SPEECH RECOGNITOIN USING HIDDEN MARKOV MODEL AND ARTIFICIAL NEURAL NETWORKS

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

RAGHAV

11614243

Supervisor

SWEETY SEHGAL

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

November 2017

November 2017

SPEECH RECOGNITION USING HIDDEN MARKOV MODEL AND ARTIFCIAL NEURAL NETWORKS

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

RAGHAV

11614243

Supervisor

SWEETY SEHGAL

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

November 2017

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0350

Supervisor Name : Sweety Sehgal **UID :** 11115 **Designation :** Assistant Professor

Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Raghav	11614243	2016	K1637	9023542800

SPECIALIZATION AREA : Programming-I **Supervisor Signature:** _____

PROPOSED TOPIC : Speech Recognition using Hidden Markov Model with Artificial Neural Networks

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	Parameter	Rating (out of 10)
1	Project Novelty: Potential of the project to create new knowledge	6.40
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.20
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.00
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.40
5	Social Applicability: Project work intends to solve a practical problem.	7.00
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	6.80

PAC Committee Members		
PAC Member 1 Name: Kewal Krishan	UID: 11179	Recommended (Y/N): Yes
PAC Member 2 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 3 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 4 Name: Dr. Pooja Gupta	UID: 19580	Recommended (Y/N): Yes
PAC Member 5 Name: Kamlesh Lakhwani	UID: 20980	Recommended (Y/N): NO
PAC Member 6 Name: Dr.Priyanka Chawla	UID: 22046	Recommended (Y/N): Yes
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): Yes

Final Topic Approved by PAC: Speech Recognition using Hidden Markov Model with Artificial Neural Networks

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11024::Amandeep Nagpal **Approval Date:** 10 Nov 2017

12/5/2017 12:26:03 PM

Abstract

Speech Recognition is the smart and easy way to interact with machines. Physically challenged people can also use the technology with the help of speech recognition systems. It gives freedom to the users to interact with any kind of systems. Speech recognition is in research from many past years, many works and models were created by the different researchers. In this research, the proposed technique to create speech recognition model is Neural Networks. Neural Nets have the capacity and power to learn the behaviours and decision making capabilities. Although many types of neural nets are available now, we can modify them or can propose our own model to create a best speech recognizer. Hidden Markov Models will be used for matching the patterns of sounds. HMM has the capability to compare the different special length of data using the probabilistic methods. With the help of hybrid model created using HMM and ANN, we try to built accurate speech recognizer.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation proposal entitled "SPEECH RECOGNITOIN USING HIDDEN MARKOV MODEL AND ARTIFICIAL NEURAL NETWORKS" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mrs.Sweety Sehgal. I have not submitted this work elsewhere for any degree.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

.....

Raghav

R.No: 11614243

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech dissertation proposal entitled "**SPEECH RECOGNITOIN SYSTEM USING HIDDEN MARKOV MODEL AND ARTIFICIAL NEURAL NETWORKS"**, submitted by **Raghav** at **Lovely Professional University, Phagwara, India** is a bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

.....

Sweety Sehgal

Date: 01 Dec 2017

Counter Signed by:

1) Concerned HOD:

H.O.D.'s Signature: _____

H.O.D. Name: _____

Date: _____

2) Neutral Examiners:

External Examiner

Signature: _____

Name: _____

Affiliation: _____

Date: _____

Internal Examiner

Signature: _____

Name: _____

Date: _____

ACKNOWLEDGEMENT

I would like to thanks Mrs. Sweety Sehgal for her timely help and guidance for this dissertation. I would also like to thanks to my classmates and friends for their continuous support for my work. Most importantly I would like to thank my family for their support and motivation for this whole work.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER – 1

INTRODUCTION

1.1 Speech Recognition Systems: An Introduction

We interact with computers and machines using many kind of input device such as keyboard, mouse, joystick, camera etc. Whereas interacting with such systems using voice is unique and much easier concept now those days, because we know communication through voice/speech is much easier among human beings. To interact with computers through voice we need a microphone and a good written program to translate speech frequencies into text. But interacting with machines using any one of natural languages as human spoke is very difficult task. To enable computer systems to understand the natural language is very difficult task for developers. Because voice can be handled using computers, but understanding the meaning of that speech, is not an easy task.

Fig 1.1 Speech Recognition System

Speech recognition is the process to convert speech into text. It is considered as a sub-field of computer linguistics, which enables the computers to understand the speech of natural language. It is the collective knowledge and research of linguistic, computer science and electrical field. The tech giant companies like Google, Microsoft, Apple, IBM, Amazon, Baidu etc. are working in this field from last many decades. There are basically two type of speech recognizing systems that are in practice now days:

- Speaker dependent speech recognition system
- Speaker independent speech recognition systems

1.1.1 Speaker Dependent Speech Recognition System

Speaker dependent system are that systems that were used for the dictation purpose, in which an individual speaker read and speak the given content for system's learning purpose. Such kinds of system are very accurate to a single speaker. These systems are more reliable than speaker independent but the limitation of such systems is that they are able to translate the single person's speech only. Speaker dependent systems are learned to translate speech of single speaker only, his way of speaking, rate of speech, and other factors of the voice are considered by the speaker dependent speech recognition systems. These kinds of systems are best for single user speech recognition system only. It will give extreme accuracy to translate single speaker's speech into text. To use such kind of system each user has to train his own system only for his usage purpose.

1.1.2 Speaker independent speech recognition systems

Speaker independent systems are independent of speaker's individuality. These systems are basically available now days in Smartphone. Such systems can understand and translate speech from different speaker as well as. This system needs a lot of training to understand natural language and spoken words with different styles, accents and different rate of speech. Rate of speech is the speaking speed of speaker; some people speak very fast whereas some speak very slowly. That will not be fare if speech recognizer can understand the speech that is spoken as a manner of news reporter or some kind of read speech, but can't work with the normal speakers. So, this is a very difficult task. Another difference came in speech as of age of the speaker, because children speech is pretty much different than of an adult and from an old age person. So, try to recognize speech by the different age group is also a difficult task for speaker independent speech recognition systems. By the style, I mean to say the speaking style of a particular person. Each person speaks in his different style or accent, Nonnative English speakers can't speak in the same style as the native speakers. So, it is also a difficult task in building these kinds of speech recognition systems. To overcome such situations, adequate learning examples are required, so our system can work efficiently with all the users to recognize their speech and convert it into text.

1.2 Speech Recognition System Difficulties

- **Style:** The style of the speech a major difficulty in SR systems. Casual speech differs than of read speech. Casual speech is the speech that is used by the humans normally to interact with

one another, whereas read speech means reading and giving speech for example politicians address the people with some written speech. Another style of speech is isolated words and continuous words. Continuous speech is speech in which words are over fitted with each other and can make confusion to the computer system, whereas isolated speech meaning words are spoken properly with pauses in between every two words.

