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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

A Review on Speech Recognition by Machines

Amit Chugh
Assistant professor, Dept CSE
Manav Rachna international university
Faridabad, India

K. Swarna Krishnan
Student, Dept. CSE
Manav Rachna international university
Faridabad, India

K. Jerusha
Student, Dept. CSE
Manav Rachna international university
Faridabad, India

***Abstract:-* This paper discusses about the importance of speech recognition and also shows how the speech recognition evolved in these many years. This also shows the technological perspectives of the and progress in the important field of speech communication. These many years of research made the automatic speech recognition one of the challenging domains and made it an important research thing. The speech recognition at the time of invention or its early stage was not as effective as it is now, therefore many researchers worked on this domain and made it one of the exceptional aspects. This paper shows some of the important works or updates that was done by researchers to make this automated speech recognition (ASR) work as it is now. This paper discusses on how the speech recognition changed the way the world it is now with the help of mobile personal assistants and how the automatic speech recognition system works. In this paper we discussed briefly about the speech recognition techniques, feature extraction methods, types of speech of recognition. This paper also consists of the applications of speech recognition since its existence. This paper also shows the approaches of speech recognition and also shows the matching techniques like whole word match and sub word match. Literary survey shows the chronical order inventions and discoveries in the field of speech recognition. This paper gives the complete idea about speech recognition.**

Keywords-Automatic speech recognition system (ASR), Speech classifiers, Feature extraction, Acoustic signals, Language model, Neural networks.

I. INTRODUCTION

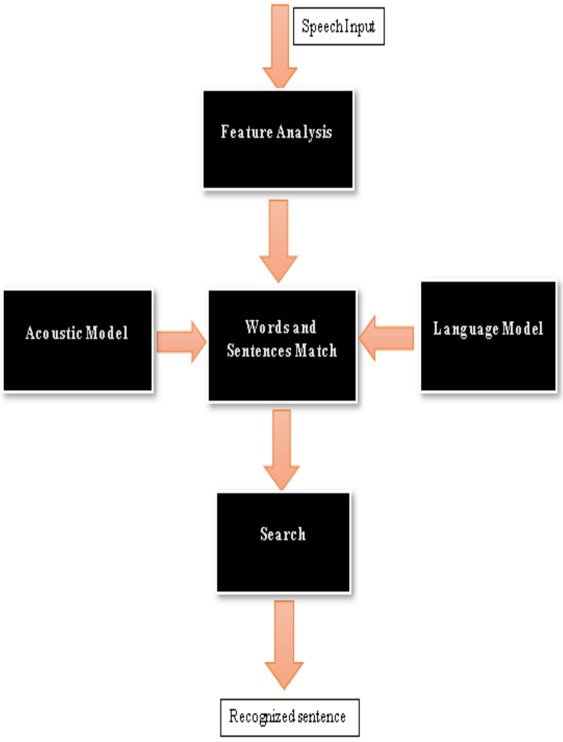
A. DEFINITION OF SPEECH RECOGNITION:

Speech Recognition is a process in which the speech signals are converted into words or sentences or sequence of words with the help of algorithms which are implemented as the computer programs.

B. SIMPLE MODEL OF SPEECH RECOGNITION:

The basic and natural form of human communication is called as speech where speech processing is the most exciting topics of processing of signals. Speech recognition made easy for computers to follow and understand the human voice commands and languages. There are certain systems of speech recognition that are to be trained to work effectively which means a human has to interact with the system and give some voice inputs with which the system trains and understand the commands. As everyone knows that speech is the primary way in which humans communicate with each other. Hence the speech recognition systems made the human life easier like these days we people have mobile assistants which help and notifies about our doubts, schedules that we have made on our phone and many more activities. The customer services use speech recognition to help the customers with problems and doubts. Other examples where speech recognition is used is banking, voice dictation, data entry, helps handicapped people, railway reservations etc. We see all these applications in our daily lives because of the tremendous and excellent increase in the statistical modelling of speech recognition. This paper focusses on the change and applications and updates that have taken place in the field of speech recognition to provide the technical concept and perspective to the people. This paper also ensures and helps as a reference in a few areas in the speech recognition which are a bit problematic or that needs to be tackled. In this speech recognition, for a system to understand the each and every word that is spoken by the human to the machine, *it* undergo*es* a certain process.

