International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 3 Issue 4, April-2014, pp: (505-511), Impact Factor: 1.252, Available online at: www.erpublications.com

Speech Recognition Techniques-A Review

Meenal Goel[1], Sukhwinder Singh[2]
[1]Student, E & E C Department, PEC University of Technology, Sector 12, Chandigarh, INDIA
[2]Assistant Professor, E & EC Department, PEC University of Technology, Sector 12, Chandigarh, INDIA

Abstract: This paper presents a brief survey on Automatic Speech Recognition and discusses the major themes and advances made in the past years of research, 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 and performance evaluation. Hence author hopes that this work shall be a contribution in the area of speech recognition. 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: Automatic Speech Recognition, Feature extraction, Performance evaluation, Robust speech recognition, Statistical Modelling, Word error rate.

I. INTRODUCTION

The speech is primary mode of communication among human being and also the most natural and efficient form of exchanging information among humans. Speech Recognition can be defined as the process of converting speech signal to a sequence of words by means of algorithm implemented as a computer program. Speech processing is one of the exciting areas of signal processing. The goal of speech recognition area is to develop the technique for speech input to machine. Based on major advancement in statistical modelling of speech, automatic speech recognition today finds widespread application in task that require human machine interface such as automatic call processing. Since the 1960s computer scientists have been researching ways and means to make computers able to record, interpret and understand human speech.

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. The style of computation that can be performed by networks of such units bears some similarity to the style of computation in the nervous system. Parallel distributed processing or massively distributed processing are terms used to describe these models. Connectionist models rely critically on the availability of good training or learning strategies.

Throughout the decades this has been a daunting task. Even the most rudimentary problem such as sampling voice was a huge challenge in the early years. It took until the 1980s before the first systems arrived which could actually decipher speech. Machine recognition of speech involves generating a sequence of words that best matches the given speech signal. Some of known applications include virtual reality, multimedia searches, auto-attendants, travel information and reservation, translators, natural language understanding and many more applications.

Definition of Speech Recognition: Speech Recognition (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.

Basic model of Speech Recognition: The recognition process is shown below (Fig .1).

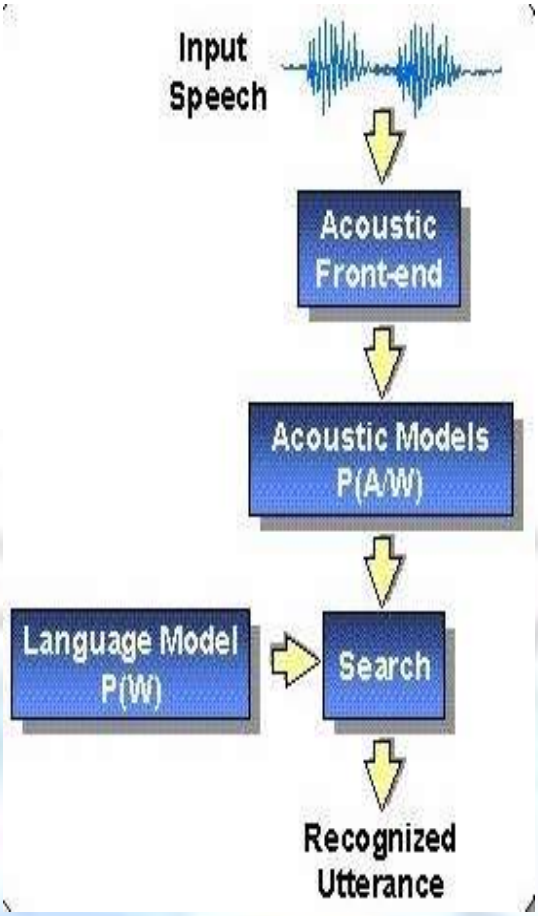

Fig.1: Basic model of speech recognition

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:

i. 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.

ii. Connected Words: Connected word systems are similar to isolated word systems but allow separate utterances to be 'run-together' with a minimal pause between them.

iii. Continuous Speech: Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize special methods to determine utterance boundaries.

iv. Spontaneous Speech: At a basic level, it can be thought of as a 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 and even slight stutters.

Automatic speech recognition system classification: The following tree structure emphasizes the speech processing applications. Automatic Speech Recognition systems can be classified as shown in figure 2.

