Speech Recognition Techniques: A Review

Praphulla A. Sawakare, Ratndeep R. Deshmukh, Pukhraj P. Shrishrimal

Abstract— Speech is the natural and the fundamental way of communication for most humans. Human beings are developing systems which can understand, interpret and accept the command via speech signals. Speech recognition is the process of converting an acoustic waveform into the text similar to the information being conveyed by the speaker .This paper present the basic idea of speech recognition system for fundamental progress of speech recognition and also gives overview technique developed in each stage of speech recognition.

Index Terms— Analysis, Artificial Neural Networks ,Aumatic speech recognition (ASR), Feature Extraction, Feature Matching, hmm,Mel Frequency Cepstral Coefficient (MFCC), Modeling, SVM

1 INTRODUCTION

Speech is the most primary,efficient and widely used mode of communication between humans beings. There are large number of different spoken languages which are used throughout the world. The communication among the human is mostly done by vocally for information exchange[1]. Thus it is natural for people to hope speech interfaces with computer. For a real-time a intelligent applications, it is essential that the machine can hear, interpret , analysis and act upon input information from speaker, and also give immediate response to complete the information transfer .This can be carried out by developing an Automatic Speech Recognition (ASR) system which is a procedure of translating an acoustic signal into a written text or a command without understanding what has been recognized[2].

Since, 1960s many computer scientists have been researching on different means to make computer record, interpret, analys and understand human speech. Research on automatic speech recognition by machine has attracted much attention over the last five decades. The survey shows that the agencies like AT & T Bell Labs, DARPA, IBM, and Mi-crosoft have sponsored number of programs for research in this area in the last 50 years [3]. Still a lot of research work is being done in this area. Although ASR is still lagging far behind commercial speech recognition systems are being used in many applications, like , auto-attendants, virtual reality, electronic devices, dictation, controlling the various programs, automatic telephone call processing system and query based information system such as travel information system , weather forecast information system etc. Though Speech recognition systems has gained new heights and era but robustness and noise tolerant recognition systems are few of the problems which make them difficult to handle. Many researcher around the world are trying to develop a robust and noise tolerant speech recognition systems.

2 CLASSIFICATIONS OF SPEECH RECOGNITION SYSTEMS:

The speech recognition systems can be divided into different type depending upon type of utterances, vocabulary size and speaker dependency.

2.1 Classification on the basis of utterances:

2.1.1 Isolated Words:
Isolated word recognition is to recognize an isolated speech signal as a single word, where the signal is well segmented and the output is unique .The recognition is mainly based on the acoustic models of all possible hypotheses, with little consideration, if any, for language models. Isolated word recognizers usually need each of the utterance to have quiet on both sides of the sample window. It mean that it require a single utterance at a time. These systems generally have "Listen/Not-Listen" states, because they require the speaker to wait between utterances [4].

2.1. 2 *Connected Words:*
Connect word systems are similar to isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them.

2.1.3 Continuous Speech:
Continuous speech recognition deals with a signal without knowing the number of contained word units and the segmentations; it needs to recognize the most likely string of units and determine the boundaries of the recognized units simultaneously.

2.1.4 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

2.2 Classification on the basis of Vocabulary size:

The complexity, processing requirements and the accuracy of the system depends upon the vocabulary size of a speech recognition system. With increase in the size of vocabulary, the taskof recognition become difficult is.Thus ASR system are classified based on the vocabulary as follow

i. Small vocabulary - 10 words
ii. Medium vocabulary – 100 words
iii. Large vocabulary – 1000 words
iv. Very-large vocabulary - 10000 words
v. Out-of-Vocabulary- Mapping a word from the vocabulary into the unknown word.

2.3 Classification on the basis of Speaker mode:

2.3.1 Speaker Independent:
In thespeaker-independent speech recognition systems there is no need of training of the system to recognize a appropriate

speaker and thus the stored word patterns must be representative of the collection of speakers expected to be used the system. The word templates are derived from a large number of sample patterns which are a cross-section of talkers of different , age-group and dialect,ascent ,sex .After this a clustering is used to form a representative pattern for each word[5].

