Automated Speech Recognition

Lokesh Purohit[1] Vikas Thakur[2] Kuldeep Maurya[3] Mr DK Mishra[4]
[4]Assistant Professor
[1,2,3,4]Department of Information Technology
[1,2,3,4]HMR Institute of technology & management, India

Abstract— Speech, the ability to express thoughts and feelings by sounds has been one of the foremost research topics. The ability of computer to understand human language and to reproduce human speech artificially by computers has been made possible by the efforts of researchers. It deals with interpreting the context of the speech generated by humans. This paper provides an overview of ASR along with its methodologies, techniques and its background. One of the most interesting research challenge in ASR technology is its accuracy and interpreting the speech in noisy environment. Automated speech recognition will continued to be one of the most important area of research in the development of artificial intelligent systems. Speaker identification is one other research area in ASR which can be pretty useful in the context of security.

Key words: Speech Recognition

I. INTRODUCTION

Automated speech recognition is the ability and adeptness of the machine or software to recognize and understand what the speaker has to say. This include getting the voice input from the user and then to interpret it so that the machine could understand what the speaker wants the machine to do [1]. Recent trends also revolving around speech recognition, there are various new technologies coming around such as google assistant, siri who has the capability to follow human commands through their voices. This speech recognition technology is growing superfast in this recent times. There have been various practices for the machine to train to learn human languages which includes natural language processing and others artificial intelligence methods [3]. The ASR systems are designed in such a way that they can communicate to human in real time and give responses to the human efficiently. Although there are speech recognition systems that are intelligent enough to communicate with human efficiently and gives dynamic answers to the exact same question, but still a fully automated speech recognition system which is not token or keyword based is still a proposed work [5]. We have obviously came closer to a fully automated system but still there are a lot of more work has to be done to achieve that level of intelligent system. ASR technology as a general term has numerous applications, and if this technology is used in a coherent manner it can not only solves plenty of human problems but also furnish human with a luxury to talk to machine without even giving typed commands. One such application of automated speech recognition which we are discussing here is car motion through ASR. Everything is becoming automated, and with this mindset if car motion get controlled by human voices, then it'll be a huge perk for the drivers.

II. RELATED WORK

The development in the field of automated speech recognition has travelled a long way. The development of speech recognition technology is in some sense similar to the growth of a baby. Eventually and gradually, ASR developed from the level of manual token based keyword searching to having a rich vocabulary library and giving quick and dynamic replies to different questions, an example of such latest ASR application is google assistant which gives instant and efficient replies to the user. Now listening to Siri and google assistant made us wonder how far this technology has reached and it has became an integral part of our life. From making calls to booking doctor's appointment by just uttering some words, because of ASR these things looks easy. Virtual assistants are becoming more and more popular because of this ease that they provide to humans. This is definitely making human life easier. Here's the overview of development of ASR in past few decades [6].

In 1950s and 1960s, the ASR was capable of recognizing only digits. Digits being simple to speak and discrete to identify, it makes proper sense. At that period, inventors and engineers were focusing on numbers. "Audrey" system was designed in Bell Laboratories which could recognize digits. That was a huge accomplishment at that time. That was the first milestone in the progress of speech recognition system. Then came "Shoebox" machine which could understand 16 words of English dictionary. The machine was trained for 16 English words. That was the first attempt to make machine understand human language [9].

Attempts were made in different counties Labs such as in United States, Japan and England for the improvement resulting in the enhancement in recognition to four vowels and some consonants. That was such an acceleration in this fields. Regardless of how primitive computers were at that time, improvements in this technology continued.

Because of the funding from U.S Department of defense, Speech recognition technology reached to new heights in 1970s. Harpy speech recognition system was developed as a result of this much attention to ASR system, I could understand 1011 words. It had the vocabulary stored in it as much as of a child. It could understand basic words which people use in day to day life. The path for a fully automated system was still very long but that system was a good start.

Beam search approach was used in Harpy, that's what made it more significant because it introduced a more efficient search approach. This approach proved that there can be finite network of possible sentences. In 70s only few other milestones were completed in the field of speech recognition technology. The very first commercial speech recognition companies are set up in 70s. Threshold technology and Bell Laboratories were founded in 70s only.

It could interpret multiple people's voices. Different simple accents were interpretable at that time.

In 1980's new approaches were founded to understand people voices. Drastic changes in the improvement of this technology was seen over the next decade. The vocabulary became wider, actually it jumped from about few hundred to few thousand words. Recognizing numbers was not a big deal at that time. Because of this new methods, the system was able to understand unlimited number of numbers. Hidden Markov Method was one of the new statistical method that came into existence over that period. It made the recognizing task much easier. Instead of token searching for words, Hidden Markov Method even considered the probability of different unknown sound being words. 80's was the time when the boom in this technology came in real sense. The extension in vocabulary turned into commercial applications very quickly for business and other industries. In 1987, worlds of wonder's Julie doll came into existence, which could be trained by the children to respond to their voices. ASR technology has travelled a long path from recognizing some finite numbers to becoming day to day need of humans. This technology has now even entered on the phones, smart TV and other smart applications.