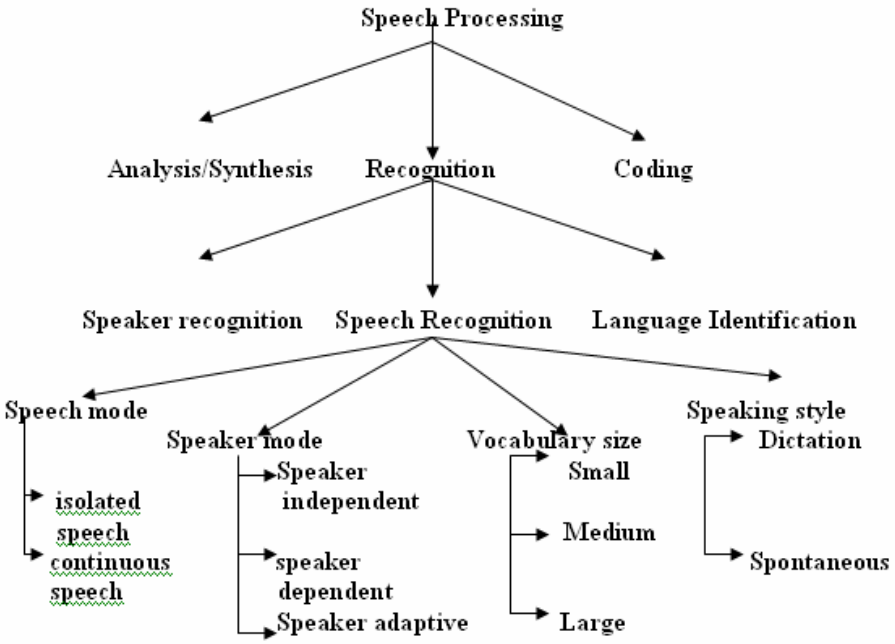

Fig.2 Speech processing classification

II. APPROACHES TO SPEECH RECOGNITION

There exist three approaches to speech recognition:

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

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 (Hemdal and Hughes 1967), which postulates that there exist finite, distinctive phonetic units (phonemes) in spoken language and that these units are characterized by a set of acoustic properties that are expressed in the speech signal over time. Even though, the acoustic properties of phonetic units are highly variable, both with speakers and with neighbouring sounds (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. The first step in this 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 segmentation and labelling in which the speech signal is divided into stable acoustic regions, followed by attaching one or more phonetic labels to each divided 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 segmentation to labelling [1]. 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. This approach has not been widely used in most commercial applications.

Pattern recognition approach: The pattern-matching approach [1] 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 labelled training samples through 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 speech according to the matching of the patterns. The pattern-matching approach has become the predominant method for speech recognition in the last six decades. In this, there exists two methods namely template approach and stochastic approach.

i. **Template based approach:** Template based approach [2] has provided a family of techniques that have advanced the field considerably. A collection of typical speech patterns are stored as reference patterns representing the dictionary of candidates' words. Recognition is then carried out by matching an unknown spoken utterance with each of these reference templates and selecting the category of the best matching pattern. This has the advantage that the errors due to division or classification of smaller acoustically more variable units such as phonemes can be avoided. Each word must have its own full reference template. Template preparation and matching becomes impractical as vocabulary size comes out be beyond a few hundred words. One idea in template based approach is to derive a typical sequence 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.

ii. **Stochastic approach:** Stochastic approach involves 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, contextual effects, and homophones. Thus, stochastic approach is suitable for speech recognition. The most popular stochastic approach today is hidden Markov modelling. A hidden Markov model is characterized by a finite state Markov model and a set of output distributions. Compared to template based approach, hidden Markov modelling is more general and has a stronger mathematical foundation. HMMs do not provide much insight on the recognition process, so it is often difficult to analyse the errors of an HMM system in an attempt to improve its performance. However, careful incorporation of knowledge has significantly improved HMM based systems.

Dynamic time warping (dtw): Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. For e.g., similarities in walking patterns would be detected, even if in one video, the person was walking slowly and if in another, he or she was walking more quickly, or even if there were accelerations and decelerations in the other. This approach can be applied to any data which can be turned into a linear representation. To cope with different speaking speeds, Dynamic Time Warping finds application in automatic speech recognition. Dynamic Time Warping is a method that allows a computer to find an optimal match between two given sequences with certain restrictions. The sequences are "warped" non-linearly in the time dimension to determine the measure of their similarity. DTW is a method particularly suited to matching sequences with missing information, provided there are long enough segments for matching to occur. The optimization process is performed using dynamic programming, hence the name.

Vector Quantization (VQ): Vector Quantization (VQ) [2] is often applied to speech recognition systems. Since transmission rate is not a major issue for ASR, the use of VQ lies in the efficiency of using compact codebooks for reference models and codebook searcher in place of more costly evaluation methods. Each vocabulary word gets its own VQ codebook, based on training sequence of several repetitions of the word. The test speech is assessed by all codebooks and ASR chooses the word whose codebook provides the lowest distance measure. In basic VQ, codebooks have no explicit time information as 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). Such segments are thus more likely to specify code words than less frequent consonant frames. Code words, however, exist for constant frames because such frames would otherwise contribute large frame distances to the codebook. This relative emphasis that VQ puts on speech transients can be an advantage over other ASR comparison methods for vocabularies of similar words.

