International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Study on Speech Recognition

Mridula Shanbhogue[1], Shreya Kulkarni[2], R Suprith[3], Tejas K I[4], Nagarathna N[5]
Dept. Of Computer Science & Engineering, B.M.S. College of Engineering, Bangalore, India

Abstract— Among human beings, speech is a primary medium of interaction. Speech is the most natural and easy way of information exchange. As modern devices are aimed to be user friendly, voice interaction with machines is the most easy and convenient way of communicating. This paper concentrates on Speech recognition technology that allows for hand-free interface and provides better use of present day technologies. Thus speech recognition technology will help remove technological barriers and provides solution for easy human machine interaction. This paper discusses about the various types of speech and the methods to extract and analyze the structure of the speech. The paper summarizes about speech recognition techniques and various algorithms for speech-to-text conversion.

Keywords— Automatic Speech Recognition (ASR), Feature Extraction, Acoustic Model, Speech Analysis, Hidden Marcov Model (HMM), Mel Frequency Cepstral Coefficient (MFCC), Principal Component Analysis (PCA), Linear Discriminate Analysis (LDA).

I. INTRODUCTION

Information exchange and communication are important parts of our day-to-day lives. Speech is a natural and primary mode of interaction. Among human, efficient information exchange is achieved through speech. As technology is developing, many modern devices are developed to ease manual work and such devices require human –machine interaction. For this purpose, many interfaces are developed. As speech interaction is the most natural and general way of communicating, natural language speech recognition is the latest technological development for human machine interaction. Speech recognition is also known as "Automatic speech recognition (ASR)" which is defined as the process of translating spoken words into corresponding meaningful texts. Using various algorithms which are implemented as computer programs, sequence of words are generated from speech signals. As spoken language is a primary mode of communication, speech interfaces will meet the expectation of people. It is comfortable to use speech interface than using other primitive interfaces like keyboard, mouse etc. People with motor impairment, visually impaired people face difficulty using computers and other modern day devices. They face difficulty in using the current primitive interfaces and such interfaces also require certain level of literacy from user. Speech recognition technology will help deal with such issues and build speech interfaces for different applications. Speech recognition technology provides an option for speech input/output to use such systems efficiently. Voice interaction will help users to multitask. Elderly, physically challenged people and general community will be benefited by speech recognition technology which will keep them closer to Information technology revolution. Speech recognition technology can be incorporated in building voice interfaces for ATM machines, computer systems, web browsers, cellular phones, household appliances and many more. Speech recognition system acquires speech input signal and pre-processes it to extract useful information. Speech recognition is done using the extracted information and translated to suitable text.

II. SPEECH RECOGNITION

Speech Recognition refers to the process of recognizing the speech and translating spoken words into meaningful text. "Automatic speech recognition (ASR)" or "speech-to-text" are the other common names used for this technology. The spoken words which consist of the sound wave, are captured by a microphone and converted into electrical signals. The obtained electrical signals are then converted into digital form which can be understood by the system. Speech signals are then converted into discrete sequence of feature vectors, containing relevant information about the spoken word [1]. Decoding is performed for finding the best match for the feature vectors using the knowledge base [2] and corresponding text is generated. Fig 1 explains the outline of speech recognition system.

Volume 4 Issue III, March 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

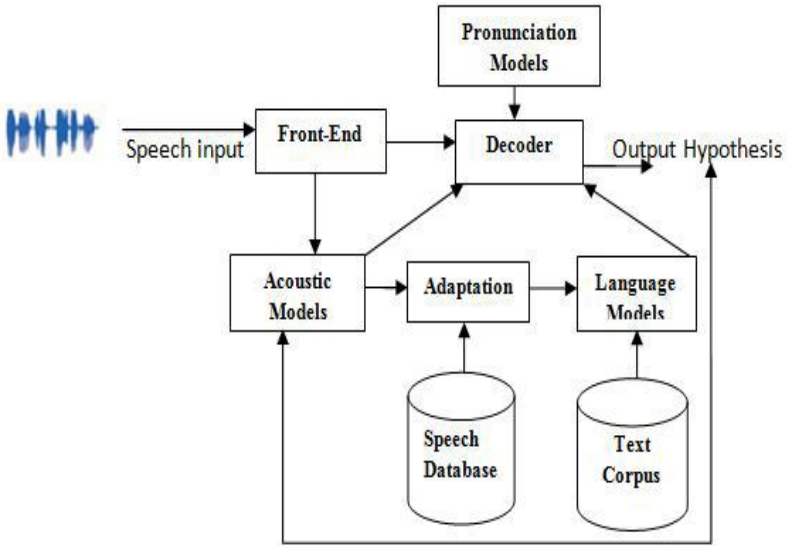

Fig 1 Speech Recognition System Diagram [2]