- **Environment:** By the environment I mean to say the place where the user is present and using the SR system. The main conditions including background noise, channel condition, room acoustics etc. Channel means the medium from our system is getting voice/sound i.e. microphone. Background noise is any kind of disturbance in background that is making the sound robust for the system i.e. traffic noise, other person voices etc.

- **Speaker characteristics:** Speaker characteristics includes rate of speech, accent etc. Some speakers speak very faster, some people annunciate more. Some people speaks in different accent than others. So, these constraints are also difficult to tackle. Also, the age of the user is considered as a difficulty, because different age groups people have different voice, a child voice is pretty much different than of an adult and an old age person.

- **Task specifics:** The number of words in the vocabulary is also a constraint to discuss. If we talk about medium size or a small sized SR system then the data needed for these kind of system is also limited, that will not so difficult. But if we talk about the larger system i.e. full English language with all the grammar rules, then it will be a time consuming and resource rich task, that is pretty much difficult.

1.3 History of Speech Recognition

- In 1952 at Bell Labs, three researchers developed a single speaker speech recognition system, which can perform digit recognition only. The total length of vocabulary of that system was only ten.
- In 1960’s Raj Reddy became the first person to take on continuous speech recognition as graduation at Stanford University, Earlier the system was made to recognize either digits or single words, but Raj Reddy's system can recognize the continuous speech that was actually a commanding system to the chess.
- In 1960’s Soviet researchers also invented Dynamic Time Warping (DTW) algorithm that was

capable of translating 200-word vocabulary. DTW uses the concept of segmentation and processing each segmented frame.

- In 1971 Defense Advanced Research Project Agency (DARPA) gave funds to encourage the research in this field, DARPA's main goal was to achieve to recognize up to 1000 words vocabulary. BBN, IBM, Carnegie Mellon and Stanford Research Institute were the participants in this competition. Out of them CMU's HARPY system won the competition.
- At Institute of Defense Analysis, Leonard Baum created Markov Chains in 1960s. Hidden Markov Model (HMM) used by the Raj Reddy's students James Baker and Janet N. Baker for speech recognition. James Baker studied HMM during his studies.
- At IBM Fred Jelinek invented Tangora, that was operated through voice. In 1980 Tangora was able to handle up to 20,000 words of vocabulary. Jelinek also used statistical approach same as HMM.
- HMM were considered as the highly useful model to model speech. HMM replaced DTW. In 1980 n-gram language model was introduced. In 1987 Katz used n-grams of multiple length to use language model.
- Actual speed in speech recognition arrives when the computing power increased. Back on days in 1976, when DARPA ended its research competition, the best computer available at that time was included 4MB of RAM only. Using such computers, it is very time-consuming process, because it takes 100 of minutes to decode 25 seconds of speech. But after a few decades, computing power increased tens of thousands of the earlier computers. Then the researchers took advantage and started building speech recognition system with much more vocabulary, speaker independent systems etc.
- In 1987, first commercial recognizer was introduced by Kurzweil Applied Intelligence. In 1990 Dragon Dictate was introduced.
- In 1992 AT&T used a voice Recognition Call Processing system, which can reroute calls by itself without human efforts. That was developed by the Lawrence Rabiner at Bell Labs.
- Raj Reddy's another student Xuedong Huang, developed Sphinx system at CMU. In 1992 this was considered as the best system by DARPA. After that Xuedong Huang in 1993 went to Microsoft and founded speech recognition group of Microsoft.
- Raj Reddy's another student Kai-Fu Leen in 1992, went to APPLE, and developed CASPER that was a speech interface prototype.
- In 2000's speech recognition is still using HMM but in a hybrid manner along with feedforward

Artificial Neural Networks. Today Deep Neural Networks takes all the speech recognition process, known as Long short-term memory (LSTM).

- In 2007 Connectionist Temporal Classification used with LSTM used to get better results than earlier techniques in some applications. Google used CTC-trained LSTM and got a very dramatic outcome, that is now used by Google Search in all of its smart phones.

1.4 Structure of a Typical Speech Recognition System

Speech recognition system is combination of different models, which performs different task and gives outcome as the text to the corresponding speech. It includes Acoustic Analysis, Acoustic Model, Pronunciation Model and Language Model. The all the models work together and performs speech recognition using some learnt mechanism. Each model plays its role, and give output as input to the next model, based upon the input from the previous model, each model computes its output and pass over the system. The decoder is the main part of the speech recognition model, which

