Automatic Speech Recognition: A Review

[1]Iqbaldeep Kaur, [2]Navneet Kaur, [3]Amandeep Ummat, [4]Jaspreet Kaur, [5]Navjot Kaur
[1,2,3,4,5]Dept. of CSE, Chandigarh Engineering College, Landran, Punjab, India

Abstract
This research study aims to present automatic speech recognition system and discuss the major themes and advances made in the past 60 years of research, so as to provide a technological perspective and an appreciation of the fundamental progress that has been accomplished in this important area of speech communication. After years of research and development the accuracy of automatic speech recognition remains one of the important research challenges. The design of Speech Recognition system requires careful attentions to the following: Definition of various types of speech classes, speech recognition process, ASR design issues, and speech recognition techniques. The objective of this review paper is to summarize and compare some of the well-known methods used in various stages of speech recognition system and identify research topic and applications which are at the forefront of this exciting and challenging field.

Keywords
Automatic Speech Recognition, Models, Speech Classes

I. Introduction

"Automatic speech recognition system can be defined as independent, computer-driven transcription of spoken language into readable text in real time."

ASR is technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it to written text. Having a machine to understand fluently spoken speech has driven speech research for more than 60 years. Although ASR technology is not yet at the point where machines understand all speech, in any acoustic environment, or by any person, it is used on a day-to-day basis in a number of applications and services. The ultimate goal of ASR research is to allow a computer to recognize in real-time, with 100% accuracy, all words that are intelligibly spoken by any person, independent of vocabulary size, noise, speaker characteristics or accent. Today, if the system is trained to learn an individual speaker's voice, then much larger vocabularies are possible and accuracy can be greater than 90%. Commercially available ASR systems usually require only a short period of speaker training and may successfully capture continuous speech with a large vocabulary at normal pace with a very high accuracy. Most commercial companies claim that recognition software can achieve between 98% to 99% accuracy if operated under optimal conditions. `Optimal conditions' usually assume that users: have speech characteristics which match the training data, can achieve proper speaker adaptation, and work in a clean noise environment (e.g. quiet space).

The main goal of speech recognition area is to develop techniques and systems for speech input to machine. Speech is the primary means of communication between humans. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to desire to automate simple tasks which necessitates human machine interactions and research in automatic speech recognition by machines has attracted a great deal of attention for sixty years [3-4]. Based on major advances in statistical modeling of speech, automatic speech recognition systems today find widespread application in tasks that require human machine interface, such as automatic call processing in telephone networks, and query based information systems that provide updated travel information, stock price quotations, weather reports, Data entry, voice dictation, access to information: travel, banking, Commands, Automobile portal, speech transcription, Handicapped people (blind people) supermarket, railway reservations etc. Speech recognition technology was increasingly used within telephone networks to automate as well as to enhance the operator services. This report reviews major highlights during the last six decades in the research and development of automatic speech recognition, so as to provide a technological perspective. Although many technological progresses have been made, still there remain many research issues that need to be tackled.

Fig. 1 shows speech recognition system in simple equations which contain front end unit, decoder unit, model unit, language model unit, and database. The recognition process is shown below (Fig .1).

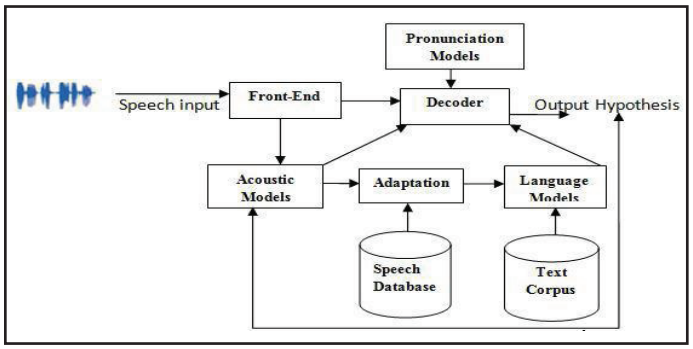

Fig. 1: Basic Model of Speech Recognition

Fig. 1 shows the basic model of speech recognition system and description of models and speech database is given below:

A. Models

- According to the speech structure, three models are used in speech recognition to do the match:
- An acoustic model contains acoustic properties for each senone. There are context-independent models that contain properties (most probable feature vectors for each phone) and context-dependent ones (built from senones with context).
- A phonetic dictionary contains a mapping from words to phones. This mapping is not very effective. For example, only two to three pronunciation variants are noted in it, but it's practical enough most of the time. The dictionary is not the only variant of mapper from words to phones. It could be done with some complex function learned with a machine learning algorithm.
- A language model is used to restrict word search. It defines which word could follow previously recognized words (remember that matching is a sequential process) and helps to significantly restrict the matching process by stripping words that are not probable. Most common language models used are n-gram language models-these contain statistics of

word sequences-and finite state language models-these define speech sequences by finite state automation, sometimes with weights. To reach a good accuracy rate, your language model must be very successful in search space restriction. This means it should be very good at predicting the next word. A language model usually restricts the vocabulary considered to the words it contains. That's an issue for name recognition. To deal with this, a language model can contain smaller chunks like subwords or even phones. Search space restriction in this case is usually worse and corresponding recognition accuracies are lower than with a word-based language model.

B. Speech Databases

Speech databases have a wider use in Automatic Speech Recognition. They are also used in other important applications like, Automatic speech synthesis, coding and analysis including speaker and language identification and verification. All these applications require large amounts of recoded database. Different types of databases that are used for speech recognition applications are discussed along with its taxonomy.

1. Taxonomy of Existing Speech Databases

The intra-speaker and inter-speaker variability are important parameters for a speech database. Intra-speaker variability is very important for speaker recognition performance. The intra-speaker variation can originate from a variable speaking rate, changing emotions or other mental variables, and in environment noise. The variance brought by different speakers is denoted inter-speaker variance and is caused by the individual variability in vocal systems involving source excitation, vocal tract articulation, lips and/or nostril radiation. If the inter-speaker variability dominates the intra-speaker variability, speaker recognition is feasible. Speech databases are most commonly classified into single-session and multisession. Multi-session databases allow estimation of temporal intra-speaker variability. According to the acoustic environment, databases are recorded either in noise free environment, such as in the sound booth, or with office/home noise. Moreover, according to the purpose of the databases, some corpora are designed for developing and evaluating speech recognition, for instance TIMIT, and some are specially designed for speaker recognition, such as SIVA, Polycost and YOHO. Many databases were recorded in one native language of recording subjects; however there are also multi-language databases with non-native language of speakers, in which case, the language and speech recognition become the additional use of those databases.

II. Types of Speech Recognition

Speech recognition systems can be separated in several different classes by describing what types of utterances they have the ability to recognize. These classes are classified as the following:

A. Isolated Words

Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on both sides of the sample window. It accepts single words or single utterance at a time. These systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses). Isolated Utterance might be a better name for this class.

B. Connected Words

Connected word systems (or more correctly 'connected utterances') are similar to isolated words, but allows separate utterances to be 'run-together' with a minimal pause between them.

C. Continuous Speech

Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content.(Basically, it's computer dictation). Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize special methods to determine utterance boundaries.

D. Spontaneous Speech

At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters.

III. Automatic Speech Recognition System Classification

The following tree structure emphasizes the speech processing applications. Depending on the chosen criterion, Automatic Speech Recognition systems can be classified as shown in fig. 2.