In speech to text technology the user interacts with system through speech input and user friendly front-end helps in doing so. This speech waveform is statistically represented by acoustic model. Adaptation reduces environmental noise in statistical representation and increases accuracy. The language model provides sequence of words that probably match the actual word. Language model uses text corpus which is a structured set of texts. Decoder with the help of pronunciation model selects the best match and gives text output.

III. TYPES OF UTTERANCES

The Speech recognition systems are classified based on their ability of recognizing types of utterances/words [3]. They are classified as i) Isolated word ii) Connected word iii) Continuous speech iv) Spontaneous speech.

A. Isolated Word

In this system, each utterance should have quiet bandwidth on either sides of sample windows. This system accepts single utterance or single word at a time .There exists two states namely “Listen and Non Listen state”. This class can be also called as isolated utterance.

B. Connected Word

Connected word system recognizes separate utterances which can be “run together” with minimum pause between each utterance. Speech that has to be recognized can contain single word, or a collection of words, a single sentence, or multiple sentences. Every possible region of the connected word-input pattern is mapped against each word-pattern stored. This is carried out in word-level stage. A region in which each word in the connected word-pattern may start and end is defined by an adjustment window.

C. Continuous Speech

In this system users are allowed to speak naturally and continuous speech system analyses and determines the spoken content. Recognizer with continuous speech utilizes special method to determine utterance boundaries and thus such systems are difficult to build.

D. Spontaneous Speech

Natural unrehearsed speech is known as spontaneous speech. Speech can consist of false-starts, mispronunciations and non-words. Variety of natural features of speech such as many words being run together, non-words and wrong pronunciations are handled by ASR systems with spontaneous speech ability [4].

Isolated word and connected word speech recognition systems are widely developed and used for different applications like ATM machines, household appliances etc. Continuous speech and Spontaneous speech systems are complex to build and research is going on for developing efficient ASR systems.

IV. TYPES OF VOCABULARY

The vocabulary size affects the complexity, processing necessities, performance and precision of the speech recognition system. Some of the applications require only few words and others require very large and sophisticated dictionaries. The different types of vocabularies can be classified as follows.

A. Small Vocabulary – tens of words
B. Medium vocabulary – hundreds of words
C. Large Vocabulary – thousands of words
D. Very-large Vocabulary – millions of words
E. Out-of-Vocabulary – mapping a word from the vocabulary into the unknown word

Various other characteristics like environment variability, speaker style, sex, age, speed of speech and the variability in the signal also makes the speech recognition system more complex. Thus efficient speech recognition systems are developed based on the accuracy needs of the incorporating applications.

V. STAGES IN SPEECH RECOGNITION

For a machine to be able to "listen", "perceive" and "act in accordance with" spoken information is the goal of speech recognition. Bell Laboratories were the first to attempt developing recognition system in the 1950s [5]. Single digit recognition system was built by Balashek, Biddulph and Davis. This system can be used by a single user (speaker) [6]. The speaker recognition system can be divided into four working stages [7]:

A. Speech Analysis

Speech analysis is the first stage in speech recognition. Selection of suitable frame size to segment the input speech signal is the first part of speech recognition. This is carried out in speech analysis stage. Further analysis and extraction is done using segmented speech [9], [10]. The speech analysis can be done with following three methods/techniques:

1) *Segmentation Analysis:* Suitable frame size and shift in the range of 10-30ms is used to segment the input speech signal and analyze it. It is used in speaker recognition to extract vocal tract information.
2) *Sub Segmental Analysis:* In sub segmental analysis, speech signal is divided using frame size and shift in range of 3-5ms and then used to analyze speech. Characteristic of the excitation state is analyzed and extracted using this analysis [11].
3) *Supra Segmental Analysis:* Frame size suitable for analysis is used to segment the speech signal and analyze it. This method is mainly used to analyse the behaviour character of the speaker.

The above three speech analysis techniques are categorized by using the frame size and range of the speech under consideration. These methods are mainly used to analyze and characterize the speech.