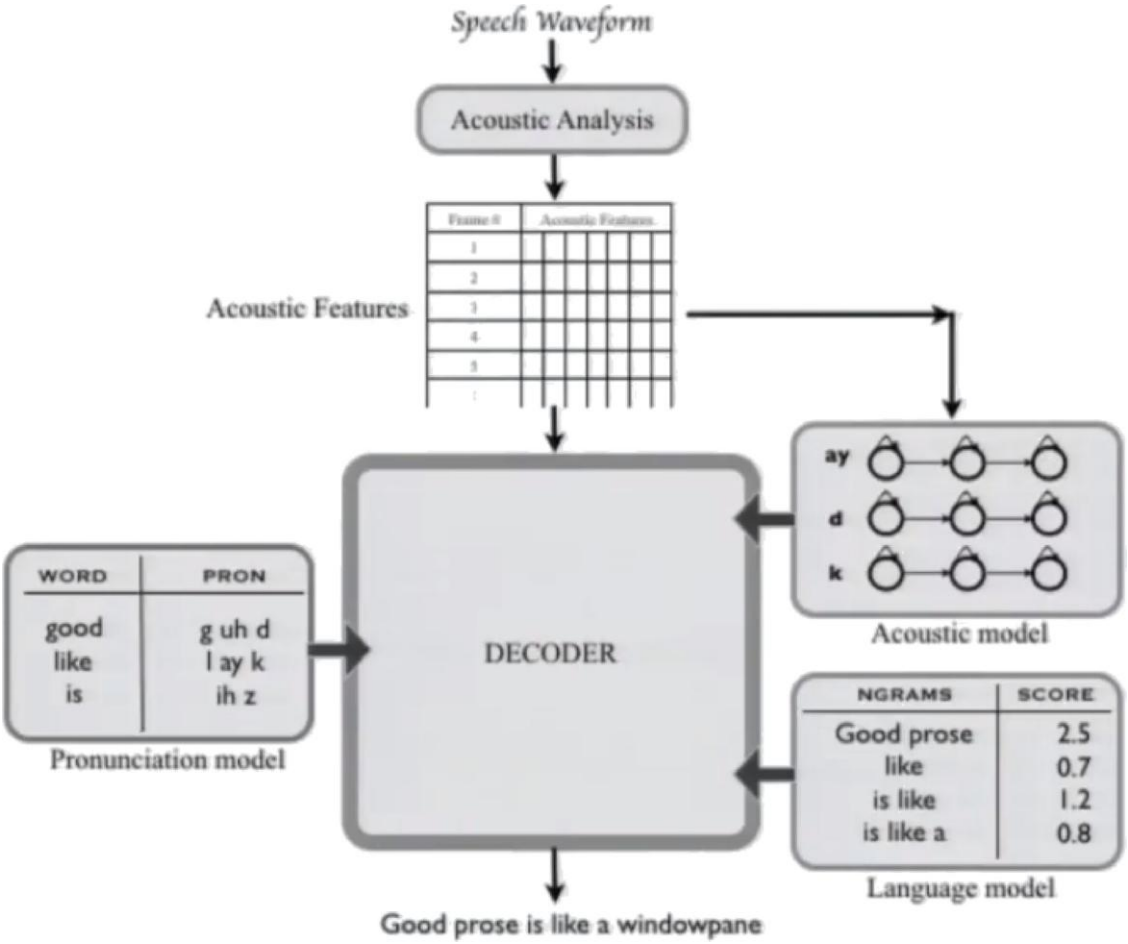

Fig 1.2 Structure of Typical Speech Recognition System

1.4.1 Acoustic Analysis

In this phase we have to analyze the raw speech signal and then discretized, because we can't work with raw data. Then we have to make samples of this raw data. Each sample is of 10-15 milliseconds of speech. Out of which each frame is extracted as a feature, feature extraction is very involved process, if the data is not stationary, then we can't get features of that data. This is the model that uses the same fundamental as of our ears uses. It also known as Mel frequency Cepstral Coefficients (MFCC)

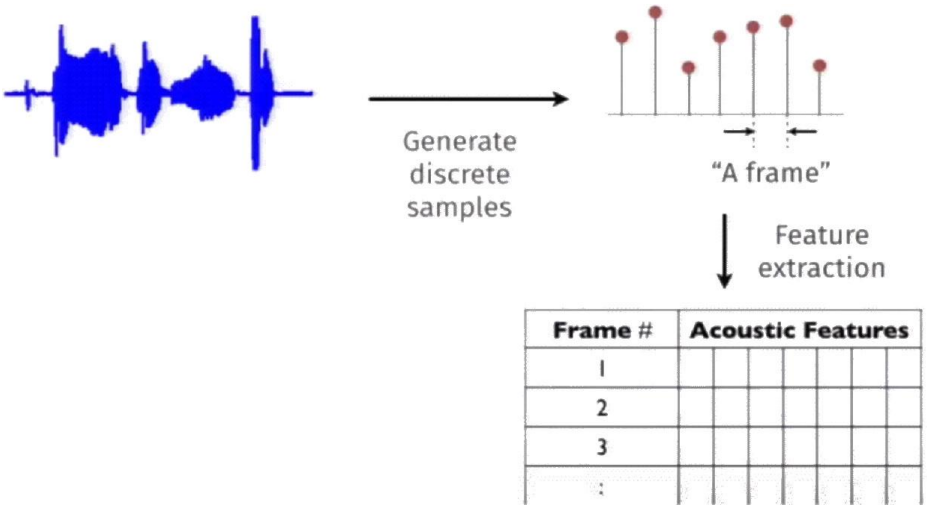

Fig 1.3 Process of Acoustic Analysis in Speech Recognition System

1.4.2 Acoustic model

The basic unit of acoustic information is phoneme. A phoneme is basically a discrete and distinctive unit of language that can be used to differentiate different words. Most of the natural language is having 20-60 phonemes.

Fig 1.4 Acoustic Model of Speech Recognition System (String on the Beads Method)

The pronunciation of the word is pretty much different than of original written word. These phonemes were written by the linguistic experts of the language. The most popular dictionary that is used by the researchers all over the world that is CMUDict, that is a free open source dictionary offered by the CMU. It is very toughest work to create our own acoustic model, because these are created by the linguistic experts.

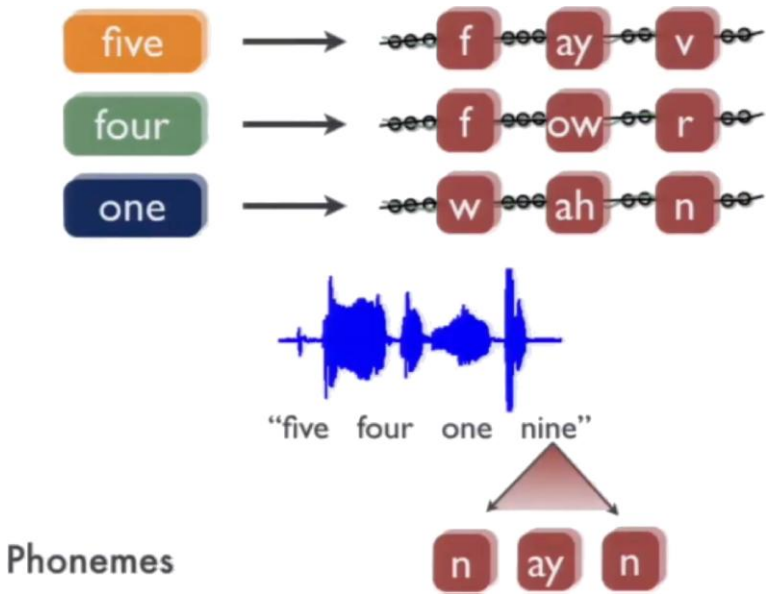

Fig 1.5 Acoustic Model Prediction of Unknown Occurrence

1.4.2.1 Mapping the Acoustic Feature to Phoneme

Mapping the acoustic feature got from the first section with the likely phonemes to know the particular speech occurrence. For this purpose, Hidden Markov Model is the paradigm that is used to learn this mapping.