2.3.2 Speaker Dependent:

Speaker adaptive speech recognition systems uses the speaker dependent data and adapt to the best suitable speaker to recognize the speech and decrease the error rate by adaption.

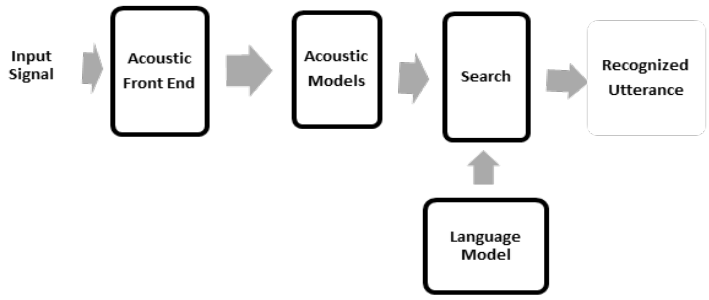

Fig 1: System Architecture of for Automatic Speech Recognition System [6].

3 SPEECH RECOGNITION TECHNIQUES:

The speaker recognisation system may be viewedas working in a four stages

1. Analysis
2. Feature extraction
3. Modeling
4. Testing

3.1 Speech analysis:

Speech data include different type of information that shows a speaker identity. This speaker specific information is obtained from the vocal tract, excitation source and behavior feature. This information is embedded in signal and can be used for speaker recognition process. The speech analysis stage compromise with stage with suitable frame size for segmenting speech signal for further analysis and extracting[7].The speech analysis technique done with following three techniques.

3.1.1 Speech analysis technique:

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

3.1.2 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.

3.1.3 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.

3.1.4 Performance of System:

The performance of speaker recognisation system depends on thetechnique employed in the various stages of speaker recognisation system. The state of art of speaker recognisation system mainly used segmental analysis, Mel frequency Spectral coefficients (MFFCs), Gaussian mixture model (GMM) and feature extraction, modeling and testing stage. There are practical issues in the speaker recognisation field other technique may also have to be used for resulting a good speaker recognisation performance some of practical issues are as follows…

- Nonacoustic sensor provide an exciting opportunity formultimodal speech processing with application to areas such as speech enhancement and coding .this sensor provide measurement of function of the glottal excitation and can supplement acoustic waveform.
- A universal background Model (UBM) is a model usd in a speaker verification system to represent general person independent the feature characteristics to be compared against a model of person specific feature characteristics when making an accept or reject decision.
- A Multimodel person recognisation architecture has been developed for the purpose of improving overall recognition performance and for addressing channel specific performance. This multimodal architecture includes the fusion of speech recognisation system with the MIT/LL GMM/UBM speaker recognition architecture.
- Many powerful for speaker recognition have introduced in high level features, novel classifiers and channel compression methods.
- SVMs have become a popular and powerful tool in text independent speaker verification at the core of any SVM type system give a choice of feature expansion [8].
- A recent areas of significant progress in speaker recognition is the use of high level features-idiolect, phonetic relations, prosody. A speaker not only has distinctive acoustic sound but uses language in a characteristic manner.

3.2 Feature Extraction Technique:

In speech recognition system, Feature Extraction plays an important role to separate one speech from other. The main focus of feature extractor is to keep the relevant information and discard irrelevant one from the speech.In fundamental formation of speaker identification and verification system, that the number of training and test vector needed for the classification problem grows with the dimension of the given input so we need feature extraction of speech signal. But extracted feature should fulfill some criteria while dealing with the speechsignal such as easy to measure, should not be susceptible to mimicry, should show little fluctuation from one speaking environment to another, should be stable over time, should occur frequently and naturally in speech.