B. Feature Extraction

The extraction of the features from speech signal is an important task to produce a better recognition performance. It is the process discarding unwanted and redundant information like background noise to extract useful information. The speech signals are converted to digital form and signal characteristics are measured to obtain acoustically identifiable components. Following are the various feature extraction methods:

1) *Principal Component analysis (PCA):* PCA is used to obtain set of values called principal components from set of observations. It is a statistical procedure. The total number of Principal components is comparatively lower than that of original variables. The transformation is done such that the first component is able to account for as much as of the variability in the data as possible. This nonlinear feature extraction method produces linear components and is fast and eigenvector-based.
2) *Linear Discriminate Analysis (LDA):* It is used to obtain set of objects by converting set of observations. These objects are split into groups based on the features that describe them. LDA works when the measurements made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis. This is a nonlinear feature extraction method which is fast and eigenvector based and better than PCA for classification.
3) *Linear Predictive Coding:* It is a static feature extraction method. Speech signal is analyzed by estimating the formants. It removes the effect of formants from signal to produce useful features. It has 10 to 16 lower order coefficients and hence extracts the features at lower order.

Volume 4 Issue III, March 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

4) *Mel Frequency Cepstral Coefficient (MFCC):* In this method, the sound is represented as a short-term power spectrum on a non-linear Mel scale of frequency. From the cepstral representation of sound some coefficients are derived that collectively form Mel Frequency Cepstrum. Finding the Fourier Transform of the sound signal is the first step in deriving the coefficients. The powers of the Fourier Transform are then mapped onto the Mel scale. This mapping is known as filtering. At each of the Mel frequencies logarithms of the powers are calculated. The discrete cosine transform values of Mel log powers are taken. This will result in a spectrum having MFCCs as amplitudes. These MFCCs are the extracted features of sound signal. Thus Feature vector for MFCC is obtained from the original speech signal.

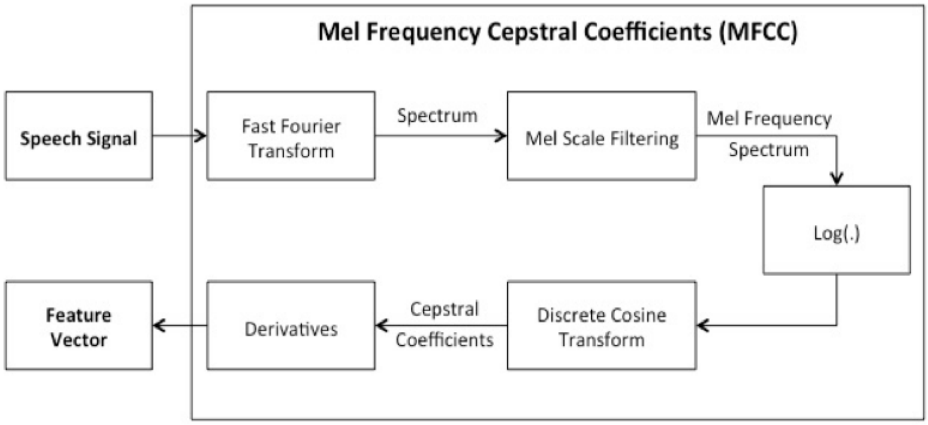

Fig 2 Mel Frequency Cepstral Coefficient Generation [11]

The MFCC is the most superior method used for feature extraction. It is the most prevalent method as it provides the accurate results when compared to other existing feature extraction methods.

C. Modelling Technique

Modelling is the process of generating the speaker model using the feature vector. Different kinds of modelling approaches are used for this purpose. Some of them are as follows:

1) *Acoustic-Phonetic approach:* This kind of approach deals with acoustic aspects of spoken words. It analyses some of the properties of sound wave like amplitude, duration, fundamental frequency. Based on the observed features and with the knowledge of the relationship between acoustic features and phonetic symbols the spoken words are decoded to obtain sequence of phonemes and other linguistic units. The first step involves the parameter measurement where the spectral representation of the speech signal is obtained. In the next stage spectral measurements are converted into a set of features that describe the acoustic property of the various phonetic units. Numbers of feature detectors are used for this purpose. The obtained features are combined and decision logic is applied on it to recognize spoken word. Finally hypothesis tester checks whether sufficient amount of features are available using vocabulary features and gives recognized speech as output. The recognized speech will be the best matching word or sequence of words.