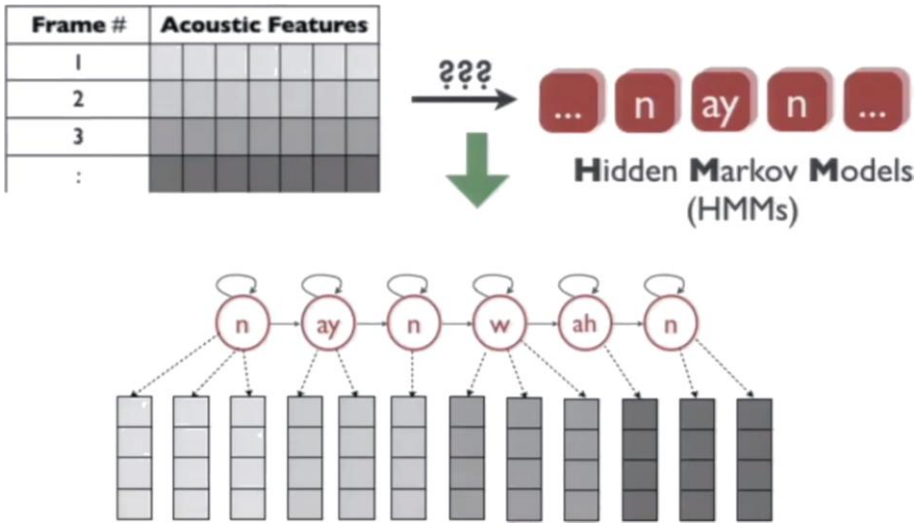

Fig 1.6 Mapping the Acoustic Feature with data using HMM

Now those days Deep Neural Networks are used for the similar mapping than HMM. That also uses probability distribution. In earlier technique using HMM the probability distribution calculated through Gaussian Mixture Model (GMM), but in case of DNN, it itself used to generate probability distribution.

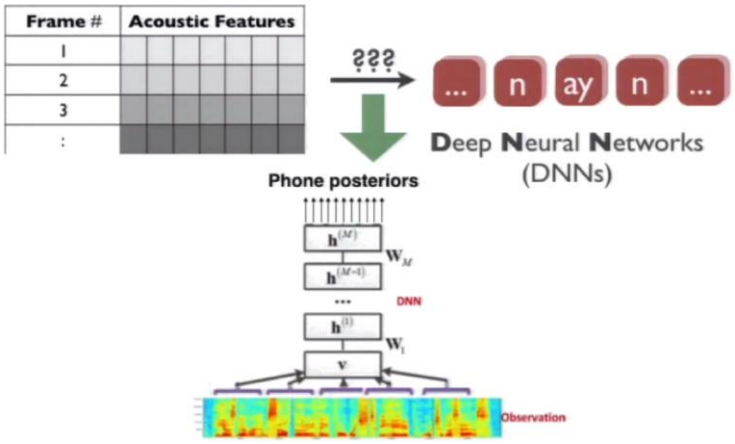

Fig 1.7 Mapping the Acoustic Feature with data using DNN

1.4.3 Pronunciation Model

Pronunciation model provides the link between sub-word units (phonemes) and the actual word. So typically, a large dictionary of pronunciation is maintained in this model. This is the only module in speech recognition system that is not learnt. Pronunciation model is actually expert derived data; expert gives us the mapping between these. In this data all the pronunciation variations, phonetic transcriptions are present to map the word with the library data.

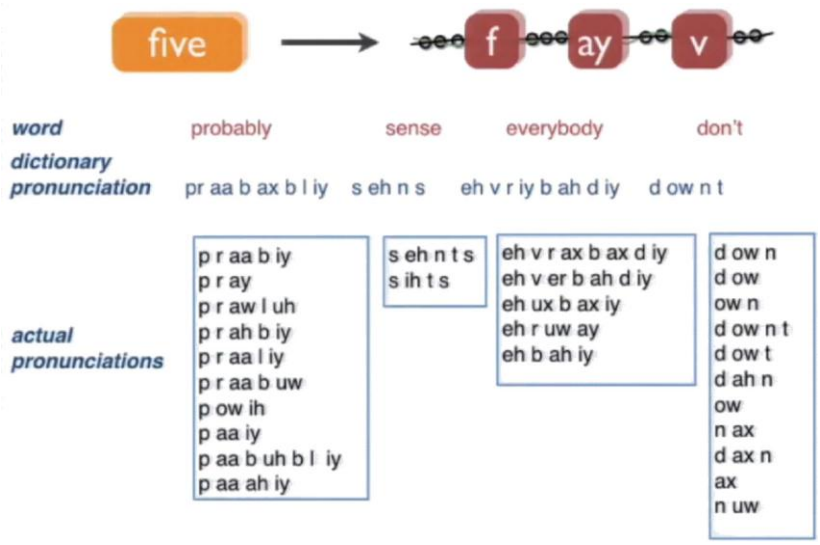

Fig 1.8 Pronunciation Model of Speech Recognition System

1.4.4 Language Model

This model predicts the correct order of the words that is derived from pronunciation model. So, it is a learnt model that used lots of text and grammar rules to train. It finds the occurrences of the words to-

gether. It uses probabilistic approach to predict the order of the speech.

- For example, “the dog” ran?
  can?
  pan?

The language models also differentiate the similar acoustics. For example, the utterance is

- “Is the baby crying” vs. “Is the bay bee crying”.

Language model only can be used in n number of applications. For example, Speech Recognition, Machine Translation, Handwriting Recognition, Optical Character Recognition, Spelling Correction of sentences etc. We can use n-gram language model to compute the correct order of the given words. For this we can increase the order of the n-grams. We can use bi-gram, tri-gram or even n-gram with higher number also, if we are running into a large data of language. It is advised to use a correct number of n-gram for better outcomes.

1.4.5 Decoder

The main component of that structure is the decoder, which is very important. Here we have to face searching problem. We have to found the very likely phoneme sequence and the word sequence, and then finally we can get the most likely word sequence corresponding to speech utterance. To do this we have to put all the later section together and look from the entire space. There could be a very large search graph, from where we have to found most likely phonemes and its values. So, this is how a search graph looks like:

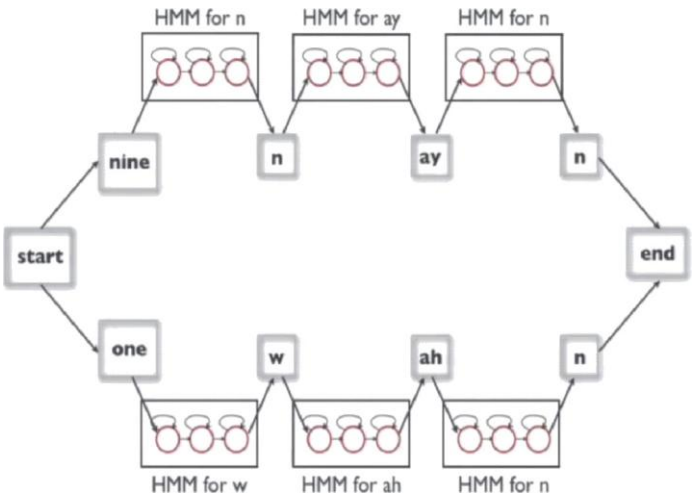

Fig 1.9 Decoder’s Work

Here starting from the start, and then we assume we can get only two words one or nine. Here in graph each arc is having some weight. Here each word is having its own related phoneme; each phoneme corresponding to it's HMM. As we can see this is a pretty large graph that is built for only two words, as of if we want to do same search with 20-40 thousand of words, which will be really complex one.

