Volume 3, Issue 3, March 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

An Overview of Technical Progress in Speech Recognition

Sanjivani S. Bhabad
Department of E & TC
Pune university, India

Gajanan K. Kharate
Department of E & TC
Pune university, India

Abstract: This paper presents a brief survey on Speech Recognition and discusses the major themes and advances made in the past few 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 (e.g. variations of the context, speakers, and environment).The design of Speech Recognition system requires careful attentions to the following issues: Definition of various types of speech classes, speech representation, feature extraction techniques, speech classifiers, and database and performance evaluation. The problems that are existing in SR and the various techniques to solve these problems constructed by various research workers have been presented in a chronological order. 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.

Keywords: Speech Recognition, Statistical Modelling, Robust speech recognition, Noisy speech recognition, classifiers, feature extraction, performance evaluation, Data base.

I. INTRODUCTION

A. *Defination of Speech Recognition*

Speech Recognition (is also known as Automatic Speech Recognition (ASR) or computer speech recognition) is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program.

1) Basic Model of Speech Recognition: Research in speech processing and communication for the most part, was motivated by people's desire to build mechanical models to emulate human verbal communication capabilities. Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of the signal processing. 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. This paper reviews major highlights during the last few decades in the research and development of 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 a mathematical representation of speech recognition system in simple equations which contain front end unit, model unit, language model unit, and search unit.

The recognition process is shown below (Fig.1).

The standard approach to large vocabulary continuous speech recognition is to assume a simple probabilistic model of speech production whereby a specified word sequence, W, produces an acoustic observation sequence Y, with probability P(W,Y). The goal is then to decode the word string, based on the acoustic observation sequence; so that the decoded string

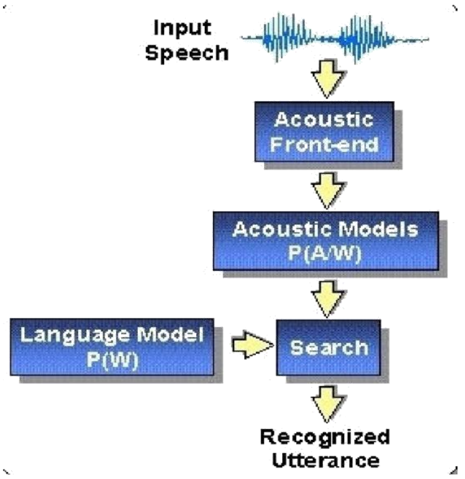

Fig.1 Basic model of speech recognition

has the maximum of equation (1) is posterior (MAP) probability.

$$\hat{} \qquad P(W/A) = \arg \max w_{W} \; P(W/A) \quad ...(1)$$

Using Bay's rule, equation 1can be written as

$$P(W/A) = \frac{P(A/W)\,P(W)}{P(A)} \qquad ...(2)$$

Since P(A) is independent of W, the MAP decoding rule of equation(1) is

$$\hat{W} = \arg \max_{w} \hat{P}(A/W)P(W) \quad ...(3)$$

The first term in equation (3) P(A/W), is generally called the acoustic model, as it estimates the probability of a sequence of acoustic observations, conditioned on the word string. Hence P(A/W) is computed. For large vocabulary speech recognition systems, it is necessary to build statistical models for sub word speech units, build up word models from these sub word speech units, (using a lexicon to describe the composition of words), and then postulate word sequences and evaluate the acoustic model probabilities via standard concatenation methods. The second term in equation (3) P(W), is called the language model. It describes the probability associated with a postulated sequence of words. Such language models can incorporate both syntactic and semantic constraints of the language and the recognition task.

2) Types of Speech Recognition:

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

.

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.

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.

3) 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 figure 2.