III. PERFORMANCE MEASURES

Speech recognition is still a field of ongoing research. A model and approach so that the machine could interpret as much speech as possible. The area is still wide and open for research to get 100% effective system. It is still one of the main field of interest of many researchers. The performance of the automated speech recognition system can be measured in terms of following:

1) Accuracy: The efficiency of the recognition system is measured in terms of accuracy. More than 80-85% accuracy is called to be as good. The more the word it interprets, better the accuracy.
2) Memory: Memory constraint is another issue in the ASR system as well. The size or the total memory requirement of the system should be as small as possible.
3) Speed: The response time is the biggest factor in recognizing the sound. The faster the response, the better the system.

Speaker identification would only allow the system to accept the commands of the known speaker or the owner. Speaker identification is another application of speech recognition by which it can recognize the speaker.

Now a days this technology is also integrated with the smart phone lock system. Speech recognition for car motion can be pretty useful for physically disabled people which'll help them to drive the vehicle with ease. It can also play a significant role for new drivers or the ones who can't drive vehicle.

IV. TYPOLOGIES OF SPEECH RECOGNITION SYSTEM

[1] Speaker Dependent System: In this system, it require a user to train the system according to his or her voice before the recognition process.
[2] Speaker Independent System: In this system, it do not require a user to train the system that is, they are developed to operate for any speaker. Before training also for a particular voice, speech recognition system works.
[3] Isolated word recognizers System: In this system from the speech, each word is isolated. It accept one word at a time. Continuous speech can be recognized by this system.
[4] Connected word systems: It is a planned speech system which allow speaker to speak slowly. Each word should be spoken with a short pause.
[5] Spontaneous recognition systems: It is a method of recognition spontaneously. It allow us to speak spontaneously [8].

Now it is the time to articulate the research work with ideas gathered in above steps by adopting any of below suitable approaches:

V. APPROACHES TO ASR

There are mainly three approaches to automated speech recognition:

A. Acoustic Phonetic Approach:

Acoustic phonetic approach is also pretty popular by the name of rule-based approach. For the searching process in acoustic phonetic approach, the knowledge of phonetics and linguistic is used. There are rules in this approach that guides the searching of the speech that defines everything. It is not very common method to use, it is kind of outdated because it has poor performance. Individual phonemes, words and sentences are identified by this approach. It deals with acoustic aspects of sound. Acoustic phonetics approach deals with time domain characteristics like mean squared amplitude of wave, its time duration frequency. It also deals with the frequency domain of the speech like frequency spectrum. Acoustic phonetics and its study was improved in late 19th century because of Edison phonograph. Edison phonograph provided the facility to record the signal and then later process and analyze it. A spectrogram could be found by filtering the speech signal again and again with different band pass filter.

B. Pattern Recognition Approach:

Pattern recognition approach is the approach which basically deals with first training the machine based on the stored data to get a model that can recognize the pattern in the new data to identify it. This approach has mainly two steps that is training the machine and then to recognize the pattern by pattern comparison. In the first phase i.e., reference phase or training phase, the reference pattern is used to create the model. Once the model is created of the known pattern or the reference pattern, it can be used to compare with the test pattern. Before that, a number of measurements are made upon the input signal or the test signal. The unknown pattern is then compared with the known pattern with each sound reference pattern to measure the level of similarity score. In pattern recognition also, there are two types of matching approaches, Template Matching Approach and stochastic based approach [7].

C. Artificial Intelligence Recognition Approach:

Artificial intelligence recognition approach is the approach that is made by considering the other two approaches. It is the

combination of the other two recognition approaches i.e., acoustic approach and pattern recognition approach. Neural network is used in this approach to implement the expert system to recognize sounds and to classify them.

An automatic speech recognition system, keywords that were connected with the processing of telephone calls was developed. This system is still under development but the accuracy level on the basis of present data is always better through this.

VI. SPEECH RECOGNITION PROCESS

MFCC Feature Extraction: Features of a speech is a distinctive attribute or characteristics of the speech that can help in distinctly identifying or recognizing the speech. Feature extraction refers to extracting characteristics or attributes of the acoustic signals. It is important in recognition process. The efficiency of feature extraction is the most forehand task to develop an efficient speech recognition model. MFCC works like humans in hearing perceptions [3]. It cannot perceive frequency over 1KHz. MFCC and Logfbank are the two methods to extract features from the acoustic signals. In MFCC, there are two types of filters available that can worked linearly at frequency below 1000Hz. It can work on logarithmic spacing above 1000Hz. It focuses on capturing significant attributes of the phonetics.

Pre-emphasis is a system process which is used in MFCC to enhance the magnitude of some frequencies with respect to some others in order to improve the overall SNR. Energy of the signal is increased in this process.