1.5 Structure of Speech Recognition system using Neural Nets

There is new direction that is in practice now days. In the previous section we saw the typical speech recognition system, where different models are used to accomplish its task. But with the help of neural networks we can skip to those models. Using neural nets system can learn the directly mapping from the acoustic features to the characters. This system's main advantage is that we don't need to map data with pronunciation model and we don't have need to look over phonemes. But to train such systems we need a lot of speech and corresponding text. And we need lots of data to get quite a good outcome from the system.

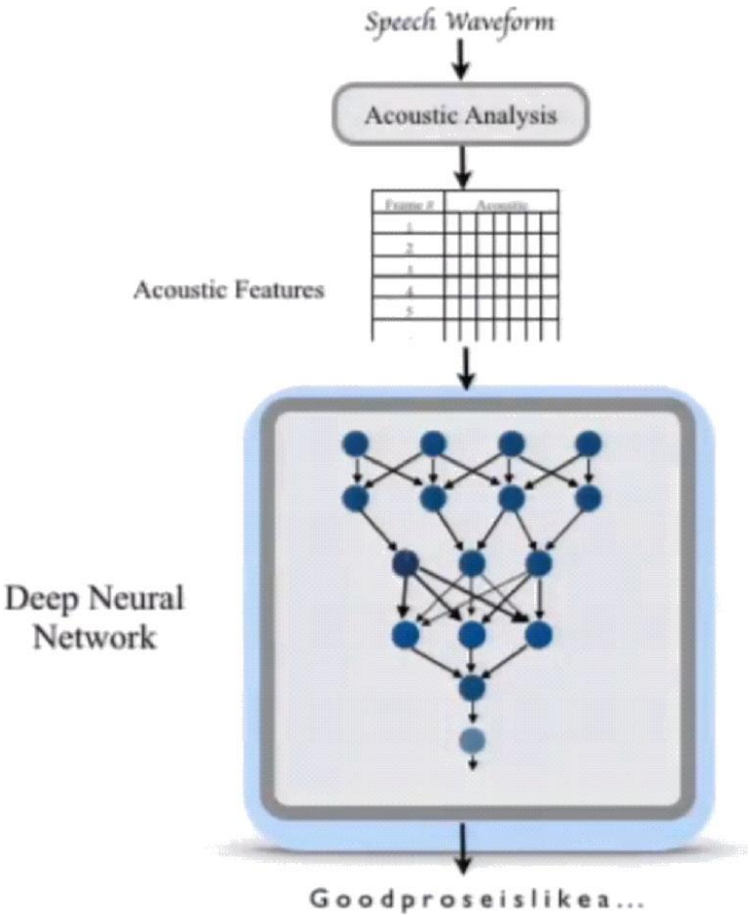

Fig 1.10 Structure of Speech Recognition System Based on Deep Neural Networks

1.6 Applications of Speech Recognition

- **Dictation:** This is one application of the speech recognition system. People related to law, medical and business are using the dictation which uses speech recognition. Some dictionaries are also created for special purpose or user based, such kind of dictionary system included with special vocabulary related to the user's purpose.

- **Command based Control systems:** Speech recognition systems are created for commanding purposes also, for example commanding your mobile to 'call home'. Such kind of systems are created using speech recognition.

- **Telephony:** Speech recognition also used for the automatic call answer systems, that give freedom to their user to give instruction through voice, instead of pressing the keys.

- **For Disabled Peoples:** Some people are unable to write, because of some natural phenomena, speech recognition could be used as the input to gadgets to be operated by such people who are unable to type.

- **Embedded:** Speech recognition can be used for the embedded system to operate them with the voice, as now we operate our mobile through voice, but what if we interact with washing machine, other home appliances using our voice, for example, 'switch on TV', 'start washing machine' etc.

CHAPTER – 2

REVIEW OF LITRATURE

For the clarification of the topic and the different purposed work and solutions given by the researchers all over the world in the field of speech recognition, it is important to read literature available to me. I read many research papers to be more specific for the speech recognition and for my work. The study of such literature gives me idea of my work plan during my dissertation.

2.1 Ashwin Bellur et al (2017) [1], it is very hard to compare two different datasets with one having some noise, but similar to the first one and predict if they belong to each other or not. In this condition the system will classify one dataset as different one compared to the other one. The data during training and the actual data for recognition, can be tiny different because of the environmental factors. The author provides a solution to this problem by using 2-d Gabor filter bank [1], that will help to distinguish between the speech and non-speech sound. Then he also purposed an advanced technique to use the Genetic algorithm to enhance the outcome of the Gabor filter, that were used to retune the Gabor filters. This technique gave him a better result than the conventional system up to some value [1]. Inspired by the biological hearing system of human, author purposed this technique to overcome the error that arrives because of the environmental noises.

2.2 Karen Ullrich et al (2014) [2], Boundary recognition in the voice data is very important task in voice structure analysis. Author want to automatically detect the boundaries of the signal, so it can predict the pauses occurred between each word in the music. To create such system, he used Convolutional Neural Networks. These convolutional neural nets were trained on the Mel-scaled magnitude spectrogram directly. They used F-measure [2], which denotes the ratio of the speech signal. By this technique the author will able to detect the boundary of the sound pattern. Same as this technique if we use this technique to speech recognition, I think the accuracy of the speech recognition can also be increased. Because the speech also contains continuous connected words in between the speech, because some people speaks very fast, so in such cases the F-measure can be changed to according to the speaker. We can train our system to automatically generate a well-suited F-measure for the user so that the speech recognition can be done easily without less amount of error. This seems a good practice to use neural nets to learn the human speaking behavior. Neural nets are capable of learning new things by itself. When we talk about neural nets in speech recognition, it is very clear, we are going to skip some models, using in the previous technique. Neural nets itself is capable of learning the speech to text process,

but adequate data for training is needed. Training data is part to think here, and it takes lots of time to train data with a bigger dataset.