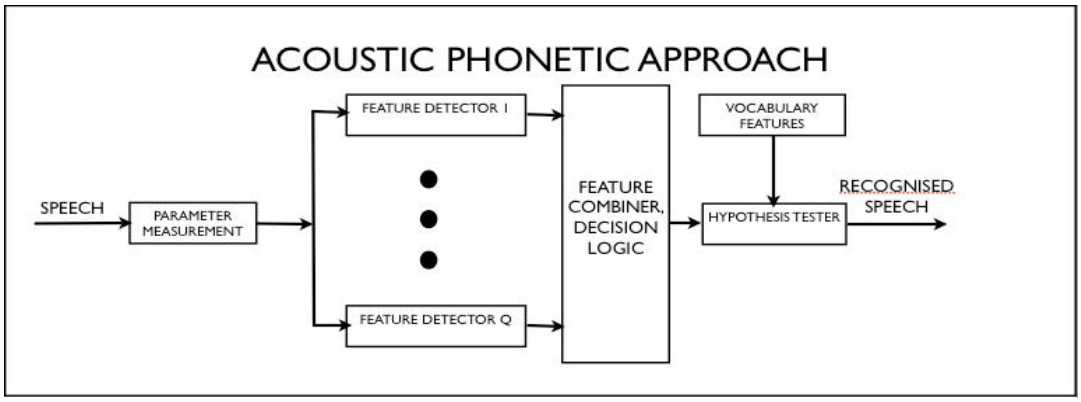

Fig 3 Speech Recognition by Acoustic Phonetic Approach [13]

Volume 4 Issue III, March 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

2) *Pattern Recognition approach:* This approach deals with speech patterns that are directly used without segmentation and explicit feature determination. This method involves two major stages – i) training of speech patterns, ii) recognition of the patterns by the way of pattern comparison. On the input speech signal, sequence of measurements is made to get the test pattern. This phase is known as parameter measurement phase. The generated test pattern is compared with each sound reference pattern and similarity between every test and reference pattern pair is measured. The calculated value is known as the similarity score which is used by decision rule for deciding the reference pattern that best matches the test pattern.

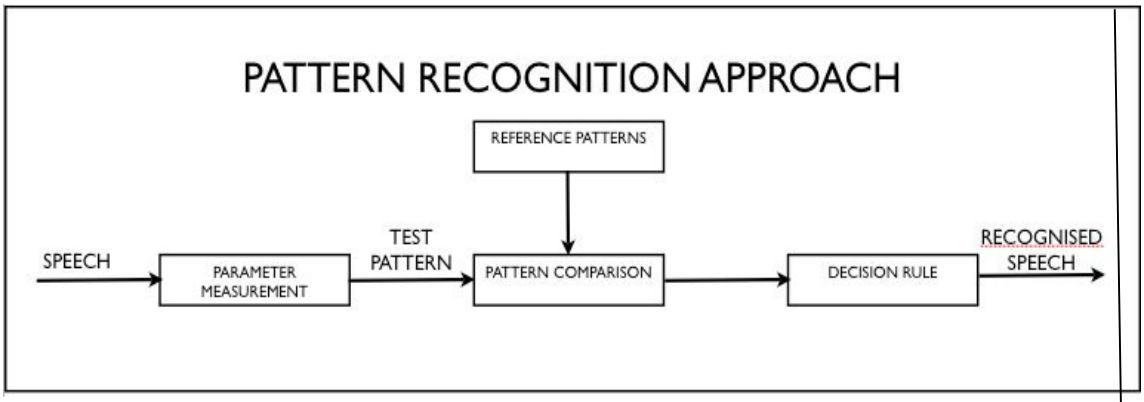

Fig 4 Speech Recognition by Pattern Recognition Approach [13]

3) Artificial Intelligence approach: This approach is based on both pattern recognition approach and acoustic-phonetic approach. Artificial intelligence approach is a composition of previous approaches. An expert system or a self-organizing (learning) system implemented by neural network is used in this approach. This system gains knowledge from a variety of knowledge sources and stores it for the future use. Whenever it receives a voice input with the help of pre-known knowledge it classifies sound.

4) Hidden Marcov Models (HMMs): This is a statistical model where the system is considered a Marcov process with unobserved (hidden) states. The input is represented as a set of connected states where each state is associated with a probability distribution. The actual state is always hidden for an external observer. Hence it is known as Hidden Marcov Model. HMM contains N states, an initial state distribution, a state transition matrix and an emission probability distribution. The initial state distribution determines the initial probabilities of HMM states. The emission probability distribution gives the emission probability at a state and the state transition matrix provides state-to-state transition probabilities. With the help of these a sequence of statistical operations are performed on the input which gives the required output.[16]-[20]

D. Matching Technique

Matching is the process of matching a detected word to a known word. There are two techniques to achieve this.