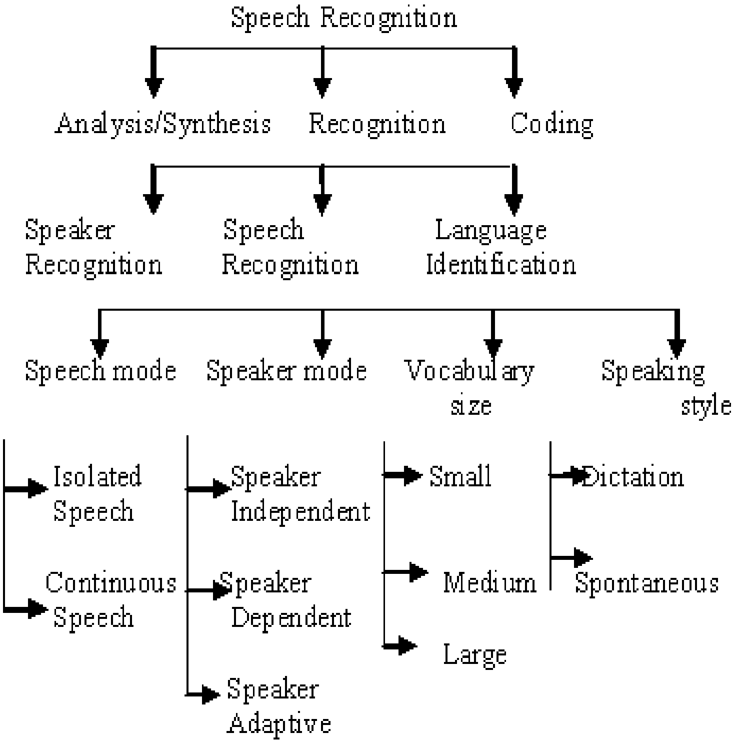

Fig. 2 Speech Processing Classification

4) 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
Vocabulary	Characteristics of available training data; specific or generic vocabulary;

Table 2 Speech Recognition Techniques

Techniques	Representation	Recognition Function
Acoustic Phonetic Approach	Spectral analysis with feature detection Phonemes/ segmentation and labelling	Probabilistic lexical access procedure
Pattern Recognition approach • Template • DTW • VQ	Speech, samples, pixels and curves Set of sequence of spectral vectors Set of spectral vectors Features	Correlation distance measure Dynamic warping Optimal algorithm Clustering function
Neural Network	Speech features/ perceptrons/ Rules/ Units/Procedures	Network function
Support Vector Machine	Kernel based features	Maximal margin hyperplane,Radial basis
Artificial intelligence approach	Knowledge based	

II. Approaches To Speech Recognition

Basically there exist three approaches to speech recognition. They are

- Acoustic Phonetic Approach
- Pattern Recognition Approach
- Artificial Intelligence Approach

A. Acoustic Phonetic Approach

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, 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 neighbouring sounds, it is assumed in the acoustic-phonetic approach that the rules governing the variability are straightforward and can be readily learned by a machine. The first step in the acoustic phonetic approach is a spectral analysis of the speech combined with a feature detection that converts the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. The next step is a segmentation and labelling phase in which the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech. The last step in this approach attempts to determine a valid word (or string of words) from the phonetic label sequences produced by the segmentation to labelling. In the validation process, linguistic constraints on the task (i.e., the vocabulary, the syntax, and other semantic rules) are invoked in order to access the lexicon for word decoding based on the phoneme lattice. The acoustic phonetic approach has not been widely used in most commercial applications ([1]).The following table 2 broadly gives the different speech recognition techniques.

B. 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 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 ([1] Refer fig.2). In this, there exists two methods namely template approach and stochastic approach.

1) Template Based Approach: Template based approach [2] to speech recognition have provided a family of techniques that have advanced the field considerably during the last decades. A collection of prototypical speech patterns are stored as reference patterns representing the dictionary of candidate's words. Recognition is then carried out by matching an unknown spoken utterance with each of these references templates and selecting the category of the best matching

pattern. Each word must have its own full reference template; template preparation and matching become prohibitively expensive or impractical as vocabulary size increases beyond a few hundred words. One key idea in template method is to derive typical sequences of speech frames for a pattern (a word) via some averaging procedure, and to rely on the use of local spectral distance measures to compare patterns. Another key idea is to use some form of dynamic programming to temporarily align patterns to account for differences in speaking rates across talkers as well as across repetitions of the word by the same talker.