2.3 Ossama Abdel-Hamid et al (2014) [3], purposed a CNN based solution to achieve more accuracy in speech recognition than the earlier hybrid Deep Neural Networks (DNN)- Hidden Markov Method (HMM) [3]. The DNN-HMM model also proves itself better than earlier Gaussian Mixture Model (GMM) - HMM model. Convolution Neural Nets are the modified type of neural networks, in such kind of neural networks the hidden layers are changed with pooling ply and convolutional ply [3]. The weight learning process is dome after the pooling ply and convolution ply. In the CNN the authors purposed further modification, they used limited-weight-sharing technique for enhanced modeling of the system. Using limited-weight-sharing speech patterns are handled in a better way by the CNN, and smaller units in pooling ply, that helped the system to use less computational power with less complexity. Using CNN, the error rate reduced by 6%-10% compared to earlier purposed solution. So, CNN is also a good technique to be used in the speech recognition. If Hidden Markov Model is combined with the CNN, then the accuracy of the system can be increased further.

2.4 Avery Li-Chun Wang (2010) [4], started building a new system in 2000, named as Shazam. The main idea behind the Shazam is to connect people with music using any device. This algorithm is able to recognize the real-time broadcasting of music from environment and match it with the stored database and get the name of the song directly within few minutes. There was no earlier work of this kind or not any algorithm was purposed by any researcher. So, Li-Chun Wang started working on this by himself. He created a system and now it is capable of doing song recognition out of 2 million songs dataset. It used 10-15 seconds of sound from the environment and it can predict the actual song name with its title in less than 15 milliseconds [4]. The solution purposed by author is very much similar with the fingerprinting. The purposed a solution to create a fingerprint hash value of the songs in their database and after that whenever a user want to use it, he gives the system a 10-15 second of sound and the system suddenly computes the hash value and match that hash value with the stored hash value table. If match found then the output is sent back to the user's device i.e. mobile phone. The accuracy the system is too high that it can recognize the song with the help of radio quality sample input in less than 15 milliseconds. Before producing hash values, this system used the robust cancellation technique to handle noise in the sample. After the noise removal the fingerprint is created and then only matched with the dataset's fingerprints. From this literature, I got an idea if one wants to recognize dataset available dataset with sample data, hash values is the best medium to get accuracy and speed. But limitation here

is that the style of speech should be same as the training data, and corresponding song's hash should be available in the dataset, so data in this system plays a big role. But for pattern matching, this is best technique.

2.5 Dharmendra P Kanejiya [5], proposed a framework to perform speech recognition using Hidden Markov Model and Artificial Neural Network. He stated that HMM are very good to handle the data with temporal variation but are not able to do classification. Temporal variation means sometimes people spoke a word with elastic effect. Such kind of speech also can be handled through HMM, but as said HMM can't do classification. Whereas neural nets are capable of generation posterior probability, neural nets don't make assumptions, it works on stats. Neural nets are unable to work with variant speech. But the combination hybrid approach of HMM and ANN can give us more than we accepted. Author used Multilayer Perceptron to learn the learning rate of the system for better classification. From his work I got idea of hybrid models such as HMM and ANN. Both are considered as the masters in the different fields, so to skip Acoustic Model, Pronunciation Model and language model we can use HMM with the ANN where we don't have need of some of the computation, but only learning is required.

2.6 Mohamad Adnan Al-Alaoui et al (2008) [6], the authors purposed a technique to classify the isolated word speech in Arabic Language. They proposed Cepstral Feature extraction for the extraction of features from acoustic signals. Then they used HMM for finding patterns from the speech. And for classification they used artificial neural networks (ANN), that classify the words with the help of learning data. The authors were made a system that can help the new language learner can learn and read the Arabic language quickly. They used K-Nearest Neighbour Classifier model of Neural Nets to classify the patterns. The languages which have more common speaking styles of different words, are classified only using some best classification technique. Here in this case the authors used KNN for classification.

2.7 Dhavale Dhanashri et al (2017) [7], they used a Deep Neural Network for speech recognition for isolated speech recognition. Whereas the earlier model was Hidden Markov Model (HMM) - Gaussian Mixture Model (GMM). The neural network takes the place of GMM, because GMM are used to calculate probability on the basis of assumptions only. Deep neural networks are basically feed forward networks with more than one hidden layer in between them. Author also used Deep Belief Network to pretrain the DNN. DNN's initial values will come from DBN. So, it is the combinatorial approach of HMM and DNN for speech recognition, the experiments gave them good results, but the accuracy can

be increased using more hidden layers in between the input layer and output layer. Deep neural network is the next generation machine learning tool that is capable of learning through the iteration. These nets are very much accurate in speech recognition also. Today's tech giant like Amazon, Google Search, Microsoft and Apple, all are working on the Deep neural networks only.

2.8 Bhushan C. Kamble (2016) [8], explains very briefly about the speech recognition and SR using neural networks. The author discussed each and every step, which are very simple. Then the main part of the structure came, that is speech classification. This paper was about all the neural nets. So, author tried to explain each of them. Artificial neural networks are set of neurons and designed to work as the human brain works. Human brain is much more complex in decision making than of logical approach to the computer programs. So neural networks were designed to do so, or let computers think like humans. Then author explains the 4 most used approached of neural networks; Feedforward network, Recurrent Neural Networks, Modular Neural Networks and Kohonen Self Organizing Maps (KSOM). From the study of literature, it found that the RNN are better than Multi-Layer Perceptron, the only complexity in neural nets is training. Artificial Neural Networks are the coming future of the computing.

2.9 Kapure Vijay Ramesh (2013) [9], purposed hybrid HMM-ANN approach to solve long time running problem of speech recognition. HMM are good at patter matching, whereas ANN are good at classification or learning. So, author purposed a hybrid approach by combination of both the techniques. He purposed that the feature extraction and mapping is done with the help of ANN such as Back Propagation Algorithm (BPA). He introduced Multilayer Pattern Mapping Neural Network. Those MPMNN were worked as of BPA. At the end speaker recognition is done through HMM. In this paper author tries to recognize speech as well as speaker recognition. The scope of this paper is for multiple features included speech recognition, speaker recognition and feature extraction using ANN.

2.10 Supriya S. Surwade et al (2012) [10], explains the ANN and HMM in term of speech recognition as in the speaker identification or biological identification based on the speech, they worked with HMM, but it lacks over the real-world environment. So, they opted to ANN, nut these systems also fails to map long sequences under some circumstances. So, they used a hybrid approach with the combination of HMM and ANN. MATLAB 2012 is used as tool in this research. The hybrid model shows the results as the accuracy if recognizing isolated words is increased with this model.