Figure 1: Simple Model of Speech Recognition



The above figure1 shows the simple model of speech recognition. In this whenever the speech input is given by a human to the machine is detected by the machine the input features are analysed in the next step after which the words and sentences match takes place with the help of acoustic models which includes the unit models and the lexicon , and language models which includes syntax and semantics to frame a sentence. After which the framed sentence is searched and then the sentence or the voice input given by the human is recognized. This is the basic model and working of the speech recognition system. The feature analysis is used to depict the properties of time varying voice or speech signals. The feature analysis works as an acoustic model front end in the above figure. For the large vocabulary sentences we assume a simple probabilistic condition where the words sequence is specified by W, and the sentences that are observed based on word sequences is specified by S with the probability P (W, S). Now our aim is to find the word strings with the help of sentences observed and the decoded or observed string has maximum a posteriori probability.

$$P\left(\frac{W}{A}\right) = \arg maxw\, P(\frac{W}{A}) \qquad (1)[1]$$

Using Baye's rule the (1) equation can be written as

$$P\left(\frac{W}{A}\right) = \frac{P\left(\frac{A}{W}\right)P(W)}{P(A)} \qquad (2)[1]$$

As we know that P(A) is independent of W the maximum a posteriori equation is

$$W = \arg maxw\, P\left(\frac{A}{W}\right) P(W) \qquad (3)[1]$$

In the equation (3) P(A/W) is called as acoustic model. Therefore P(A/W) is calculated. In the same equation P(W) is called as language model. Therefore, to find the word strings we need to search for the result from the acoustic model and the language model as shown in the figure1.

C. SPEECH RECOGNITION TYPES:

Speech recognition is divided into several types based on the utterances observer recognised. The classified types are:

i. ISOLATED WORDS:

This type lacks the audio signal. In this when there is a sample window then both sides should be quiet during the utterance. It accepts single word at a time. This type of system consists of listen and not listen states where the speaker or the person who is speaking has to be quiet after uttering words. Hence whenever the speaker is in silent state or the system is in not listen state that time the system processes the word that is uttered by the speaker before. This type of system is also called as discrete speech recognition systems.

ii. CONNECTED WORDS:

This type of system allows separate utterances to run together with minimum amounts of pauses. This kind of systems are also called as connected words system. This type of system is same as that of isolated words system.

iii. CONTINUOUS SPEECH:

Speaker can almost speak naturally in this type of speech recognition as while the system recognizes the content that the speaker is speaking. This is like computer dictation. These systems are made with difficult utterance boundaries which allows the speaker to speak continuously.

iv. SPONTANEOUS SPEECH:

These are very used for the speakers as there is no particular way to speak with it as the system can even recognize the sounds that are made by the speaker in between while thinking like 'UMM'. The automatic speech recognition systems with spontaneous speech are really difficult to code. This type of system allows the natural sounding and no rehearsed speeches from the speaker.

v. NATURAL LANGUAGE:

In this type of system, they not only recognize the words of utterances by the speaker but also give the reply to the questions and the doubts from the speaker.

Basically, the speech recognition systems are considered as two types which are speaker dependent systems and speaker independent systems.

vi. SPEAKER DEPENDENT SYSTEM:

In this, the systems need to be trained in order to recognize or understand or process the utterances by the speaker.

vii. SPEAKER INDEPENDENT SYSTEM:

In this, there is no need for the system to be trained as it recognizes and understand most user's voices.

D. APPLICATIONS OF SPEECH RECOGNITION:

There are several applications for speech recognition due to the popularity and flexible use of speech recognition in our daily lives. Speech recognition is used in many areas in our life right from knowing a doubt or a query to booking or reserving things that are needed. Some of the areas in which speech recognition are military, speech, telecommunication, recognition, communication sector, education sector, domestic sector etc.

a. COMMUNICATION SECTOR:

In this sector without the telephone directory people are able to make calls with the help of mobile assistants by using the voice and commands as the input.
Input form: speech in wave form.
Classes pattern: spoken words.

b. EDUCATION SECTOR:

In this sector this type of system is helpful for the students who have no hands and those who can't use keyboard to type things, those students can use the voice as input and get the results back in the speech form, for example with the help of automated speech recognition systems. To teach the foreign students the subject that particular teacher can use this type of system for the correct pronunciation and by narrating the topic they want to do research on and directly getting outputs on that exact topic rather than getting the human errors.
Input form: speech in wave form.
Classes pattern: spoken words.

c. TRANSLATION:

This type of system is an advanced system which is used to translate the speech from one language to another language. It can be used by anyone who wants to learn a new language or speak a new language and so on.
Input form: speech in wave form.
Classes pattern: spoken words.

d. HEALTH CARE:

These days people are using applications in their mobiles where the medical transcriptions are given based on the symptoms by the trained, database related speech recognition systems. In hospitals also we can see certain machines that work based on speech recognition.
Input form: speech in wave form.
Pattern classes: spoken words.
There are many other sectors in which speech recognition is used like health care, military and so on.