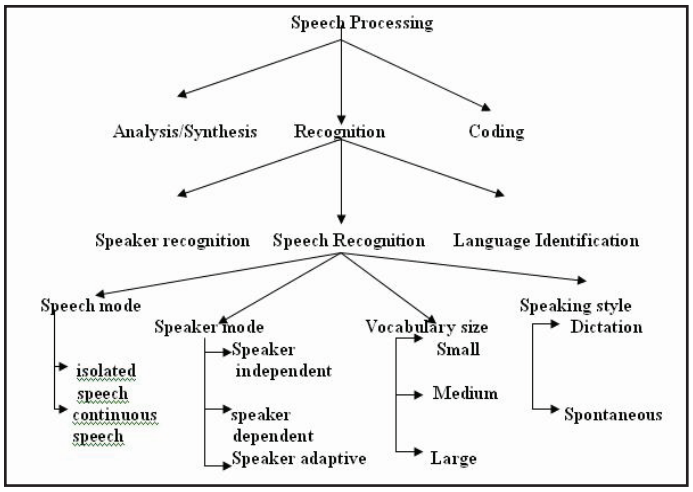

Fig. 2: Speech Processing Classification

IV. Relevant Issues of ASR Design

Main issues on which recognition accuracy depends have been presented in the Table 1.

Table 1: Relevant Issues of ASR Design

Environment	Type of noise; signal/noise ratio;
	working conditions
Transducer	Microphone; telephone
Channel	Band amplitude; distortion; echo
Speakers	Speaker dependence/independence Sex, Age;
	physical and psychical state
Speech styles	Voice tone(quiet, normal, shouted);
	Production(isolated words or continuous
	speech read or spontaneous speech)
	Speed(slow, normal, fast)
Vocabulary	Characteristics of available training
	data; specific or generic vocabulary

Table 1 shows some factors due to which some problems occur in ASR design process.

V. Speech Recognition Techniques

The goal of speech recognition is for a machine to be able to "hear," understand," and "act upon" spoken information. The earliest speech recognition systems were first attempted in the early 1950s at Bell Laboratories, Davis, Biddulph and Balashek developed an isolated digit Recognition system for a single speaker [2]. The goal of automatic speaker recognisation is to analyze, extract characterize and recognize information about the speaker identity. The speaker recognition system may be viewed as working in a four stages:

- Analysis
- Feature extraction
- Modeling
- Testing

A. Speech Analysis Technique

Speech data contain different type of information that shows a speaker identity. This includes speaker specific information due to vocal tract, excitation source and behavior feature. The information about the behavior feature also embedded in signal and that can be used for speaker recognition. The speech analysis stages deals with suitable frame size for segmenting speech signal for further analysis and extracting. The speech analysis technique done with following three techniques

1. Segmentation Analysis

In this case speech is analyzed using the frame size and shift in the range of 10-30 ms to extract speaker information. Study made in used segmented analysis to extract vocal tract information of speaker recognition.

2. Sub Segmental Analysis

Speech analyzed using the frame size and shift in range 3-5 ms is known as Sub segmental analysis. This technique is used to mainly analyze and extract the characteristic of the excitation state.

3. Supra Segmental Analysis

In this case, speech is analyzed using the frame size this technique is technique is used mainly to analyze and characteristic due to behavior character of the speaker.

B. Feature Extraction Technique

Theoretically, it should be possible to recognize speech directly from the digitized waveform. However, because of the large variability of the speech signal, it is better to perform some feature extraction that would reduce that variability. Particularly, eliminating various source of information, such as whether the sound is voiced or unvoiced and, if voiced, it eliminates the effect of the periodicity or pitch, amplitude of excitation signal and fundamental frequency etc. The reason for computing the short-term spectrum is that the cochlea of the human ear performs a quasi-frequency analysis. The analysis in the cochlea takes place on a nonlinear frequency scale (known as the Bark scale or the Mel scale). This scale is approximately linear up to about 1000 Hz and is approximately logarithmic thereafter. So, in the feature extraction, it is very common to perform a frequency warping of the frequency axis after the spectral computation.