2) Stochastic Approach: Stochastic modelling [2] entails the use of probabilistic models to deal with uncertain or incomplete information. In speech recognition, uncertainty and incompleteness arise from many sources; for example, confusable sounds, speaker variability s, contextual effects, and homophones words. Thus, stochastic models are particularly suitable approach to speech recognition. The most popular stochastic approach today is hidden Markov modeling. A hidden Markov model is characterized by a finite state markov model and a set of output distributions. The transition parameters in the Markov chain models, temporal variabilities, while the parameters in the output distribution model, spectral variabilities. These two types of variabilites are the essence of speech recognition.

C. Dynamic Time Warping (DTW)

Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video, the person was walking slowly and if in another, he or she were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW has been applied to video, audio, and graphics indeed, any data which can be turned into a linear representation can be analyzed with DTW. A well known application has been automatic speech recognition, to cope with different speaking speeds. In general, DTW is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in the context of hidden Markov models.

D. Vector Quantization (VQ)

Vector Quantization (VQ) [2] 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. The test speech is evaluated by all codebooks and ASR chooses the word whose codebook yields the lowest distance measure. In basic VQ, codebooks have no explicit time information, since codebook entries are not ordered and can come from any part of the training words. However, some indirect durational cues are preserved because the codebook entries are chosen to minimize average distance across all training frames, and frames, corresponding to longer acoustic segments (e.g., vowels) are more frequent in the training data. Such segments are thus more likely to specify code words than less frequent consonant frames, especially with small codebooks. Code words nonetheless exist for constant frames because such frames would otherwise contribute large frame distances to the codebook. Often a few code words suffice to represent many frames during relatively steady sections of vowels, thus allowing more codeword to represent short, dynamic portions of the words. This relative emphasis that VQ puts on speech transients can be an advantage over other ASR comparison methods for vocabularies of similar words.

E. Artificial Intelligence Approach (Knowledge Based Approach)

The Artificial Intelligence approach [2] is a hybrid of the acoustic phonetic approach and pattern recognition approach. In this, it 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. On the other hand, a large body of linguistic and phonetic literature provided insights and understanding to human speech processing. 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 modelling. 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.

F. Connectionist Approaches (Artificial Neural Networks)

The artificial intelligence approach ([2], Lesser et al. 1975; Lippmann 1987) attempts to mechanize the recognition procedure according to the way a person applies intelligence in visualizing, analysing, and characterizing speech based on a set of measured acoustic features. Among the techniques used within this class of methods are uses of an expert system (e.g., a neural network) that integrates phonemic, lexical, syntactic, semantic, and even pragmatic knowledge for segmentation and labelling, and uses tools such as artificial NEURAL NETWORKS for learning the relationships among phonetic events. The focus in this approach has been mostly in the representation of knowledge and integration of knowledge sources. This method has not been widely used in commercial systems. Connectionist modelling of speech is the youngest development in speech recognition and still the subject of much controversy. In connectionist models, knowledge or constraints are not encoded in individual units, rules, or procedures, but distributed across many simple computing units. Uncertainty is modelled not as likelihoods or probability density functions of a single unit, but by the pattern of activity in many units. The computing units are simple in nature, and knowledge is not programmed into any individual unit function; rather, it lies in the connections and interactions between linked processing elements. Because the style of computation that can be performed by networks of such units bears some resemblance to the style of computation in the nervous system. Connectionist models are also referred to as neural networks or artificial neural networks. Similarly, parallel distributed processing or massively distributed processing are terms used to describe these models.

Not unlike stochastic models, connectionist models rely_critically on the availability of good training or learning strategies. Connectionist learning seeks to optimize or organize a network of processing elements. However, connectionist models need not make assumptions about the underlying probability distributions. Multilayer neural networks can be trained to generate rather complex nonlinear classifiers or mapping function. The simplicity and uniformity of the underlying processing element makes connectionist models attractive for hardware implementation, which enables the operation of a net to be simulated efficiently. On the other hand, training often requires much iteration over large amounts of training data, and can, in some cases, be prohibitively expensive. While connectionism appears to hold great promise as plausible model of cognition, may question relating to the concrete realization of practical connectionist recognition techniques, still remain to be resolved.