E. CLASSIFICATION OF AUTOMATIC SPEECH RECOGNITION SYSTEM (ASR):

Depending on the speech processing the automated speech recognition system is divided into Analysis, Coding and Recognition where recognition is again divided into Speaker recognition, Language Identification and Speech recognition. In which Speech recognition is again classified into Speech mode, Speaker mode, Vocabulary size and Speaking style. For a system to be more effective and functional it has to understand and process all the kind of dialects that people are giving commands with to the machine. The system must take care of the vocabulary that a speaker is using based on its size for example speaker can be using small words, large words and medium words. The system processes these words based on the input signals in the wave form which is why it has to take care of size of vocabulary for perfect and accurate results. Speaker mode can be speaker dependent, speaker independent and speaker adaptive. We see that a few people wish to use the voice recognition mode in some applications whereas others use the typing mode, no matter what a speaker chooses the system has to meet the requirements of the user by processing them with the required information. As elaborated in the types of speech recognition above they fall under speech modes of speech recognition where different kinds of systems perform different modes of operations as demanded by the speaker.

F. SPEECH RECOGNITION APPROACHES:

In speech recognition there are specifically three approaches. They are:

- Acoustic phonetic approach
- Pattern recognition approach
- Artificial intelligence approach

a) ACOUSTIC PHONETIC APPROACH:

One of the old approaches that is used is acoustic phonetic approach which is used in early systems that were designed called as speech understanding system in 1975 by Stanford research institute [2]. In 1979, a system was developed which is named as word verification system with the help of pre-stored lexical entries that was tested for continuous speech recognition for speakers in which the energy was taken from the band of frequencies that possess the higher frequency [3]. The detailed process in which the above discussed system works is mentioned in this [4]. To create an acoustic phonetic speech recognition system on knowledge based the detailed explanation and study on semi vowel sounds was done on [5]. The acoustic phonetic approach is dependent on the theory of acoustic phonetic and postulates [6]. Spectral analysis is the first step of acoustic phonetic approach where the speech and feature detection are combined and convert those spectral measurements into set of features which described the broad acoustic properties of various phonetics units. After the first step the spectral signals are segmented and different one or more phonetic units are attached to those segmented spectral signals which results in phoneme lattice characterization of speech. This step is also called as segmentation and labelling. In the final of this approach determines the final word from the phonetic label sequences which occurs from the above step. In the validation process the linguistic constraints are added or used to ensure the lexicon for word decoding based on the phoneme lattice. When it comes to the commercial applications these acoustic phonetic approach has not been used [7].

b) PATTERN RECOGNITION APPROACH:

The two main and important steps in this pattern recognition approach is pattern training and pattern comparison. This approach uses special feature which is a mathematical framework and establishes speech pattern representations for best pattern comparison the process of pattern recognition was developed and received much attention. In this pattern recognition approach the pattern representation should either be in statistical model or speech template. This approach can be applied on sound, phrase or a word. As said earlier the pattern comparison and pattern training are most important because in the pattern comparison stage of this approach the speeches which are unknown are directly compared with each pattern that was observed during the pattern training stage. After this process the unknown speech is identified with the goodness of the match of patterns. From the past onwards this pattern recognition approach became dominant in speech recognition [8].

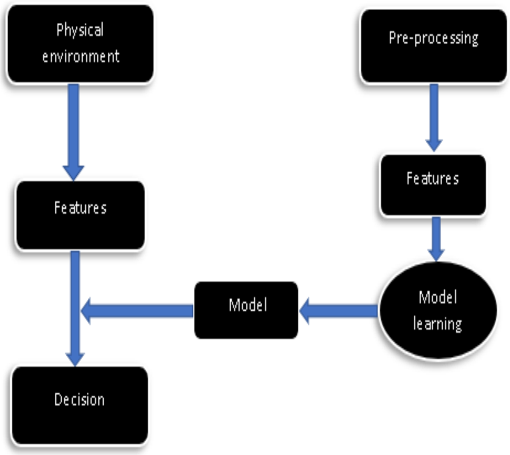


Figure 2: general pattern recognition system

The four pattern recognition approaches are:

- Template Matching Approach
- Statistical Classification Approach
- Syntactic Matching
- Neural Networks

1. TEMPLATE MATCHING APPROACH:

The template matching is done for the similarity between two similar entities like points, curves etc for the same type. The techniques that were mentioned in [9] are advanced in the field of speech recognition from the past decades. In this type of approaches the templates and prototypes are to be identified and are required to match the pattern with template. During this process of matching different operations can be performed to obtain the expected pattern. The operations that are allowed is rotation, translation and scale changes. For the entire words the templates are constructed. With the increase of vocabulary or the words the template preparation and template matching processes become expensive. For this template matching to be effective one of the key ideas is to use the dynamic programming to maintain the patterns to the account. One of the major drawbacks of this approach is that it has the fixed pre-defined templates which is why speech will only be modelled as per word many templates. This is how it can be modelled. This kind of becomes impractical, this is explained in [10].