Knowledge based approach: The Artificial Intelligence approach [3] is an amalgam of the acoustic phonetic approach and pattern recognition approach. Some speech researchers developed recognition systems that used acoustic phonetic knowledge to classify speech sounds. While template based approach has been very effective in the design of a variety of speech recognition systems; they provided little information about human speech processing, hence making error

analysis and development of knowledge-based system difficult. Knowledge based approach involves the incorporation of expert's speech knowledge into a recognition system. This knowledge is usually developed from careful study of spectrograms and is included using rules or procedures. However, this approach had only limited success 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 asynchronous knowledge sources still remains an unsolved problem. Knowledge has also been used to enhance the design of the models and algorithms of other techniques and this knowledge based enhancement has contributed considerably to the design of all successful strategies reported.

Connectionist approaches (artificial neural networks): The artificial intelligence approach (knowledge based approach) attempts to mechanize the recognition procedure according to the way a person applies intelligence in 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 incorporates phonemic, lexical, syntactic, semantic and 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 is in the representation of knowledge and integration of knowledge sources. 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. The style of computation that can be performed by networks of such units bears some similarity to the style of computation in the nervous system. Parallel distributed processing or massively distributed processing are terms used to describe these models. Connectionist models rely critically on the availability of good training or learning strategies. Connectionist model tries to organize a network of processing elements. Multilayer neural networks can be trained to generate complex nonlinear classifiers or mapping function. The simplicity and uniformity of the underlying processing element makes connectionist models attractive for hardware implementation. Training often requires much iteration over large amounts of training data and can be excessively expensive. While connectionist model appears to hold great promise as a reasonable model of speech recognition, but question relating to the realization of practical connectionist recognition techniques, still remain to be resolved.

Support Vector Machine (SVM): 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.

II. FEATURE EXTRACTION

In speech recognition, aim of the feature extraction [4] step is to figure out a 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 spectro-temporal 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 rule 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 differentiate between different, though, similar 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 invariant across speakers and speaking environment. Some of the methods of feature extraction are: Principal Component

Analysis(PCA), Linear Discriminant Analysis(LDA), Independent Component Analysis (ICA), Linear Predictive Coding, Cepstral Analysis, Mel-frequency scale analysis, Filter bank Analysis, Mel-frequency cepstrum (MFFCs), Kernel based feature extraction method, Dynamic feature Extractions, Spectral Subtraction, Cepstral mean Subtraction, RASTA filtering, Integrated Phoneme subspace method, etc.

IV. PERFORMANCE OF SPEECH RECOGNITION SYSTEMS

The performance of speech recognition systems [5] is given in terms of accuracy and speed. Accuracy is 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. The Word Error Rate is derived from the Levenshtein distance, working at the word level instead of the phoneme level. Word error rate can then be computed as:

$$WER = \frac{S + D + I}{N} \quad (1)$$

Where :

- 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:

$$WRR = 1 - WER = \frac{N - S - D - I}{N} = \frac{H - I}{N} \quad (2)$$

Where

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

V. DISCUSSIONS AND CONCLUSIONS

Speech is the primary, and the most convenient means of communication between people. Whether due to technological curiosity to build machines 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. 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. This paper attempts to provide a research on speech recognition.

Although significant progress has been made in the last two decades, there is still work to be done, and I believe that a robust speech recognition system should be effective under full variation in: environmental conditions, speaker variability etc. Speech Recognition is a challenging and interesting problem. 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. I hope this paper brings about awareness and inspiration amongst the researchers of ASR.

ACKNOWLEDGMENT

The author remains thankful to Prof. Sukhwinder Singh for his useful discussions and suggestions during the preparation of this technical paper.

REFERENCES

[1]. 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.

[2]. D. Raj Reddy "Speech Recognition by Machine: A Review"PROCEEDINGS OF THE IEEE, VOL. 64, NO. 4, APRIL 1976.

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

[4]. 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.

[5]. Pradeep Kumar Jaisal, Pankaj Kumar Mishra "A Review of Speech Pattern Recognition: Survey" IJCST Vol. 3, Issue 1, Jan. - March 2012.

[6]. Preeti Saini, Parneet Kaur "Automatic Speech Recognition: A Review" International Journal of Engineering Trends and Technology- Volume4 Issue2, 2013.

[7]. Rajesh Kumar Aggarwal and M. Dave, "Acoustic modelling problem for automatic speech recognition system: advances and refinements Part (Part II)", Int J Speech Technol, pp. 309–320, 2011.

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

[9]. Anusuya, M. A., & Katti, S. K. Front end analysis of speech recognition: A review. International Journal of Speech Technology, Springer, vol.14, pp. 99–145, 2011.

[10]. Wiqas Ghai and Navdeep Singh,"Literature Review on Automatic Speech Recognition", International Journal of Computer Applications vol.41– no.8, pp. 42-50, March 2012.