G. Support Vector Machine (SVM)

One of the powerful tools for pattern recognition that uses a discriminative approach is a SVM [2]. SVMs use linear and nonlinear separating hyper-planes for data classification. However, since SVMs can only classify fixed length data vectors, this method cannot be readily applied to task involving variable length data classification. The variable length data has to be transformed to fixed length vectors before SVMs can be used. It is a generalized linear classifier with maximum-margin fitting functions. This fitting function provides regularization which helps the classifier generalized better. The classifier tends to ignore many of the features. Conventional statistical and Neural Network methods control model complexity by using a small number of features (the problem dimensionality or the number of hidden units). SVM controls the model complexity by controlling the VC dimensions of its model. This method is independent of dimensionality and can utilize spaces of very large dimensions spaces, which permits a construction of very large number of non-linear features and then performing adaptive feature selection during training. By shifting all non-linearity to the features, SVM can use linear model for which VC dimensions is known. For example, a support vector machine can be used as a regularized radial basis function classifier.

H. Taxonomy Of Speech Recognition

Existing techniques for speech recognition have been represented diagrammatically in the following figure 3.

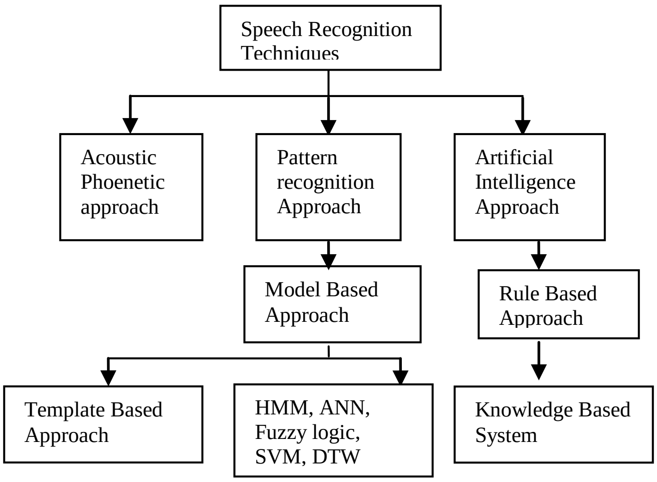
Speech Recognition Techniques
Acoustic Phoenetic approach
Pattern recognition Approach
Artificial Intelligence Approach
Model Based Approach
Rule Based Approach
Template Based Approach
HMM, ANN, Fuzzy logic, SVM, DTW
Knowledge Based System

Fig 3. Taxonomy Of Speech Recognition

III. Feature Extraction

In speech recognition, the main goal of the feature extraction step is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal. The feature extraction is usually performed in three stages. The first stage is called the speech analysis or the acoustic front end. It performs some kind of spectrotemporal analysis of the signal and generates raw features describing the envelope of the power spectrum of short speech intervals. The second stage compiles an extended feature vector composed of static and dynamic features. Finally, the last stage (which is not always present) transforms these extended feature vectors into more compact and robust vectors that are then supplied to the recognizer. Although there is no real consensus as to what the optimal feature sets should look like, one usually would like them to have the following properties: they should allow an automatic system to discriminate between different through similar sounding speech sounds, they should allow for the automatic creation of acoustic models for these sounds without the need for an excessive amount of training data, and they should exhibit statistics which are largely invariant across speakers and speaking environment.

3.1. FEATURE EXTRACTION METHODS

Various methods for Feature Extraction in speech recognition are broadly shown in the following table 3.

Table 3: feature extraction methods

Method	Property	Comments
Principal Component Analysis(PCA)	Non linear feature extraction method, Linear map; fast; eigenvector-based	Traditional, eigenvector based method, also known as karhuneu-Loeve
Linear Discriminant Analysis(LDA)	Non linear feature extraction method, Supervised linear map; fast; eigenvector-based	Better than PCA for classification;
Independent Component Analysis (ICA)	Non linear feature extraction method, Linear map, iterative non- Gaussian	Blind course separation, used for de-mixing non- Gaussian distributed sources(features)
Linear Predictive coding	Static feature extraction method,10 to 16	
	lower order co- efficient,	
Cepstral Analysis	Static feature extraction method, Power spectrum	Used to represent spectral envelope
Mel-frequency scale analysis	Static feature extraction method, Spectral analysis	Spectral analysis is done with a fixed resolution along a subjective frequency scale i.e. Mel-frequency scale.
Filter bank analysis	Filters tuned required frequencies	
Mel-frequency cepstrum (MFFCs)	Power spectrum is computed by performing Fourier Analysis	
Kernel based feature extraction based	Non linear	Dimensionality reduction leads to better classification and it is used to remove noisy and redundant features, and improvement in classification error