2. STOCHASTIC APPROACH:

This approach uses the probabilistic models for uncertain information or the problematic information [9]. Generally, in speech recognition there will be incompleteness of information or the uncertainty of information because of the unknown sounds and others. In middle of all these problems to get the effective output we use this stochastic approach. Hidden markov model is the most and well used model in speech recognition in this generation and also the past. Set of output distributions and a finite state of markov model are used to characterize the hidden markov model. The spectral variabilities and temporal variabilities of this hidden markov model is the essence and essential part of speech recognition. The hidden markov model has the best and firmer mathematical base compared to the template matching based approach.

3. SYNTACTIC MATCHING:

In speech recognition most of the time there will be interconnected features in between the patterns, at that time we often use the hierarchical patterns where one pattern consists of sub-patterns which again consists of sub-patterns. This process of arranging patterns is considered as syntactic approach of pattern recognition. We often use array, strings, graphs and trees kind of data structures for the pattern representation [11]. Since we use these data structures it will be easy to define the relation between fundamental patterns and also easy to form the hierarchical patterns. By defining them it will be easy for the recognition by comparing the unknown patterns with the known or stored patterns as the information or the complex unknown patterns are formed from the simple and known patterns. This type of comparison is used to find the similarity between the unknown inputs and the known patterns.

4. DYNAMIC TIME WARPING:

Dynamic time warping is the process of finding the similarity between the sequences of the utterances by the speaker based on time and speed of utterances. This type of approaches is basically used in any data which can be converted to the linear representation of data. Basically, what a dynamic time warping does is that it finds the similarity between two different sequences which are having certain restrictions. In this dynamic time warping the sequence is warped non-linearly using a time dimension to find the similarity using certain non-linear variations. For example, when a person is walking slowly in one video and in other that same person is moving quickly or there is much difference in the accelerations in both the video using this approach it can be detected very easily. From the above example we will know one thing that the continuity of time is not very important in this approach. As we can also understand that using this approach one can find out the missing sequences easily.

Approach type	Representation of the approach	Function of recognition	Criteria
Template matching	Pixels Samples Curves	Distance measure Correlation	Error in classification
Statistical	Feature	Discriminant	Error in classification
Syntactic	Primitives	Grammar	Error in acceptance
Neural networks	Pixels Samples Features	Network	Mean square error

Table 1:types of pattern recognition approach

Table 1 represents the differences between the types of pattern recognition approach.

5. VECTOR QUANTIZATION:

Vector quantization technique is mainly used by the speech coders for the efficient data reduction. This technique is applied for automatic speech recognition. In this approach rather than considering the transmission rate of the speech in automatic speech recognition we search for the efficient use of codebooks for codebook searchers and reference models rather than searching for the evaluation methods. In interactive web response each vocabulary is provided with the separate VQ codebooks. In automatic speech recognition technique whenever a speech is given to the system then the system measures the lowest distance codebook from the training session of all the test speeches or words. In the basic vector quantization, the codebook entries can from any training words and they are not ordered because they don't follow any time information. The advantage of using vector quantization when compared to any other automatic speech recognition techniques for vocabularies of similar word is that VQ has speech transients which makes it easy to recognise the words. [9]

c) ARTIFICIAL INTELLIGENCE APPROACH:

Artificial intelligence approach is also called as knowledge base approach. Knowledge based approach is the combination of acoustic recognition approach and pattern recognition approach. This technique uses the spectrogram, phonetics and linguistic. To develop the rules speech sounds some of the speech researchers created an acoustic phonetic knowledge system. Meanwhile the approach based on template is used to provide a variety of speech recognition systems which are very effective by providing a basic knowledge or information about the human speech processing by enhancing the knowledge base system and clearing and making error analysis. This knowledge can be ensured with the study of spectrogram and by adding certain rules and procedures. In expert systems pure knowledge engineering is much welcomed. To quantify the expert knowledge is a very hectic and difficult job to do which is why these expert systems are having limited success. In this since it's a combination of two approaches and works based on the knowledge, to integrate the human knowledge and add it to a system is highly complicate. Integration of phonetics, syntax of utterance, semantics, access to lexical, pragmatics etc. In this type of systems algorithms helps to solve problems whereas knowledge helps the algorithms to work properly. [9]