2.11 Alex Graves at all (2014) [11], purposed a model that can do directly conversion of speech into

text without the help of any model. The system is created using combination of deep bi-directional LSTM Recurrent neural networks and Connectionist Temporal classification objective function. It achieves a hall of fame 27.3%-word error rate without any dictionary or any language model. The example of this system is

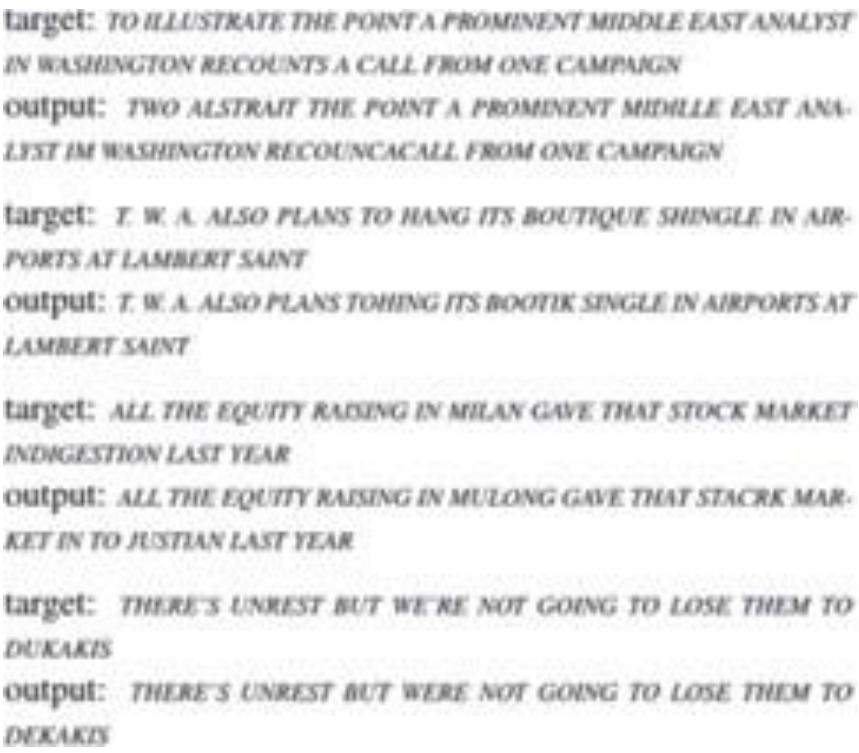

Fig 2.1 Input and output samples for the Alex Grave et al

As we can see the error is in the produced results, there would be phoneme errors, linguistic errors but apart form this the system arranged to get the accuracy without any prior model. The accuracy of the system can be increased using training data.

2.12 Andrew L. Maas et al (2015) [12], further improvement to the [11] is done by these researchers. They purposed only neural network model to map speech into characters. The work of these researchers makes the speech recognition model much easier than earlier one. They used LVCSR system that was built ion two neural networks. the outcome of this system is then compared with the GMM-HMM system model using beam search decoding.

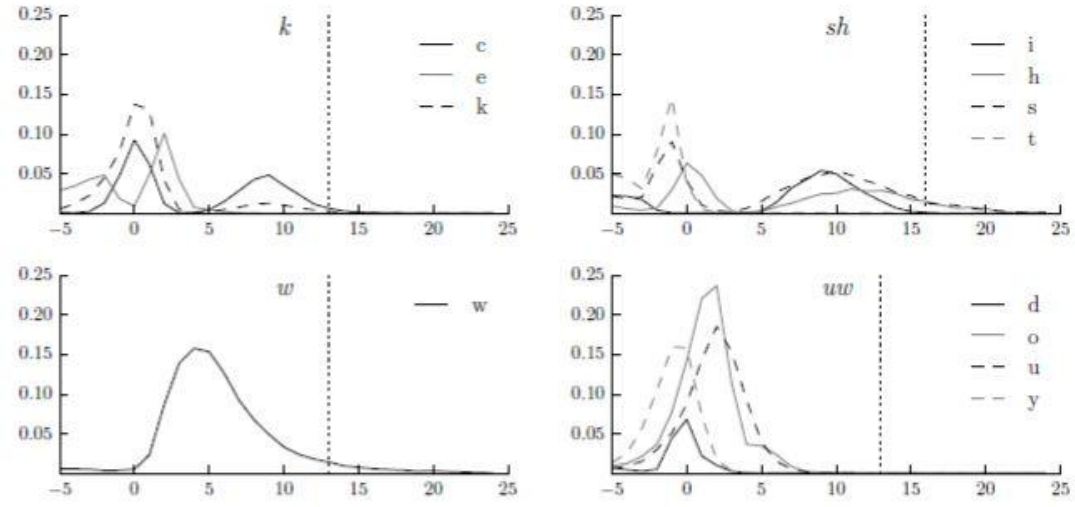

Fig 2.2 Character probability from the CTC based neural network and GMM-HMM based system

Using this technique, the first-pass LVCSR system outputs more than the HMM-GMM model. As we know DNN proves itself in the field of speech recognition. This work free the CTC based DNN from the HMM.

2.13 William Chan et al (2016) [12], purposed a recognizer that can perform Listen, Attend and Spell (LAS). This work was purposed without usage of pronunciation model. Using ANN, the pronunciation, language and acoustic models were subsumed, that's why it is also called end-to-end model. Where directly from the speech waveform, analysis some feature and such features are extracted from them. And these features directly used for the DNN, that were basically feedforward neural nets with more than one hidden layer. The working model of this system shows a greater achievement. Without any dictionary, it managed to get 10.3% of accuracy, earlier it was only 8.0% that was achieved using CLDNN-HMM. In this research author purposed a model that can directly convert acoustic signal into characters. Neural nets can be trained to map speech directly to its related word/spelling. This is an end-to-end model for speech recognition.

CHAPTER - 3

PROBLEM DEFINITION

Speech recognition is very important research in today's world. Interacting with machines using voice, can make the human life much easier than earlier technologies. If just saying a command to your refrigerator, it adjusts its temperature, then what we need further.

There are many ways proposed earlier for speech recognition, some of them are very famous and studied till many years by the students in this field. One of them is Hidden Markov Model, a probabilistic approach to detect the occurrences of any behavior, can be detected using HMM. But alone these models didn't perform so well in the past. Gaussian Mixture Model was used with HMM to give initial distributed probabilities to the HMM, this hybrid model HMM-GMM give more accurate results to match patterns of speech than simple HMM. But when neural networks came into the view of researchers of the speech recognition, the game totally changed from then. Now many researches have done research on speech recognition using neural networks and many papers were published related to these techniques. The work with the combination of different models are very dominant, those are called as hybrid models. Researchers used HMM-ANN Perceptron model, HMM-ANN feedforward neural network model etc. Now Deep neural networks are very popular among all the tech giants including Google, Apple etc. I want to do my work in this field only. Deep Neural Nets (DNN) are nothing but modified feed forward neural nets with more than one hidden layer. HMM are considered as dominant probabilistic model that is still in race with neural nets. The only difficulty in these systems is the learning process. Neural networks trained itself with training data provided to them. As much as training, the more accuracy can we get. The future in the speech recognition is bright. And I want to perceive my dissertation in this field using Neural Nets and Hidden Markov Model.