C. Modeling Technique

The objective of modeling technique is to generate speaker models using speaker specific feature vector. The speaker modeling technique divided into two classification speaker recognition and speaker identification. The speaker identification technique automatically identify who is speaking on basis of individual information integrated in speech signal The speaker recognition is also divided into two parts that means speaker dependent and speaker independent. In the speaker independent mode of the speech recognition the computer should ignore the speaker specific characteristics of the speech signal and extract the intended message .on the other hand in case of speaker recognition machine should extract speaker characteristics in the acoustic signal. The main aim of speaker identification is comparing a speech signal from an unknown speaker to a database of known speaker .The system can recognize the speaker, which has been trained with a number of speakers. Speaker recognition can also be divided into two methods, text- dependent and text independent methods. In text dependent method the speaker say key words or sentences having the same text for both training and recognition trials, whereas text independent does not rely on a specific texts being spoken. The Following are the modeling which can be used in speech recognition process:

1. The Acoustic-Phonetic Approach

This method is indeed viable and has been studied in great depth for more than 40 years. This approach is based upon theory of acoustic phonetics and postulates. The earliest approaches to speech recognition were based on finding speech sounds and providing appropriate labels to these sounds. This is the basis of the acoustic phonetic approach (Hemdal and Hughes 1967). Which postulates that there exist finite, distinctive phonetic units (phonemes) in spoken language and that these units are broadly characterized by a set of acoustics properties that are manifested in the speech signal over time? Even though, the acoustic properties of phonetic units are highly variable, both with speakers and with neighboring sounds (the so-called co articulation effect), it is assumed in the acoustic-phonetic approach that the rules governing the variability are straightforward and can be readily learned by a machine [5]. Formal evaluations conducted by the National Institute of Science and Technology (NIST) in 1996 demonstrated that the most successful approach to automatic language identification (LID) uses the phonotactic content of a speech signal to discriminate among a set of languages.

2. Pattern Recognition Approach

The pattern-matching approach (Itakura 1975; Rabiner 1989; Rabiner and Juang 1993) involves two essential steps namely, pattern training and pattern comparison. The essential feature of this approach is that it uses a well formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm. A pattern recognition has been developed over two decade received much attention and applied widely too many practical pattern recognition problem. A speech pattern representation can be in the form of a speech template or a statistical model (e.g., a HIDDEN MARKOV MODEL or HMM) and can be applied to a sound (smaller than a word), a word, or a phrase. In the pattern comparison stage of the approach, a direct comparison is made between the unknown speeches (the speech to be recognized) with each possible pattern learned in the training stage in order to determine the identity of the unknown

according to the goodness of match of the patterns. The pattern-matching approach has become the predominant method for speech recognition in the last six decades.

3. Knowledge Based Approaches

An expert knowledge about variations in speech is hand coded into a system. This has the advantage of explicit modeling variations in speech; but unfortunately such expert knowledge is difficult to obtain and use successfully. Thus this approach was judged to be impractical and automatic learning procedure was sought instead. Vector Quantization (VQ) is often applied to ASR. It is useful for speech coders, i.e., efficient data reduction. Since transmission rate is not a major issue for ASR, the utility of VQ here lies in the efficiency of using compact codebooks for reference models and codebook searcher in place of more costly evaluation methods. For IWR, each vocabulary word gets its own VQ codebook, based on training sequence of several repetitions of the word. The test speech is evaluated by all codebooks and ASR chooses the word whose codebook yields the lowest distance measure.