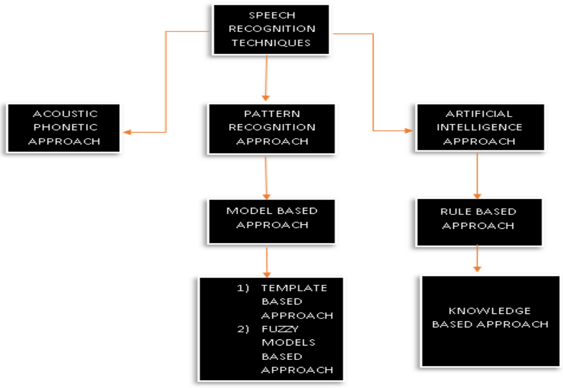

SPEECH RECOGNITION TECHNIQUES
ACOUSTIC PHONETIC APPROACH
PATTERN RECOGNITION APPROACH
ARTIFICIAL INTELLIGENCE APPROACH
MODEL BASED APPROACH
RULE BASED APPROACH
1) TEMPLATE BASED APPROACH
2) FUZZY MODELS BASED APPROACH
KNOWLEDGE BASED APPROACH

Figure 3:techniques of speech recognition

Figure 3 represents the speech recognition techniques that were discussed above.

G. FEATURE EXTRACTION TECHNIQUE:

This process chooses the input for the pattern recognition system. There are several ways and methods to extract the pattern. The feature should be selected in such a way that it should match the pattern or the task that one is performing. Feature extraction is done for different levels and based in the level for which the feature is extracted are responsible for the amount of error and amount of necessary pre-processing that is done to the features extracted. These features are represented in different ways some of them are continuous, discrete and discrete binary variables. Phase of the recognition process is measured at the time of feature extraction. These one or more measurements of feature function represents the characteristics of the object. Since there are many transformations these transformations are used to produce different features. These feature transformations are responsible for the strong reduction of information.

Method	Property	Implementation process
Principle component analysis (PCA)	Non-linear feature extraction, Fast, Eigen vector based, Linear map.	It is a traditional eigen vector base method called as karhuneu-loeve expansion. It is good for guassian data
Linear discriminate analysis (LDA)	Non-linear feature extraction, Fast, Eigen vector based, Supervised linear map.	This method is better than PCA. [1]
Independent component analysis (ICA)	Non- linear feature extraction, Linear map, Iterative non-guassian.	It is used for de-mixing of non-guassian fistributed features. Used for blind course separation.
Linear predictive coding	Static feature extraction, 10-16 lower order coefficient.	Mainly used for lower order feature extraction.
Cepstral analysis	Static feature extraction, Power spectrum.	It is used for the representation of spectral envelope.
Mel- frequency scale analysis	Static feature extraction, Spectral analysis.	Mel-frequency scale and fixed resolution are used to perform the spectral analysis.
Filter bank analysis	Frequencies tuned by filters.	It is transformed using a fourier transform. These amplitudes are highly correlated. [12]
Mel-frequency cepstrum (MFFC)	Fourier analysis is used to perform power spectrum.	It is used for pitch detection, voice identification and many more. The spectrum is firstly converted using mel scale to get the mel-frequency.cepstrum also called as MFC. [15]

Kernel based feature extraction	Non-linear transformations.	It is used for redundant features and also used to reduce errors. Dimensionality is used for better classification. [13]
Wavelet	It has better time resolution when compared to fourier transform.	It produces better time resolution at high frequencies. It can replace the fixed band width of fourier transform.
Dynamic feature extraction **a) LPC** **b) MFCC**	Accelerations and delta coefficients. 2nd and 3rd order derivatives of LPC and MFCC coefficients.	It is used for dynamic features.
Spectral subtraction	Robust feature extraction	It is used based on spectrogram. [14]
Cepstral mean subtraction	Robust feature extraction	It works on mean statistically parameter.
RASTA filtering	Noisy speech	It can recognize the feature in a noisy data.
Integrated phoneme subspace (compound method)	Transformation based on combination of PCA, LDA, ICA.	It has higher accuracy than the other existing methods. [6]

Table 2: feature extraction methods

Table 2 represents the feature extraction methods. This table shows the extraction methods along with the implementation method.