TABLE 1
DIFFERENT FEATURE EXTRACTION METHODS USED IN SPEECH RECOGNITION SYSTEM

Sr. No.	Methods	Property	Comments
1	Principal Component analysis (PCA)	Non linear feature extraction method, Linear map, fast, eigenvector-based	Traditional, eigenvector base method, also known as karhuneu-Loeve expansion; good for Gaussian data
2	Linear Discriminate Analysis(LDA)	Non linear feature extraction method, Supervised linear map; fast, eigenvector-based	Better than PCA for classification[9]
3	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)
4	Linear Predictive coding	Static feature extraction method,10 to 16 lower order coefficient,	It is used for feature Extraction at lower order
5	Cepstral Analysis	Static feature extraction method, Power spectrum	Used to represent spectral envelope
6	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
7	Filter bank analysis	Filters tuned required frequencies	
8	Mel-frequency cepstrum (MFFCs)	Power spectrum is computed by performing Fourier Analysis	This method is used for find our features
9	Kernel based feature extraction method	Non linear transformations	Dimensionality reduction leads to better classification and it is used to redundant features, and improvement in classification error.
10	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
11	Dynamic feature extractions i)LPC ii)MFCCs	Acceleration and delta coefficients i.e. II and III order derivatives of normal LPC and FCCs coefficients	It is used by dynamic or runtime Feature
12	Spectral subtraction	Robust Feature extraction method	It is used basis on Spectrogram
13	Cepstral mean subtraction	Robust Feature extraction	It is same as MFCC but working on Mean statically parameter
14	RASTA filtering	For Noisy speech	It is find out Feature in Noisy data
15	Integrated Phoneme subspace method (Compound Method)	A transformation based on PCA+LDA+ICA	Higher Accuracy than the existing Methods

3.3 Modeling Technique:

The basic objective of modeling technique is to generate

speaker models using speaker specific feature vector. The speaker modeling technique divided as: speaker recognition and speaker identification. It automatically identify who is speaking on the basis of individual information integrated in speech signal .The speaker recognition is also divided into two parts that means speaker dependant and speaker independent. In Speaker independent approch of the speech recognition the computer should ignore the speaker specific characteristics of the speech signal and extract the expected message. Beside this, in speaker dependent, recognisizing 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 textindependent methods. In text-dependent method, the speaker says key words or sentences having the same text for both training and recognition trials, whereas text independent does not dependent on a specific texts being spoken. Following are approaches to speech recognition.

3.3.1 The 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 phoneticapproach, 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. Although, the acoustic properties of phonetic units are highly dangling, both with speakers and with neighboring 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 [10]. 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. Segmentation and labeling phase done in next step. In this the speech signal is segmented into stable acoustic regions which followed by linking one or more phonetic labels to each segmented region. This results in a phoneme lattice characterization of the speech. In the last step it attempts to determine a valid word from the phonetic label sequences produced by the segmentation to labeling.Llinguistic constraints on the task are preempted in order to access the lexicon for word decoding based on the phoneme lattice. This is called as validation process[11]. The acoustic phonetic approach has not been widely used in most commercial applications.

3.3.2 Pattern Recognition approach:

There are two essential steps involved in pattern recognition approach, pattern training and pattern comparison. Using a well formulated mathematical framework and initiates consistent speech pattern representation for reliable pattern comparison. A set of labeled training samples through formal training algorithm is essential feature of this approach [12]. In this, there exist two methods: Template base approach and stochastic approach. It is more suitable approach to speech recognition as it uses probabilistic models to deal with undetermined or incomplete information.There exits many methods in this approach like HMM, SVM, DTW, VQ etc, among these hidden markov model is most popular stochastic approach today.

A. Template- Based Approach

In template based approaches matching, unknown speech is compared against a set of pre-recorded words in order to find the best match which is adavanteageous to find accurate word models. But it also has the disadvantage that pre-recorded templates are fixed. Due to this variations in speech can only be modelled by using many templates per word, which eventually becomes impractical [13]. In this approach, the templates usually consists of representative sequences of features vectors for corresponding words. The fundamental aim here is to align the utterance to each of the template words and then select the word or word sequence that contains the perfect match. For each utterance, the distance between the template and the observed feature vectors are computed using various distance measure and these local distances are compile along each possible alignment path. Then the lowest scoring path identifies the optimal alignment for a word and the word template obtaining the lowest overall score interpret the recognised word or sequence of words [14].