4. The Artificial Intelligence Approach

The artificial intelligence approach attempts to mechanize the recognition procedure According to the way a person applies its intelligence in visualizing, analyzing, and finally making a decision on the measured acoustic features. Expert system is used widely in this approach. The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. In this, exploits the ideas and concepts of Acoustic phonetic and pattern recognition methods. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. Some speech researchers developed recognition system that used acoustic phonetic knowledge to develop classification rules for speech sounds. While template based approaches have been very effective in the design of a variety of speech recognition systems; they provided little insight about human speech processing, thereby making error analysis and knowledge-based system enhancement difficult. In its pure form, knowledge engineering design involves the direct and explicit incorporation of expert s speech knowledge into a recognition system. This knowledge is usually derived from careful study of spectrograms and is incorporated using rules or procedures. Pure knowledge engineering was also motivated by the interest and research in expert systems. However, this approach had only limited success, largely due to the difficulty in quantifying expert knowledge. Another difficult problem is the integration of many levels of human knowledge phonetics, phonotactics, lexical access, syntax, semantics and pragmatics. Alternatively, combining independent and asynchronous knowledge sources optimally remains an unsolved problem. In more indirect forms, knowledge has also been used to guide the design of the models and algorithms of other techniques such as template matching and stochastic modeling. This form of knowledge application makes an important distinction between knowledge and algorithms. Algorithms enable us to solve problems. Knowledge enables the algorithms to work better. This form of knowledge based system enhancement has contributed considerably to the design of all successful strategies reported. It plays an important role in the selection of a suitable input representation, the definition of units of speech, or the design of the recognition algorithm itself.

D. Matching Techniques

Speech-recognition engines match a detected word to a known word using one of the following techniques.

1. Whole-word matching

The engine compares the incoming digital-audio signal against aprerecorded template of the word. This technique takes much less processing than sub-word matching, but it requires that the user (or someone) prerecord every word that will be recognized - sometimes several hundred thousand words. Whole-word templates also require large amounts of storage (between 50 and 512 bytes per word) and are practical only if the recognition vocabulary is known when the application is developed.

2. Sub-word Matching

The engine looks for sub-words–usually phonemes and then performsfurther pattern recognition on those. This technique takes more processing than whole-word matching, but it requires much less storage (between 5 and 20 bytes per word). In addition, the pronunciation of the word can be guessed from English text without requiring the user to speak the word beforehand.

VI. Literature Survey of Speech Recognition: (Year Vise)

The progress of automatic speech recognition (ASR) technology in the past 6 decades can be summarized as follows:

A. Speech recognition systems in 1920-1960s

In the early 1920s machine recognition came into existence. The first machine to recognize speech to any significant degree commercially named, Radio Rex (toy) was manufactured in 1920.Research into the concepts of speech technology began as early as 1936 at Bell Labs. In 1939, Bell Labs demonstrated a speech synthesis machine (which simulates talking) at the World Fair in New York. Bell Labs later abandoned efforts to develop speech-simulated listening and recognition; based on an incorrect conclusion that artificial intelligence would ultimately be necessary for success. The earliest attempts to devise systems for automatic speech recognition by machine were made in 1950s, when various researchers tried to exploit the fundamental ideas of acoustic phonetics. During 1950s, most of the speech recognition systems investigated spectral resonances during the vowel region of each utterance which were extracted from output signals of an analogue filter bank and logic circuits. In 1952, at Bell laboratories, Davis, Biddulph, and Balashek built a system for isolated digit recognition for a single speaker. In an independent effort at RCA Laboratories in 1956, Olson and Belar tried to recognize 10 distinct syllables of a single talker, as embodied in 10 monosyllabic words. The system again relied on spectral measurements (as provided by an analog filter bank) primarily during vowel regions. In 1959, at University College in England, Fry and Denes tried to build a phoneme recognizer to recognize four vowels and nine consonants.

B. Speech recognition systems in 1960-1970s

In the 1960s several fundamental ideas in speech recognition surfaced and were published. In the 1960s since computers were still not fast enough, several special purpose hardware were built. However, the decade started with several Japanese laboratories entering the recognition arena and building special purpose hardware as part of their systems. On early Japanese system, described by Suzuki and Nakata of the Radio Research Lab in Tokyo, was a hardware vowel recognizer. An elaborate filter bank spectrum analyzer was used along with logic that connected the outputs of each channel of the spectrum analyzer (in a weighted manner) to a vowel decision circuit, and majority decisions logic