H. MATCHING TECHNIQUES:

There are two types of speech recognition matching techniques. The unknown word is matched with the known word by using these matching techniques [17]. The matching techniques are:

- Whole- word matching
- Sub-word matching

1) WHOLE-WORD MATCHING:

This matching technique requires much less processing when compared to sub-word matching. In this the digital audio signal or the speech audio signal of the speaker utterance is matched or compared with the pre-recorded signals of the speaker. Here the speaker has to pre-record the words or utterances that he is going to speak for processing. Sometimes he even needs to record more than 100 or 1000 words. They require large amount of storage. They only work if the recognition vocabulary is known. [16]

2) SUB-WORD MATCHING:

In this type of matching the systems searches for the sub-words or the phonemes and Then perform the pattern recognition by matching. these techniques usually take more time for processing since it has to find for the phonemes and then perform the operation when compared to the whole-word matching technique. But these techniques require less storage as there is no process of pre-recording the utterances. The pronunciation of the word can be found from the English text and not required for the speaker to record beforehand. [18][19] these papers discuss about the research in automatic speech recognition.

I. PERFORMANCE OF SPEECH RECOGNITION SYSTEMS:

Accuracy and speed are the components which determine the performance of the speech recognition systems [20][21]. Word error rate (WER) is used to find the performance accuracy and speed is the real time factor. Single word error rate (SWER) and command success rate (CSR) are others factors used for the measurement of accuracy. Vocalizations depends on many factors which are pitch, pronunciation, articulation, roughness, volume, nasality and many others which makes speech recognition by machines a very difficult job. The speech by a speaker can be distorted due to noise, echoes etc.

➢ ACCURACY:

Accuracy can be calculated with the help of word error rate, SWER, CSR. The accuracy of the speech recognition depends on many factors which are: [22]

- Vocabulary size and confusability of speech.
- Based on speaker dependence and independence.
- Isolated, discrete and continuous speech.
- Language constraints.
- Task constraints.
- Spontaneous speech.
- Environment conditions.

➢ WORD ERROR RATE:

Word error rate is a metric used to compute the performance accuracy of the speech recognition system or machine translating system. The difficulty of word error rate calculation occurs when the word length of recognized word differs from the reference word. By using dynamic word alignment first of all align the recognized word sequence and the reference word sequence then the formula to compute performance accuracy of a system is

$$WER = \frac{S+D+I}{N}$$

Let,

S = Number of substitutions,
D = Number of deletions,
I = Number of insertions,
N = Number of reference words.

To find the word recognition rate we need word error rate. To compute word recognition rate

$$WRR = 1 - WER = 1 - \frac{S+D+I}{N} = \frac{N-S-D-I-1}{N} = \frac{H-1}{N}$$

Where $H = N - (S + D)$, number of correctly recognized words.

II. LITERATURE SURVEY:

The first recognition machine came into existence on 1920 named as Radio Rex which is a toy [23]. In 1950's many researchers tried and investigated on speech recognition machines at that time and came to a conclusion that spectral resonance is observed at the regions of vowels which were extracted from the outputs from filter bank and logic circuits. Later in 1952, the invention of isolated digit recognition for a single speaker is made at bell laboratories by Davis, Biddulph and Balashek [24]. After that in 1956, at RCA laboratories Olson and Belar tried and built a machine which can recognize 10 distinct syllables spoken by a single speaker embodied in 10 monosyllabic words [25]. Fry and Denis in 1959 at University college in England built a machine that can recognize 4 vowels and 9 consonants [26].

Year of Invention	Details of Invention
1877	Thomas Edison is the first person who invented the very first device, phonograph that can record and reproduce the sound. It is very fragile and considered as prone to damage. [27]
1879	Later in 1879 Thomas Edison invented the first dictation machine which is considered as the improved version of his phonograph. [27]
1936	At bell labs a team of engineers led by Homer Dudley, invented the first speech electronic synthesizer called as Voder (Voice Demonstrator). [28]
1939	The patent was confirmed for Dudley for his invention Voder. [28]
1952	At bell labs a team designed a machine which is capable of understanding spoken digits. [29]
1962	IBM demonstrated shoebox that can understand 16 spoken words from the speaker at fair. [30]
1971	A device is invented by IBM named as Automatic call identification system using which a person can talk and receive the spoken answers from another person from the device. [31]
Early 80's	The technique named Hidden Markov Model is being put in use in machines for the first time since then. [29]
Mid 80's	IBM started working on a machine that can understand nearly 20,000 spoken words and it was named as Tangora. [31]
1987	The invention of World of Wonders a Julie doll which is toy is done and it is trained to respond to the speaker's voice and brought the speech recognition home. [29]
1990	A machine Dragon Dictate was invented and launched by Dragon company which is considered as the first speech recognition machine for customers. [29]
1993	The first built-in speech recognition and voice enabled control software was introduced to the apple computers which is known as Speakable items.
1993	In this year the first large vocabulary speech recognition system names SphinxII was invented by Xuedong Huang. [32]
1996	The first commercial product named MedSpeak came into light which is capable of recognizing continuous speech, it was invented by IBM. [31]
2007	GOOG-411 was launched by the Google company that served as the foundation for the future voice search product. It is a telephone-based directory service.
2011	This was the year when Apple company launched the digital personal assistant named as Siri. It can not only understand the speech by the user but also it does the appropriate actions based on the speech.
2014	The voice-controlled speaker called Echo which is powered by Alexa is invented by Amazon. This Echo is kind of similar to Cortana and Siri but is different in many aspects.