Wavelet	Better time resolution than Fourier Transform	It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allow better time resolution at high frequencies than Fourier Transform
Dynamic feature extractions i)LPC ii)MFCCs	Acceleration and delta coefficients i.e. II and III order derivatives of normal LPC and MFCCs coefficients	
Spectral subtraction	Robust Feature extraction method	

IV. Classifiers

In speech recognition a supervised pattern classification system is trained with labeled examples; that is, each input pattern has a class label associated with it. Pattern classifiers can also be trained in an unsupervised fashion. Once a feature selection or classification procedure finds a proper representation, a classifier can be designed using a number of possible approaches. In practice, the choice of a classifier is a difficult problem and it is often based on which classifier(s) happen to be available, or best known, to the user.

The three different approaches are identified to design a classifier. The simplest and the most intuitive approach to classifier design is based on the concept of similarity: patterns that are similar should be assigned to the same class. So, once a good metric has been established to define similarity, patterns can be classified by template matching or the minimum distance classifier using a few prototypes per class.

V. Performance Of Speech Recognition Systems

The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy may be measured in terms of performance accuracy which is usually rated with word error rate (WER), whereas speed is measured with the real time factor. Other measures of accuracy include Single Word Error Rate (SWER) and Command Success Rate (CSR).

Word Error Rate (WER): Word error rate is a common metric of the performance of a speech recognition or machine translation system. The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence (supposedly the correct one). The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level. This problem is solved by first aligning the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment.

Word error rate can then be computed as:

$$WER= \frac{S+D+I}{N} \qquad .....(4)$$

- *S* is the number of substitutions,
- *D* is the number of the deletions,

I is the number of the insertions,

- *N* is the number of words in the reference.

When reporting the performance of a speech recognition system, sometimes word recognition rate (WRR) is used instead:

$$WER= 1- WER = \frac{N- S- D- I}{N} = \frac{H-I}{N} \qquad .....(5)$$

Where *is*, N-(S+D), the number of correctly recognized words.

VI. Speech Databases

Speech databases have a wider use in 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.

Table 4: represents the characteristics of main databases used in speech recognition

Name	**No. Speakers**	**No. units**	**Speech style**	**Recording Environment**	**SR kHz**	**Transcription based on**
TI Digits	326	>2500 numbers	reading	QR	20	Word
TIMIT	630	6300 sentences	reading	QR	16	Phones
NTIMIT	630	6300 sentence	reading	TEN	8	Phones
RM1	144	15024	reading	QR	20	Sentence

		sentences				
RM2	4	10608 sentences	reading	QR	20	Sentence
ATIS0	36	10722 utterances	Reading spont.	OFC	16	Sentence
Switch Board (Credit card)	69	35 dialogues	Conv. spont	TEL	8	Word
TI-46	16	19,136 isol. Words	Reading	QR	16	Sentence
Switch Board (Complete)	550	2550 dialogues	Conv. spont	TEL	8	Word
ATC	100	30000 dialogues	Spont	RF	8	Sentence
ATIS2	351	12000 utterances	Spont	OFC	16	Sentence

Main database characteristics:
Abbreviations:
QR: Quiet Room Ofc: Office RF: Radio Frequency

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 multi-session. 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.

VII. Summary Of The Technology Progress

In the last few years, especially in the last three decades, research in speech recognition has been intensively carried out worldwide, spurred on by advances in signal processing algorithms, architectures and hardware. The technological progress in the few years can be summarized in the table 5[12].