scheme was used to choose the spoken vowel. Another hardware effort in Japan was the work of Sakai and Doshita of Kyoto University in 1962, who built a hardware phoneme recognizer. A hardware speech segmented was used along with a zero crossing analysis of different regions of the spoken input to provide the recognition output. A third Japanese effort was the digit recognizer hardware of Nagata and coworkers at NEC Laboratories in 1963. This effort was perhaps most notable as the initial attempt at speech recognition at NEC and led to a long and highly productive research program. One of the difficult problems of speech recognition exists in the non-uniformity of time scales in speech events. In the 1960s three key research projects were initiated that have had major implications on the research and development of speech recognition for the past 20 years. The first of these projects was the efforts of Martin and his colleagues at RCA Laboratories, beginning in the late 1960s, to develop realistic solutions to the problems associated with non-uniformity of time scales in speech events. Martin developed a set of elementary time normalization methods, based on the ability to reliably detect speech starts and ends,that significantly reduce the variability of the recognition scores. Martin ultimately developed the method and founded one of the first speech recognition companies, Threshold Technology, which was built, marketed and was sold speech recognition products. At about the same time, in the Soviet Union, Vintsyuk proposed the use of dynamic programming methods for time aligning a pair of speech utterances(generally known as Dynamic Time Warping(DTW) ,including algorithms for connected word recognition. Although the essence of the concepts of dynamic time warping, as well as rudimentary versions of the algorithms for connected word recognition, were embodied in Vintsyuk s work, it was largely unknown in the West and did not come to light until the early 1980 s; this was long after the more formal methods were proposed and implemented by others.

C. Speech recognition systems in 1970-1980s

In the 1970s speech recognition research achieved a number of significant milestones. First the area of isolated word or discrete utterance recognition became a viable and usable technology based on fundamental studies by Velichko and Zagoruyko in Russia, Cakoe and Chiba in Japan, and Itakura in the united States. The Russian studies helped the advance use of pattern recognition ideas in speech recognition; the Japanese research showed how dynamic programming methods could be successfully applied; and Itakura s research showed how the ideas of linear predictive coding (LPC), which had already been successfully used in low bit rate speech coding, could be extended to speech recognition systems through the use of an appropriate distance measure based on LPC spectral parameters.Another milestone of the 1970s was the beginning of a longstanding, highly successful group effort in large vocabulary speech recognition at IBM in which researchers studied three distinct tasks over a period of almost two decades, namely the New Raleigh language for simple database queries, the laser patent text language for transcribing laser patents, and the office correspondent tasks called Tangora, for dictation of simple memos. Finally, at AT&T Bell Labs, researchers began a series of experiments aimed at making speech recognition systems that were truly speaker independent. To achieve this goal a wide range of sophisticated clustering algorithms were used to determine the number of distinct patterns required to represent all variations of different words across a wide user population. This research has been refined over a decade so that the techniques for creating speaker independent patterns are now well understood and widely used. An ambitious speech understanding project was funded by the defence Advanced Research Projects Agencies(DARPA), which led to many seminal systems and technology. One of the demonstrations of speech understanding was achieved by CMU in 1973 there Heresay I system was able to use semantic information to significantly reduce the number of alternatives considered by the recognizer. CMU s Harphy system was shown to be able to recognize speech using a vocabulary of 1,011 words with reasonable accuracy. One of the particular contributions from the Harpy system was the concept of graph search, where the speech recognition language is represented as a connected network derived from lexical representations of words, with syntactical production rules and word boundary rules. The Harpy system was the first to take advantage of a finite state network (FSN) to reduce computation and efficiently determine the closest matching strings.