B. Statistical - Based Approach

In Statistical based approaches deviation in speech are modelled statistically. It uses automatic statistical learning procedure, typically the HMM. The approach represents the current state of the art. The main drawback of statistical models is that they must take priori modelling assumptions which are liable to be misleading, handicapping the system performance. In recent years, A new technique has been emerged which challenges the problem of conversational speech recognition has emerged. It is anticipated to overcome some fundamental limitations of the conventional Hidden Markov Model (HMM) approach [15]. This new approach is a radical withdrawl from the current HMM-based statistical modeling approaches. Instead of using a large number of unstructured Gaussian mixture components to account for the huge variation in the observable acoustic data of highly coarticulated natural speech. The new speech model which has been developed to provide a rich structure for the partially observed dynamics in the domain of vocal-tract resonances [16].

3.3.3 The Artificial Intelligence Approach (Knowledge Based Approach):

It is combination of acoustic phonetic approach and pattern recognition approach. The artificial intelligence approach attempts to mechanize the recognition procedure according to the way a person applies its brilliance in visualizing, analyzing. And then finally making a decision on the measured acoustic features. The artificial intelligence approach attempts to mechanise the recognition procedure according to the way a person applies its intelligence in visualizing, analysing, and finally making a decision on the measured acoustic features [17]. Highly master system are used widely in this approach. In Knowledge based approaches: An expert knowledge about variations in speech is hand coded into a system. It is benoficial of explicit modelling variations in speech; but unfortunately such expert knowledge is difficult to obtain and use successfully [18]. Thus this approach was judged to be imprac-

tical and automatic learning procedure was sought instead.

3.3.4 Connectionist Approaches (Artificial Neural Networks):

The artificial intelligence approach attempts to mechanise the The artificial intelligence approach attempts to mechanise the recognition procedure according to the way a person applies its intelligence in visualizing, analysing, and final-ly making a decision on the measured acoustic features [19]. Among the techniques used within this class of methods are uses of an expert system that integrates lexi-cal, phonemic, semantic, syntactic and pragmatic knowledge for segmentation and labeling. It uses tools such as artificial NEURAL NETWORKS for learning the relationships among phonetic events. It mainly focuses on the representation of knowledge and integration of knowledge sources [20]. This method has not been widely used in commercial systems.

3.3.5 Vector Machine (SVM):

One of the powerful tools for pattern recognition is SVM which uses a discriminative approach is a SVM. It uses linear and nonlinear separating hyper-planes for data classification. In spite of this, SVMs can only classify fixed length data vectors. This procedure may not be readily applied to task involving variable length data classification. This data has to be trans-formed to fixed length vectors before SVMs can be used [21]. It is a generalized linear classifier with maximum-margin fitting functions. This function gives regularization which helps the classifier generalized better. Traditional statistical and Neural Network methods control model complexity by using a small number of features. SVM curbs the model complexity by con-trolling the VC dimensions of its model. This method is inde-pendent of dimensionality and can employ spaces of very large dimensions spaces, which grants a construction of very large number of non-linear features and then performing ro-bust feature selection during training [22].

3.4 Matching Techniques

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

i. Whole-word matching. The engine compares the incoming digital-audio signal against a prerecorded template of the word. This technique takes much less processing than sub-word matching, but it requires that the user (or someone) pre-record 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.

ii. Sub-word matching. The engine looks for sub-words - usually phonemes - and then performs further 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 [23].

4 Performance Evaluation of Speech RecognitionSystems

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) [24].

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. 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

$$WER = (S+D+I)/N$$

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:

$$WRR = 1 - WER = 1 - (S+D+I) / N$$
$$= (H - I) / N$$

Where H = (N-S-D) is the correctly recognized words

5 Conclusion

In this paper we have reviewed, the classification and develpment of speech recognition system. We have also discused the commonly used feature extraction techniques which contributes maximum recognition accuracy in any speech recognition application.

6 Acknowledgments

The author remains thankful to Department of CS&IT, Dr.BAMU, Aurangabad, for their useful discussions and suggestions during the preparation of this technical paper. This work is supported by University Grants Commission.

References

[1] Pukhraj Shrishrimal, R.R. Deshmukh, and Vishal Waghmare "Indian Language Speech Database: A Review". International Journal of Computer Application (IJCA), Vol 47, No. 5, pp. 17-21, 2012.

[2] Michelle Cutajar, Edward Gatt, Ivan Grech, Owen Casha, Joseph Micallef, "Comparative study of automatic speech recognition techniques," IET Signal Process, Vol. 7, Iss. 1, pp. 25–46, 2013.