Table 5: Summary of the technological progress in the last few years

Sr.no	Past	Present
1	Template matching	Corpus -based modelling e.g HMM and n-grams
2	Filter bank/spectral resonance	Cepstral features, Kernel based function, group delay functions
3	Heuristic time normalization	DTW/DP matching
4	Distance –based methods	Likelihood based methods
5	Maximum likelihood approach	Discriminative approach e.g .MCE/GPD and MMI
6	Isolated word recognition	Continuous speech recognition
7	Small vocabulary	Large vocabulary
8	Context Independent units	Context dependent units
9	Clean speech recognition	Noisy/telephone speech recognition
10	Single speaker recognition	Speaker-independent/adaptive recognition
11	Monologue recognition	Dialogue/Conversation recognition
12	Read speech recognition	Spontaneous speech recognition

13	Single modality(audio signal only)	Multimodal (audio/visual) speech recognition
14	Hardware recognizer	Software recognizer
15	Speech signal is assumed as quasi-stationary in the traditional approaches. The feature vectors are extracted using FFT and wavelet methods etc.	Data driven approach does not possess this assumption i.e. signal is treated as nonlinear and non stationary. In this features are extracted using Hilbert Haung Transform using IMFs.[141]

VIII. Gap Between Machine And Human Speech Recognition

What we know about human speech processing is still very limited and we have yet to witness a complete and worthwhile Unification of the science and technology of speech. In 1994, Moore [13] presented the following 20 themes which is believed to be an important to the greater understanding of the nature of speech and mechanisms of speech pattern processing in general:

- How important is the communicative nature of speech?
- Is human-human speech communication relevant to human machine communication by speech?
- Speech technology or speech science? (How can we integrate speech science and technology)
- Whither a unified theory?
- Is speech special?
- Why is speech contrastive?
- Is there random variability in speech?
- How important is individuality?
- How much effort does speech need?
- What is a good architecture (for speech processes)?
- What are suitable levels of representation?
- What are the units?
- What is the formalism?
- How important are the physiological mechanisms?
- Is time-frame based speech analysis sufficient?
- How important is adaptation?
- What are the mechanisms for learning?
- What is speech good for?
- How good is speech?

After more than 10 years, we still do not have clear answers to these 20 questions.

IX . 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 five decades. We have also encountered a number of practical limitations which hinder a widespread deployment of application and services. In most speech recognition tasks, human subjects produce one to two orders of magnitude less errors than machines. There is now increasing interest in finding ways to bridge such a performance gap. What we know about human speech processing is very limited. Although these areas of investigations are important the significant advances will come from studies in acoustic-phonetics, speech perception, linguistics, and psychoacoustics. Future systems need to have an efficient way of representing, storing, and retrieving knowledge required for natural conversation. Although significant progress has been made in the last two decades, there is still work to be done, and we believe that a robust speech recognition system should be effective under full variation in: environmental conditions, speaker variability s etc. Speech Recognition is a challenging and interesting problem in and of itself. We have attempted in this paper to provide a comprehensive cursory, look and review of how much speech recognition technology progressed in the last few 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. We hope this paper brings about understanding and inspiration amongst the research communities of ASR.

X. Future Scope

With the help of above review we can design and implement speech recognition and rectification system for articulatory handicapped people which will be a nobel work for society. And hence we can reduce the speech communication problems faced by articulatory handicapped people in their day to day life.

References

[1] Dat Tat Tran, Fuzzy Approaches to Speech and Speaker Recognition , A thesis submitted for the degree of Doctor of Philosophy of the university of Canberra.

[2]]R.K.Moore,Twenty things we still don t know about speech , Proc.CRIM/ FORWISS Workshop on Progress and Prospects of speech Research an Technology , 1994.

[3] Anil K.Jain, et.al., Statistical Pattern Recognition: A Review , IEEE Transactions on Pattern Analysis and Machine Intelligence ,Vol.22, No.1, January 2000.

[4] L.R.Rabiner, A tutorial on hidden Markov models an selected applications in speech recognition, Proc.IEEE 77(2):257-286.February 1989.

[5] K.L.Oehler and R.M.Gray, Combining Image compression and Classification Using vector quantization, IEEE Trans. Pattern Analyssis and achine Intelligence,Vol.17, no.5,pp.461-473,1995.

[6] Q.B.Xie, C.A.Laszlo, and R.K.Ward, Vector Quantisation Technique for Nonparametric Classifier Design , IEEE Trans. Pattern Analysis and Machine Intelligence, vol.15, no.12, pp.1,326-1,330,1993.

[7] T.Kohonen, Self-Organizing Maps. Springer Series in Information Sciences, vol.30,Berlin, 1995.