Table 3: timeline of invention in speech recognition

Table 3 shows the inventions that took place since the speech recognition came into existence. Since 2014 many updates were done in the field of speech recognition. In 2002, the introduction of speech recognition was done in the office apps by Microsoft. Later in 2007 Microsoft launched windows vista which is the first version having speech recognition. Before in 2006, National security agency started using speech recognition to recognize and understand the recorded conversations which are private. The aspect which is admiring about speech recognition is that people understanding and responding to the speech or voice is a very common thing but for a machine that don't have a brain to understand, get trained, analyse and respond to the voice or the commands is a very amusing thing.

III. FUTURE SCOPE:

This paper comprises the information on speech recognition that is known or which is invented or designed or understood or explained or spoken of or discussed at length since the time it was known to have existed. What is known about speech recognition is very limited and the gap between what is known and what is yet to be explored is magnificent. This paper can help the researchers to understand the speech recognition properly also it can help to know what had happened in the field of speech recognition till now. The future inventions should be more exciting in this field for example consider an environment where large number of speakers are there and the machine's work is to understand who is speaking from those large number of speakers and do the appropriate action according to the command or the speech. This paper can help the researcher to grasp the knowledge about the technology. In [9] this R.K. Moore showed a few points which are considered to be very important to understand speech, speech processing which didn't get the answer till now, some of them are Is speech special? Why is speech contrastive? And so on, he really asked excellent questions which are really difficult to answer till this year.

IV. ADVANTAGES OF SPEECH RECOGNITION:

Speech recognition has numerous uses in our day to day lives. Using speech recognition, the call centres are able to transcribe the calls based on the common issues and common call patterns. Certain applications are introduced in the mobile phones that is based on speech recognition. For example; in certain medical apps, people first of all talk to the trained machine that questions about the symptoms, etc and then based on the information that it gets, the particular doctor is connected to that person according to the symptoms noted. There are still many more applications of speech recognition. Google's google assistant, Alexa, Siri, Cortana and other digital personal assistants work based on this technology. Using speech recognition, through conversational applications retail companies are extending their sales with customers. In this way speech recognition is reducing the cost and making it efficient and speed effective process. It's also decreasing the human effort of doing certain difficult tasks. From the customers or consumers point of view speech recognition is providing the security, convenience and accessibility. There are many other applications of speech recognition which is reducing the human effort and making the work simpler.

V. LIMITATIONS OF SPEECH RECOGNITION:

Speech recognition has a lot of advantages in the day to day life of each and every person. Along with these advantages there are a bunch of disadvantages too that needs attention. Some of the limitations of speech recognition includes:

- The conversion of speech signal into the text is neither more efficient not with good accuracy, irrespective of the environment of the speaker and certain other parameters.
- Generation of automatic word lexicons.
- Generation of automatic speech models for different new tasks.
- Theoretical limit for speech recognition implementation calculation.
- No matter how accurate the speech recognition it can never surpass a human in this category. Nut it would be fantastic if this happens.
- Algorithm of optimal utterance verification and rejection.

VI. CONCLUSION AND DISCUSSION:

Speech is one of the most effective and natural ways of communications. Due to the interest in this field, many machines were invented in the past decades that could recognize, understand and respond to the speech. As we can see there is really a tremendous growth in this area and also many applications, software and machines were invented. There are also practical limitations which hinder the use of services and applications. Since there are certain limitations in this area, it becomes one of the exciting topics to push off the limits and spread its wings. The researchers and enthusiasts are showing interest in this field to increase the performance of speech recognition. In future, the speech recognition problems or drawbacks that is now present should not be there and those future problems and limitations should be even more difficult than those of now. In this paper, we have tried to show progress has been done till now since its existence. Though this technology has increased in past four decades tremendously there is still that is to be done. Speech recognition is expected to flourish more in the areas of human- machine interaction and also to develop more in the future. Our aim is to bring about understanding through this paper to the researchers working in this field.

VII. REFERENCES:

[1] Anasuya, M. A., & Katti, S. K. (2009). Speech recognition by machine: A review. International Journal of Computer Science and Information Security, 6, 181-205.