1) *Whole word matching:* In this technique the input word is compared against pre-recorded template of words. Among the known words, the one that best matches is discovered. This technique takes less processing time and can compare several thousand words. It needs a pre-record of every word that it recognizes and large amount of memory to store them. [14]
2) *Sub-word matching:* In this technique the engine searches for the sub words like phonemes. It performs the pattern recognition for sub-words. The required processing time is more, but needs less memory when compared to whole word matching. [13], [15]

VI. CONCLUSION

In the paper, we discussed about speech recognition technique through its various stages. MFCC is the efficient feature extraction technique and it is widely used and HMM is widely used for modelling speech. This paper brings about the importance of speech recognition to build speech interfaces and how the speech recognition technique works.

VII. ACKNOWLEDGEMENT

The work reported in this paper is supported by the college through the TECHNICAL EDUCATION QUALITY IMPROVEMENT PROGRAMME [TEQIP - II] of the MHRD, Government of India.

REFERENCES

[1] Preeti Saini, Parneet Kaur, CSE Department, Kurukshetra University ACE, Haryana, India "Automatic Speech Recognition: A Review" - International

Journal of Engineering Trends and Technology- Volume4Issue2- 2013
[2] Rajesh Kumar Aggarwal and M. Dave, “Acoustic modeling problem for automatic speech recognition system: advances and refinements Part (Part II)”, Int J Speech Technol, pp. 309–320, 2011.
[3] Sanjib Das, “Speech Recognition Technique: A Review”, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 Vol. 2, Issue 3, May-Jun 2012.
[4] 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
[5] Zahi N. Karam, William M. Campbell “A new Kernel for SVM MIIR based Speaker recognition" MIT Lincoln Laboratory, Lexington, MA, USA.
[6] R. Klevansand R. Rodman, “Voice Recognition", Artech House, Boston, London 1997.
[7] Om Prakash Prabhakar, Navneet Kumar Sahu, “A Survey On: Voice Command Recognition Technique" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013 ISSN: 2277 128X
[8] B. Yegnanarayana, S.R.M. Prasanna, J. M. Zachariah, and C.S. Gupta, “Combining evidence from source, supra segmental and spectral features for a fixed-text speaker verification system”, IEEE Trans. Speech Audio Process., vol. 13(4), pp. 575-82, July 2005.
[9] Gin-Der Wu and Ying Lei “A Register Array based Low power FFT Processor for speech recognition" Department of Electrical engineering national Chi Nan university Puli ,545 Taiwan.
[10] Shikha Gupta, Mr. Amit Pathak, Mr. Achal Saraf “A study on Speech Recognition System: A literature Review" International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 8, August 2014,2192,ISSN: 2278 – 7798
[11] Nicolas Morales1, John H. L. Hansen2 and Doorstep T. Toledano1 “MFCC Compensation for improved recognition filtered and band limited speech” Center for Spoken Language Research, University of Colorado at Boulder, Boulder (CO), USA.
[12] M. A. Anusuya, S. K. Katti “Speech Recognition by Machine: A Review” International journal of computer science and Information Security 2009.
[13] L. R. Rabiner, B. H. Juang. "Fundamentals of Speech Recognition", Prentice-Hall, Inc., Upper Saddle River, NJ. 1993.
[14] S. katagiri, Speech Pattern recognition using Neural Networks.
[15] D. R. Reddy, “An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave”, Tech. Report No.C549, Computer Science Dept., Stanford University., September 1966.
[16] Shigeru Katagiri et.al, "A New hybrid algorithm for speech recognition based on HMM segmentation and learning Vector quantization", IEEE Transactions on Audio Speech and Language processing Vol.1, No.4
[17] L.R Rabiner. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." Proceedings of the IEEE. 77(2):257-286. 1989.
[18] Keh-Yih Su et. al., Speech Recognition using weighted HMM and subspace IEEE Transactions on Audio, Speech and Language.
[19] 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
[20] G. 2003 Lalit R .Bahl et. al., Estimating Hidden Markov Model Parameters so as to maximize speech recognition Accuracy, IEEE Transaction