In the process of framing, the speech samples is segmented into small frames of length within the range of 20 to 40msec. After that hamming window is used to multiply with the input signal to get the output window signal. Fourier transform of the window signal is the next step of to get the MFCC extracted features.

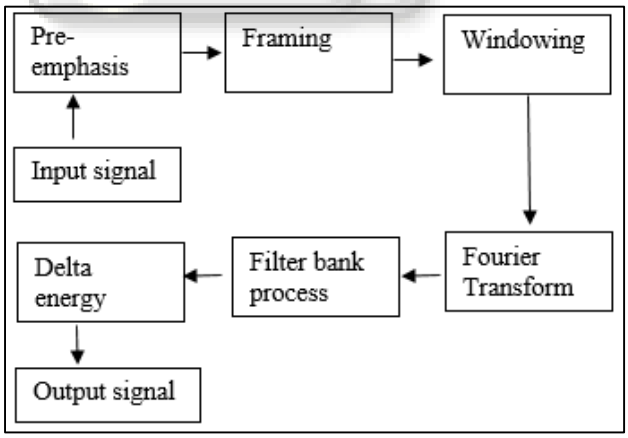

Fig. 1: MFCC features extraction

A. Feature matching:

Dynamic time warping is an algorithm based on dynamic programming. In matching two speech signals, the signals may vary in terms of time and speed. Dynamic time warping is the algorithm that handles this issue of time and speed. It measures the similarities between two series of times. It can also be used to find the proper alignment among two series if one of them may be warped non-linearly along the time axis. It could be shrinked or stretched along the time axis. This warping can be very helpful in finding the similarities between two time series. For aligning two series using dynamic time warping, an nbn-m is formed that contains the distance between the two points of both the signals. Then, by Euclidean distance, the absolute distance between the values of two sequence is calculated.

VII. SUPPORT VECTOR MACHINE

Support vector machine is one of the most powerful tools to classify different patterns. Support vector machine is also very useful for pattern recognition that makes it stands out from other classifiers to use it in the technology of speech recognition. Support vector machine (SVM) can classify both linear and non-linear data. It separates the classes by creating a hyperplane and by using two support vectors. SVM cannot be used to classify variable length data. First, the variable length data needs to be converted into a fixed length data vector, then only it can be applied on that data. It is the classifier with the highest margin fitting functions. SVM is independent of dimensionality and therefore it can utilize large dimensions space.

For applying SVMs to recognition of speech, there are a number of factors that needs to take into consideration. Segmentation needs to be specified in the features. The model needs to be trained for the features extracted in the form of Numpy matrix.

VIII. CONCLUSION

Automatic Speech Recognition Systems provide the capability to convert speech signals into understandable words. Because of its ability to convert real time speech, its application in Air traffic control and automated car environment has been studied. Hidden Markov model is used for Air traffic control in feature extraction while its phraseology is based on the commands used in air applications. Speech recognition is done used for route navigation application in car environment. In our small model we designed a robotic car which is capable of understanding various vocal commands such as forward, backward, left, right and stop. The robotic car is designed using the Arduino Uno and the code is written in machine learning using python script. The main moto to design the car was to make it capable of understanding the basic commands. The model is trained with more than thousand audio samples. SVM is used for training the model with the accuracy of more than 90 percent. At the end the car was not only capable of understanding the basic commands but is also capable of recognizing the speaker. We provided audio samples of four different person for training the model. The robotic car moves only on the command of these four individual voices. We used logfbank to extract features from the audio samples so that to make a numpy matrix of each audio sample for training.

REFERENCES

[1] Preeti Saini , Parneet Kaur, "Automatic Speech Recognition: A Review", International Journal of Engineering Trends and Technology
[2] Takialddin Al Smadi1, Huthaifa A. Al Issa2, Esam Trad3, Khalid A. Al Smadi, "Artificial Intelligence for

speech recognition based on neural networks", journal of signal and information processing, 2018,6,66-72
[3] Abhijeet kumar, "Voice command recognition system based on MFCC", Internations journal of Engineering Science and Technology, 2017,7335-7342
[4] Bharat Kumar Dhal1, "Controlling Devices Through Voice Based on AVR Microcontroller", International Journal of Scientific and Research Publications
[5] "Automatic Speech Recognition System: A Survey Report"
[6] B.H. Juang & Lawrence R. Rabiner, "Automatic Speech Recognition – A Brief History of the Technology Development "
[7] Faizan Mehmood, "An Overview on Speech Recognition System and Comparative Study of its Approaches", International Conference on Engineering & Emerging Technologies (ICEET-2014), At Superior University, Lahore, Volume: 1
[8] R K Aggarwal and M. Dave, "Markov Modeling in Hindi Speech Recognition System: A Review", CSI Journal of Computing, vol. 1, no.1,pp. 38-47, 2012
[9] Ayushi Trivedi,Navya Pant, Pinal Shah,Simran Sonik and Supriya Agrawal, "Speech to text and text to speech recognition systems-A review "