[8] P.A.Devijver and J.Kittler, Pattern Recognition:A Statistical Approach , London, Prentice Hall, 1982.

[9] E.Oja, Subspace Methods of Pattern Recognition, Letchworth, HeHertfordshire,England:Research Studies Press, 1983.

[10] K.Fukunaga, Introduction to Statistical Pattern Recognition , second, Newyork:Academic Press, 1990.

[11] J.H.Friedman, Exploratory Projection Pursuit, J.Am.Statistical Assoc.,Vol.84,pp.165-175,1989.

[12] .Angel de la Torre , Histogram Equalization of Speech Representation for Robust Speech Recognition , IEEE Transactions On Speech And Audio Processing, Vol.13, No. 3, May 2005.

[13] Yang Liu et.al,. Enriching Speech Recognition with Automatic Detection of sentence Boundaries an disfluencies , IEEE Transactions on Audio,Speech and Language processing, V.14,No.4,July 2006.

[14] K.H.Davis, R.Biddulph, and S.Balashek, Automatic Recognition of spoken Digits, J.Acoust.Soc.Am., 24(6):637-642,1952.

[15] H.F.Olson and H.Belar, Phonetic Typewriter , J.Acoust.Soc.Am.,28(6):1072-1081,1956.

[16] D.B.Fry, Theoritical Aspects of Mechanical speech Recognition , and P.Denes, The design and Operation of the Mechanical Speech Recognizer at Universtiy College London, J.British Inst. Radio Engr., 19:4,211-299,1959.

[17] J.W.Forgie and C.D.Forgie, Results obtained from a vowel recognition computer program, J.A.S.A.,31(11), pp. 1480-1489.1959.

[18] J.Suzuki and K.Nakata, Recognition of Japanese Vowels Preliminary to the Recognition of Speech , J.Radio Res.Lab37(8):193-212,1961.Preliminary to the Recognition of Speech , J.Radio Res.Lab37(8):193-212,1961.

[19] T.Sakai and S.Doshita, The phonetic typewriter, information processing 1962 , Proc.IFIP Congress, 1962.

[20] K.Nagata, Y.Kato, and S.Chiba, Spoken Digit Recognizer for Japanese Language , NEC Res.Develop., No.6,1963.

[21] T.B.Martin, A.L.Nelson, and H.J.Zadell, Speech Recognition b Feature Abstraction Techniques , Tech.Report AL-TDR-64-176,Air Force Avionics Lab,1964.

[22] T.K.Vintsyuk, Speech Discrimination by Dynamic Programming , Kibernetika, 4(2):81-88,Jan.- Feb.1968.

[23] H.Sakoe and S.Chiba, Dynamic programming algorithm optimization for spoken word recognition ,IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-26(1).pp.43- 49,1978.

[24] D.R.Reddy, An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave , Tech.Report No.C549, Computer Science Dept., Stanford Univ., September 1966.

[25] V.M.Velichko and N.G.Zagoruyko, Automatic Recognition of 200 words , Int.J.Man-Machine Studies,2:223,June 1970.

[26] H.Sakoe and S.Chiba, Dynamic Programming Algorithm Optimization for Spoken Word Recognition ,IEEE Trans.Acoustics, Speech, Signal Proc.,ASSP-26(1):43- 49,February 1978.

[27] F.Itakura, Minimum Prediction Residula Applied to Speech Recognition ,IEEE Trans.Acoustics, Speech,Signal Proc., ASSP-23(1):67-72,February 1975.

[28] C.C.Tappert,N.R.Dixon, A.S.Rabinowitz, and W.D.Chapman, Automatic Recognition of Continuous Speech Utilizing Dynamic Segmentation, Dual Classification, Sequential Decoding and Error Recove,Rome Air Dev.Cen, Rome, NY,Tech.Report TR-71-146,1971.

[29] F.Jelinek, L.R.Bahl, and R.L.Mercer, Design of a Lingusistic Statistical Decoder for the Recognition of Continuous Speech , IEEE Trans.Information Theory,IT- 21:250-256,1975.

[30] F.Jelinek, The Development of an Experimental Discrete Dictation Recognizer , Proc.IEEE,73(11):1616-624,1985.