[3] Pratik K. Kurzekar, Ratndeep R. Deshmukh, Vishal B. Waghmare, Pukhraj P. Shrishrimal, "Continuous Speech Recognition System: A Review", Asian Journal of Computer Science and Information Technology, 2014.

[4] M. A. Anusuya, S.K. Katti, "Speech Recognition by Machine: A Review", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009.

[5] X. D. Huang, "A Study on Speaker - Adaptive Speech Recognition", Proc. DARPA Workshop on Speech and Natural Language, pp. 278-283, February 1991.

[6] R.K. Aggarwal, M. Dave "Integration of Multiple acoustic and language models for improved Hindi speech Recognition system", Springer Science Business Media, LLC 2012.

[7] Santosh K. Gaikwad, Bharti Gawli, Pravin Yannawar, "A Review of Speech Recognition Technique", International Journal of Computer Applications (0975–8887) Volume 10, No.3, November 2010.

[8] Asghar .Taheri, Mohammad Reza Trihi et.al, Fuzzy Hidden Markov Models for speech recognition on based FEM Algorithm, Transaction on engineering Computing and Technology V4 February 2005, and IISN, 1305-5313.

[9] Kenneth Thomas Schutte "Parts-based Models and Local Features for Automatic Speech Recognition" B.S.University of Illinois at Urbana-Champaign (2001) S.M., Massachusetts Institute of Technology (2003).

[10] M.A Zissman,"Predicting, diagonosing and improving automatic Language identification performance", Proc.Eurospeech97, Sept.1997 vol.1, pp.51-54 1989.

[11] Ankit Kumar, Mohit Dua, "Continuous Hindi Speech Recognition using Monophone based Acoustic Modeling", International Journal of Computer Applications® (0975 – 8887), 2014.

[12] C.S.Myers and L.R.Rabiner, A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition , IEEE Trans. Acoustics, Speech Signal Proc.,ASSP-29:284- 297,April 1981.

[13] Sanjivani S. Bhabad, Gajanan K. Kharate, " An Overview of Technical Progress in Speech Recognition", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.

[14] Vimala.C, Dr.V.Radha, "A Review on Speech Recognition Challenges and Approaches", World of Computer Science and Information Technology Journal, Vol. 2, No. 1, 1- 7, 2012.

[15] L.R.Bahl et.al, A method of Construction of acoustic Markov Model for words, IEEE Transaction on Audio,speech and Language Processing ,Vol.1,1993.

[16] Gerhard Rogoll,Maximum Mutual Information Neural Networks for hybrid connectionist-HMM speech Recognition systems ,IEEE Transaction on Audio, speech and Language Processing Vol.2 ,No.1,Part II,Jan.1994.

[17] Tavel R.K.Moore, Twenty things we still don't know aboutspeech proc.CRIM/FORWISS Workshop on Progress and Prospects of speech Research and Technology 1994.

[18] M.J.F.Gales and S.J young, Parallel Model combination for Speech Recognition in Noise technical Report, CUED/FINEFENG/TRI135, 1993.

[19] Abhishek Thakur, Naveen Kumar, "Automatic Speech Recognition System for Hindi Utterance with Regional Indian Accents: A Review", International Journal of Electronics & Communication Technology, Vol. 4, April – June 2013.

[20] Thiang, and Suryo Wijoyo, "Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot", International Conference on Information and Electronics Engineering IPCSIT vol.6, IACSIT Press, pp.179-183, Singapore, 2011.

[21] J. Manikandan, and B. Venkataramani, "Design of a real time automatic speech recognition system using Modified One Against All SVM classifier" Microprocessors and Microsystems, ELSEVIER, Vol.35, pp. 568–578, 2011.

[22] Amol T., Kokane ; Guddeti, Ram Mohana Reddy," Multiclass SVM-based language-independent emotion recognition using selective speech features", International Conference on Advances in Computing, Communications and Informatics, Pp.1069 –1073, 2014.

[23] L.R.Rabiner and B.H.jaung," Fundamentles of Speech Recognition Prentice-Hall, Englewood Cliff, New Jersy, 1993.

[24] 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, May 2000.