[2] Walker, D. (1975). The SRI speech understanding system. *IEEE transactions on acoustics, speech, and signal processing*, *23*(5), 397-416.

[3] Bhagath, P., & Das, P. K. (2004). Acoustic Phonetic Approach for Speech Recognition: A Review. *Language*, *77*, 93.

[4] Diller, T. (1979, April). Phonetic word verification. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 4, pp. 256-261). IEEE.

[5] Green, P., & Wood, A. (1986, April). A representational approach to knowledge-based acoustic-phonetic processing in speech recognition. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 11, pp. 1205-1208). IEEE.

[6] Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, *25*(1-2), 21-52.

[7] Reddy, D. R. (1966). Approach to computer speech recognition by direct analysis of the speech wave. *The Journal of the Acoustical Society of America*, *40*(5), 1273-1273.

[8] Myers, C., & Rabiner, L. (1981). A level building dynamic time warping algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *29*(2), 284-297.

[9] Moore, R. K. (1994, September). Twenty things we still don't know about speech. In *Proc. CRIM/FORWISS Workshop on Progress and Prospects of speech Research on Technology*.

[10] Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, *26*(1), 43-49.

[11] Fu, K. S., & Bhargava, B. K. (1973). Tree systems for syntactic pattern recognition. *IEEE Transactions on Computers*, *100*(12), 1087-1099.

[12] Filter bank analysis (https://labrosa.ee.columbia.edu/doc/HTKBook21/node54.html)

[13] Schutte, K. T. (2009). *Parts-based models and local features for automatic speech recognition* (Doctoral dissertation, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science).

[14] Campbell_, W. M., Reynolds_, D. S. W. S. D., & Navratily, J. (2006). The MIT-LL/IBM Speaker recognition System using High performance reduced Complexity recognition.

[15] Wikipedia Cepstrum (https://en.wikipedia.org/wiki/Cepstrum)

[16] Katagiri, S. (2003). Speech pattern recognition using neural networks. *Pattern Recognition in Speech and Language Processing*, 115-147.

[17] Razak, Z., & Ibrahim, N. J. (2010). emran mohd tamil, mohd Yamani Idna Idris, Mohd yaakob Yusoff. *Quranic verse recition feature extraction using mel frequency ceostral coefficient (MFCC).*

[18] Tran, D. T. (2000). *Fuzzy approaches to speech and speaker recognition* (Doctoral dissertation, university of Canberra).

[19] Rabiner, L., & Juang, B. H. (1993). Fundamental Of Speech Recognition Prentice-hall International.

[20] Ciaramella, A. (1993). A prototype performance evaluation report. *Sundial workpackage*, *8000*.

[21] Gerbino, E., Baggia, P., Ciaramella, A., & Rullent, C. (1993, April). Test and evaluation of a spoken dialogue system. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 2, pp. 135-138). IEEE.

[22] Rathor, A. S., & Mishra, P. K. (2014). Word Recognition Using Barthannwin Wave Filter and Neural Network. In *Proceedings of the Third International Conference on Soft Computing for Problem Solving* (pp. 99-111). Springer, New Delhi.

[23] Windmann, S., & Haeb-Umbach, R. (2009). Approaches to iterative speech feature enhancement and recognition. *IEEE Transactions on audio, speech, and Language processing*, *17*(5), 974-984.

[24] Davis, K. H., Biddulph, R., & Balashek, S. (1952). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, *24*(6), 637-642.

[25] Olson, H. F., & Belar, H. (1956). Phonetic typewriter. *The Journal of the Acoustical Society of America*, *28*(6), 1072-1081.

[26] Fry, D. B. (1959). Theoretical aspects of mechanical speech recognition. *Journal of the British Institution of Radio Engineers*, *19*(4), 211-218.

[27] Newville, L. J. (1959). *Development of the Phonograph at Alexander Graham Bell's Volta Laboratory* (No. 218). Smithsonian Institution.

[28] Dudley, H. W. (1938). *U.S. Patent No. 2,121,142*. Washington, DC: U.S. Patent and Trademark Office.

[29] Speech recognition through the decade: How we ended up with siri (https://www.pcworld.com/article/243060/speech_recognition_through_the_decades_how_we_ended_up_with_siri.html)

[30] Maney, K., Hamm, S., & O'Brien, J. (2011). *Making the world work better: the ideas that shaped a century and a company*. Pearson Education.

[31] Pioneering speech recognition (https://www.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/)

[32] Lee, K. F., Hon, H. W., & Reddy, R. (1990). An overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *38*(1), 35-45.