D. Speech recognition systems in 1980-1990s

Just as isolated word recognition was a key focus of research in the 1970s, the problems of word recognition was a focus of research in the 1980s. A wide variety of the algorithm based on matching a concatenated pattern of individual words were formulated and implemented, including the two level dynamic programming approach of Sakoe at Nippon Electric Corporation (NEC), the one pass method of Bridle and Brown at Joint Speech Research Unit(JSRU) in UK, the level building approach of Myers and Rabiner at Bell Labs, and the frame synchronous level building approach of Lee and Rabiner at Bell Labs. Each of these optimal matching procedures had its own implementation advantages, which were exploited for a wide range of tasks. Speech research in the 1980s was characterized by a shift in technology from template based approaches to statistical modeling methods especially the hidden Markov model approach. Although the methodology of hidden Markov modeling (HMM) was well known and understood in a few laboratories(Primarily IBM, Institute for Defense Analyses (IDA), and Dargon systems), it was not until widespread publication of the methods and theory of HMMs, in the mid- 1980, that the technique became widely applied in virtually, every speech recognition research laboratory in the world. Today, most practical speech recognition systems are based on the statistical frame work developed in the 1980s and their results, with significant additional improvements have been made in the 1990s.

E. Speech recognition systems in 1990-2000s

In the 1990s a number of innovations took place in the field of pattern recognition. The problem of pattern recognition, which traditionally followed the framework of Bayes and required estimation of distributions for the data, was transformed into an optimization problem involving minimization of the empirical recognition error. This fundamental paradigmatic change was caused by the recognition of the fact that the distribution functions for the speech signal could not be accurately chosen or defined and the Bayes decision theory becomes inapplicable under these circumstances. Fundamentally, the objective of a recognizer design should be to achieve the least recognition error rather than provide the best fitting of a distribution function to the given (known) data set as advocated by the Bayes criterion. This error minimization concept produced a number of techniques such as discriminative training and kernel based methods. As an example of discriminative training, the Minimum Classification Error(MCE) criterion was proposed along with a corresponding Generalized Probabilistic

Descent(GPD) training algorithm to minimize an objective function which acts to approximate the error rate closely. Another example was the Maximum Mutual Information (MMI) criterion. In MMI training, the mutual information between the acoustic observation and its correct lexical symbol averaged over a training set is maximized. Although this criterion is not based on a direct minimization of the classification error rate and is quite different from the MCE based approach, it is well founded in information theory and possesses good theoretical properties. Both the MMI and MCE can lead to speech recognition performance superior to the maximum likelihood based approach. A key issue in the design and implementation of speech recognition system is how to properly choose the speech material used to train the recognition algorithm. Training may be more formally defined as supervised learning of parameters of primitive speech patterns (templates, statistical models, etc.,) used to characterize basic speech units (e.g. word or subword units), using labeled speech samples in the form of words and sentences. It also discusses two methods for generating training sets. The first uses a nondeterministic statistical method to generate a uniform distribution of sentences from a finite state machine represented in digraph form. The second method, a deterministic heuristic approach, takes into consideration the importance of word ordering to address the problem of co articulation effects that are necessary for good training.

F. Speech recognition systems in 2000-2009s

Around 2000, a variational Bayesian (VB) estimation and clustering techniques were developed. Unlike Maximum Likelihood, this VB approach is based on a posterior distribution of parameters. Giuseppe Richardi have developed the technique to solve the problem of adaptive learning, in automatic speech recognition and also proposed active learning algorithm for ASR. In 2005, some improvements have been worked out on Large Vocabulary Continuous Speech Recognition system on performance improvement. In 2007, the difference in acoustic features between spontaneous and read speech using a large scale speech data base i.e, CSJ have been analyzed. Sadaoki Furui investigated SR methods that can adapt to speech variation using a large number of models trained based on clustering techniques. In 2008, the authors have explored the application of Conditional Random Field(CRF) to combine local posterior estimates provided by multilayer perceptions corresponding to the frame level prediction of phone and phonological attributed classes. Authors proposed an alternative method for processing the Fourier transform phase for extraction speech features, which process the group delay feature(GDF) that can be directly computed for the speech signal.

VII. Applications of Speech Recognition

Various applications of speech recognition domain have been discussed in the following Table 2.

