feed inputs are being decoded followed by which character index mapping is done.
- The trained model is being saved using a saver function.
- The trained model is restored and test inputs are given and the accuracy is also computed.

5. CONCLUSION AND FUTURE WORK

In this paper, we have discussed briefly about converting speech to text. By referencing many papers, the method phenome segmentation preprocessing and feature extraction is done but the accuracy of the output gained is found to be lower than expected. So, we have incorporated an hybrid methodology which includes Long Short Term Memory(LSTM) and Connectionist Temporal Classification(CTC) inorder to produce high accuracy. We have used Mel Frequency Coefficient Cepstral(MFCC) and Librosa to extract features from the speech signals. The architecture of our proposed system includes the features being extracted from the speech data which it then fed into the training model of LSTM and CTC. This is followed by validation by calculating the CTC loss. The CTC loss is also calculated to direct the input either to be trained or to be validated repeatedly. The Label Error Rate(LER) and Word Error Rate(WER) are also computed to calculate the accuracy of the model. The trained model is being tested with input speech.

Our future work concentrate on increasing the accuracy of the model and also to complete the system as a whole for the railway system we took dataset for. Our model works only for trained users and the future work is to extent for n users and develop an interactive module. Along with this, we like to provide the system with authenticating the user by using their voice ID.

REFERNCES

[1] K. Geetha and Dr.R. Vadivel, Phoneme Segmentation of Tamil Speech Signals Using Spectral Transition Measure , Oriental journal of Computer Science and technology, Vol. 10, March 2017, pp. 114-119, ISSN:0974-6471

[2] A. Graves, A. Mohamed and G. Hinton, Speech Recognition with Deep Recurrent Neural Network, International Conference on Acoustics, Speech, and Signal Processing, May 2013, ISSN: 2379-190X

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

[3] D. Dhanashri, S.B. Dhonde, Speech Recognition Using Neural Networks: A Review, International Journal of Multidisciplinary Research and Development, Volume: 2, Issue: 6, June 2015, pp. 226-229, ISSN:2349-4182

[4] C. Ittichaichareon, S. Suksri and T. Yingthawornsuk , Speech Recognition using MFCC, International Conference on Computer Graphics, Simulation and Modeling, July 28- 29, 2012

[5] X. Liu, Deep Convolutional and LSTM Neural Networks for Acoustic Modelling in Automatic Speech Recognition, Pearson Education Inc. Vol. 8 (6), 2011

[6] S. Kim, T. Hori, S. Watanabe, Joint CTC-Attention based end-to-end speech recognition using multi-task learning, 31 Jan 2017, arXiv:1609.06773v2

[7] A. Halageri, A. Bidappa, C. Arjun, M. Sarathy, S.Sultana , Speech Recognition using Deep Learning, International Journal of Computer Science and Information Technologies, Vol. 6 (3), 2015, pp. 3206-3209,ISSN: 0975-9646

[8] I. Majeed, H. Husain, S. Abdul Samad, T.F. Idbeaa, Mel Frequency Cepstral Coefficients(MFCC) Feature Extraction Enhancement in theApplication of Speech Recognition:A Comparison study, Journal of Theoretical and Applied Information Technology, Vol.79. No.1, September 2015, pp. 38-56, ISSN: 1992-8645

[9] K. Lekshmi, Dr.E. Sherly, Automatic Speech Recognition using different Neural Network Architectures A Survey, International Journal of Computer Science and Information Technologies, Vol. 7 (6), 2016, pp.2422-2427, ISSN: 0975- 9646

[10] I. Medennikov and A. Bulusheva, LSTM-based Language Models for Spontaneous Speech Recognition, International Conference on Speech and Computer, Vol.9811, SPECOM 2016, pp. 469-475, ISBN: 978-3-319-43958-7