CHAPTER - 4

SCOPE OF STUDY

Speech recognition is dominant field of research in current world. This technology is booming very fast. As we see in every smartphone speech recognizer were installed to ease of usage. Speech recognition can be used in the dictation systems, translator systems. The universal translator can be created using speech recognition. The disabled person can use the technology with voice, who are unable to type, blind people can interact with machines without any difficulty, deaf people can use the technology to read what other said. The scope of speech recognition systems is for the benefit of the mankind. Voice controlled system can make our life much easier than now.

Neural networks are also very dominant in the technology. It can used to train any kind of system weather it's is face recognition, or voice recognition. All the complex computation can be done using these methods. It helps to create self-operated systems as well. We can depict the human decision making system with help of the neural nets. The scope of the Neural Net is also very esteem. It can be implemented in n number of systems.

Neural networks as well as speech recognition; both are dominant and excellent fields of study. We can predict the best only by using neural networks based speech recognizers.

CHAPTER - 5

OBJECTIVE OF STUDY

Speech recognition can be used in many dominant works including dictation, translation, giving commands to the smart systems etc. So, the objective of the speech recognition is dominant. My objective of the study is:

- Work with neural networks and probabilistic model like HMM.
- Determining how system is to be trained to convert the acoustic signals into characters.
- Study of acoustic signals, feature extraction and other models to be used in the speech recognition.
- Learning the concept of Pattern Matching.

CHAPTER – 6

PROPOSED RESEARCH METHODOLOGY

In this research I have to use neural networks and Hidden Markov Models. To implement these techniques I need data, firstly I will collect data related to my speech work. There is many free data resources available to use for the study purpose. These dataset are created by the linguistic experts with hundreds of hours of hard work. CMUDict is the one example of those datasets. After getting data I will build the model using different type of Neural Networks and Hidden Markov Model, and train that model using the available data. In the training phase I try to optimize the systems as much as I can. If training successfully completed, then the testing will be done on the trained model, and compare the outputs from different models I have created. The system that will give best result I will choose that system and try to optimize it.

CHAPTER – 7

EXPECTED OUTCOMES

Many earlier systems were using the ANN with HMM. My expectation should be to retrieve as much as accuracy with least training data or less computations. I will try to get least word error rate (WER) as compared to other systems. The expectation from the system is to translate the voice into text correctly in the robust environment also. The voice recognizer should be able to recognize the different speaking styles, accents. To create that system, I will use ANN and Hidden Markov Models, which are dominant in the industry. Although much work is done till date, but I expect further accuracy and the speed from the speech recognizer. The system that I will create can be used by the partially physical challenged persons i.e. deaf, they can use this system to change the acoustic signal around them into text and can behave over them. I will continue this field after my studies also and try to build a free app for the smart-phones to change the speech into text without internet.

CONCLUSION

In today's world speech recognition is dominant field of study in the field of research. Automatic Telephony, Smart appliances, Smart mobile interaction, Internet usability, all the area are changing drastically. Technology is changing human lives very fast, making everything easy on the fingertips of every person. Speech recognition can help people to interact such systems easily with voice. Many other benefits of the speech processing are there. My aim is to create speech recognition systems easily with the help of AI dominant neural networks that work as the human brain works, they can learn the patterns and the occurrences of the speech. I will make a hybrid model using HMM with those neural nets and predicted output is efficiency and accuracy from the system. Many research works done in this field earlier, I try to built the speech recognition system with Hidden Markov Model and Neural Networks. Different type of neural nets can be used in this research. I will try to take best type of neural net and try to build system with more accuracy.

REFERENCES

[1]A. Bellur and M. Elhilali, "Feedback-Driven Sensory Mapping Adaptation for Robust Speech Activity Detection", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 481-492, 2017.

[2]Karen Ullrich, Jan Schluter, and Thomas Grill, "BOUNDARY DETECTION IN MUSIC STRUCTURE ANALYSIS USING CONVOLUTIONAL NEURAL NETWORKS", *15th International Society for Music Information Retrieval Conference*, 2014

[3]O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, "Convolutional Neural Networks for Speech Recognition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533-1545, 2014.

[4]Wang A. et al. An Industrial Strength Audio Search Algorithm //ISMIR. – 2003. – T.2003. – C. 7-13.

[5] Kanejiya, Dharmendra and IIT Delhi. “Speech Recognition Using Hidden Markov Model With Neural Network Probability Estimators.” (2002).

[6]Al-Alaoui, Mohamad Adnan, Lina Al-Kanj, Jimmy Azar and Elias Yaacoub. “Speech Recognition using Artificial Neural Networks and Hidden Markov Models.” (2008).

[7]Dhavale Dhanashri and S. Dhonde, "Isolated Word Speech Recognition System Using Deep Neural Networks", *Proceedings of the International Conference on Data Engineering and Communication Technology*, pp. 9-17, 2016.

[8] Kamble, Bhushan C.. “Speech Recognition Using Artificial Neural Network – A Review.” (2016).

[9] Kapure Vijay Ramesh and Sonal Gahankari, “Hybrid Artificial Neural Network and Hidden Markov Model (ANN/HMM) for Speech and Speaker Recognition” *IJCA Proceedings on International conference on Green Computing and Technology ICGCT*(2):24-27, October 2013.

[10] Supriya S. Surwade, Dr. Y.S. Angal, June 15 Volume 3 Issue 6, “Speech Recognition Using HMM/ANN Hybrid Model”, *International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC)*, ISSN: 2321-8169, PP: 4154 - 4157, DOI: 10.17762/ijritcc2321-8169.1506133

[11]Graves, A. & Jaitly, N.. (2014), ”Towards End-To-End Speech Recognition with Recurrent Neural Networks” *Proceedings of the 31st International Conference on Machine Learning, in PMLR* 32(2):1764-1772

[12] Maas, Andrew L., Ziang Xie, Daniel Jurafsky and Andrew Y. Ng. “Lexicon-Free Conversational Speech Recognition with Neural Networks.” *HLT-NAACL* (2015).

[13] [1]W. Chan, N. Jaitly, Q. Le and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition", 2016 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[14]S. Eddy, "What is a hidden Markov model?", 2017. .

[15]Santosh, K.Gaikwad & Bharti, W.Gawali & Yannawar, Pravin. (2010). A Review on Speech Recognition Technique. *International Journal of Computer Applications*. 10. . 10.5120/1462-1976.