Table 2: Applications of Speech Recognition

Problem Domain	Application	Input pattern	Pattern classes
Speech/Telehphone/ Communication Sector/ Recognition	Telephone directory enquiry without operator Assistance	Speech wave form	Spoken Words
Education Sector	Teaching students of foreign languages to Pronounce vocabulary correctly. Teaching overseas students to pronounce English correctly. Enabling students who are physically handicapped and unable to use a keyboard to enter text verbally Narrative oriented research, where transcripts are automatically generated. This would remove the time to manually generate the transcript, and human error.	Speech wave form	Spoken Words
Outside education sector	Computer and video games, Gambling, Precision surgery.	Speech wave form	Spoken Words
Domestic sector	Oven, refrigerators, dishwashers and washing machines	Speech wave form	Spoken Words
Military sector	High performance fighter aircraft, Helicopters, Battle management, Training air traffic controllers, Telephony and other domains, people with disabilities.	Speech wave form	Spoken Words
Artificial Intelligence sector	Robotics	Speech wave form	Spoken Words
Medical sector	Health care, Medical Transcriptions (digital speech to text)	Speech wave form	Spoken Words
General:	Automated transcription, Telematics, Air traffic control, Multimodal interacting, court reporting, Grocery shops.	Speech wave form	Spoken Words
Physically Handicapped	Useful to the people with limited mobility in their arms and hands or for those with sight	Speech wave form	Spoken Words
Dictation	Dictation systems on the market accepts continuous speech which replaces menu system	Speech wave form	Spoken Words
Translation	It is an advanced application which translates from one language to another.	Speech wave form	Spoken Words

VIII. Conclusion

Speech is the primary and the most convenient means of communication between people. Whether due to technological curiosity to build machines that mimic humans or desire to automate work with machines, research in speech and speaker recognition, as a first step toward natural human-machine communication, has attracted much enthusiasm over the past six decades. Speech Recognition is a challenging and interesting problem in and of itself. I have attempted in this paper to provide a comprehensive cursory, look and review of how much speech recognition technology progressed in the last 60 years. Speech recognition is one of the most integrating areas of machine intelligence, since humans do a daily activity of speech recognition. Speech recognition has attracted scientists as an important discipline and has created a technological impact on society and is expected to flourish further in this area of human machine interaction.

Refrences

[1] Santosh K.Gaikwad, Bharti W.Gawali,"A Review on Speech Recognition Technique", International Journal of Computer Applications, Vol. 10, No. 3, November 2010.

[2] Om Prakash Prabhakar, Navneet Kumar Sahu,"A Survey On: Voice Command Recognition Technique", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 5, May 2013

[3] Sadaoki Furui,"50 years of Progress in speech and Speaker Recognition Research", ECTI Transactions on Computer and Information Technology, Vol. 1, No. 2, November 2005.

[4] B.H. Juang, Lawrence R. Rabiner,"Automatic Speech Recognition – A Brief History of the Technology Development", Georgia Institute of Technology, Atlanta and Rutgers University and the University of California, Santa Barbara.

[5] P.satyanarayana,"Short segment analysis of speech for enhancement", Institute of IIT Madras Feb 2009.

[6] IBM (2010) Online IBM Research Source:-http://www.research.ibm.com/Viewed 12 Jan 2010.

[7] Khaled M. Alhawiti,"Advances in Artificial Intelligence Using Speech Recognition", International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol. 9, No. 6, 2015

[8] John Butzberger, Hy Murveit, Elizabeth Shriberg, Patti Price,"Spontaneous Speech Effects In Large Vocabulary Speech Recognition Applications", SRI International Speech Research and Technology Program Menlo Park.

[9] Amit Verma et al.,"Survey of Distributed Raman Amplification EDFA", National Conference on Recent Trends in Communication and Broadcasting held at SUSCET- TANGORI(MOHALI)-PUNJAB Vol. 1, pp. 272-273, April, 2009. (Sponsored by The Institute of Electronics and Telecommunication Engineering, IETE)

[10] Amit Verma et al.,"Care 99: Analysis and Re-engineering ", National Conference on Emerging Trends in Communication held at SVIET-BANUR(District Patiala, PUNJAB) pp. 72, on 20th -21st February, 2009.