Automatic Speech Recognition System helping Unlettered and Visually Impaired People in Hindi Language - A Review

Harshit Savani[1] Mihir Tandel[2] Chirag Shetty[3]
[1,2,3]Department of Electronics & Telecommunication Engineering
[1,2,3]TCET, Kandivali (E), Mumbai, India

Abstract— Technology is advancing so fast that we cannot live a day without using it. We now want machines to communicate with humans and solve their problems according to their needs. Humans are fascinated by machines that can understand us. Speech is hand-free, fast, easy and does not require any technical knowledge. Speech recognition and processing has received a lot of attention during the last few decades. Speech recognition has been an important area of research. Many applications of speech recognition had been employed in different fields like in telephony, medical or command and control area. For years people have tried to develop machines that can understand and produce speech as humans do naturally. The first speech recognition system could understand only digits, but now, the technology is developed such that speech recognition system can easily be able to recognise any language. A lot of attempts have been made to develop such vocally interactive computers to realize voice/speech recognition systems.

Key words: Automatic Speech Recognition, Feature Extraction, Classifiers, HMM, MFCC, Hindi Language, Isolated Word Vocabulary

I. INTRODUCTION

For the development of the country, where most of the people live in rural areas as a whole, the technology has to reach them as well. From the initial attempts to identify isolated words using a very limited vocabulary, to the latest advancements processing continuous speech composed by thousands of words, the ASR technology has grown progressively [1]. The various computer accessories, like the keyboard and the mouse, require a certain level of expertise from the user, which cannot be expected from rural people or from the physically challenged or blind people. In India, where the literacy rate is low, the above-mentioned constraints have to be discarded, this is where speech recognition becomes useful [2]. In India, most of the population lives in rural areas and is unfamiliar with computers and English. It will enable people to interact with computers in their own language and without the use of keyboard. Due to the rapid development in this field all over the world we can see many systems and devices with voice input and output, e.g. automated information systems, personal dictation systems converting speech to text, systems for automated transcription of audio/video recordings or radio or TV [3] on-line inputs, devices in cars controlled by voice etc. The objective is to trap human voice in a digital computer and decode it into corresponding text.

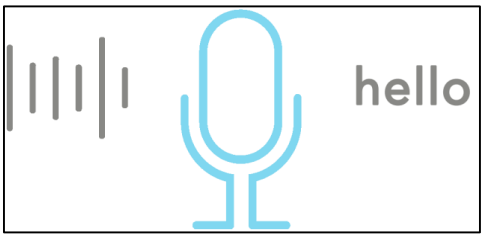

Fig. 1: Basic Speech Recognition System

II. RELATED WORK

Speech recognition process includes different stages such as preprocessing, feature extraction and feature matching. Different research papers have been referred to understand the basics of speech recognition process. Different algorithms have been studied to carry out project work related to feature extraction and feature matching techniques. A brief explanation and important results about these are discussed below.

P. Saini et al [4], used feature extraction technique as MFCC and Training model as HMM using HTK tool kit. This paper has calculated the accuracy of the system by considering the amount of states and the number of training and testing samples of voice of various persons. The work implemented in the paper is a step towards the development of such type of systems. The work may further be extended to large vocabulary size and to spontaneous speech recognition. According to their results, the system is sensitive to changing spoken methods and changing scenarios, so the accuracy of the system is a challenging area to work upon. Hence, various speech enhancements and noise reduction techniques may be applied for making system more efficient, accurate and fast.

Harpreet Kaur et al [5], implemented speech recognition system for Punjabi Language. This paper aims at developing a speech recognition system for the set of Punjabi alphabets using Mel Cepstrum Frequency Coefficients (MFCCs) for feature extraction. The system has been trained using Hidden Markov Model (HMM) for estimating the parameter. Testing has been done using the Viterbi Algorithm and the results are calculated in three scenarios. The implementation of the project was done on MATLAB software technique. The tests are conducted using the speech data of 5 female and 10 male speakers. The data set is divided into two parts, where in the first part the tests are conducted using the first 5 alphabets and the second part consists of the further 5 alphabets of Punjabi language. The system is proposed for recognition of 90 alphabets taken from Punjabi language with 9 speakers having Punjabi as their native language from the age group of 20-50 years. The accuracy of the system is evaluated in three different scenarios. The results are found to be 80%, 100% and 55% accurate for each scenario respectively.

Shreya Narang et al [6], in this paper, various feature extraction techniques like LPC, MFCC, RASTA, PLDA were compared. There has been a lot of research in the field of speech recognition but still the speech recognition systems till date are not hundred percent accurate. The systems developed so far have limitations: there are a limited number of vocabularies in the current systems, also there exists a problem of overlapping speech that is the systems cannot identify speech from multiple users, the user needs to be in a place which has background noise free for an accurate recognition, there occurs a problem with the accent and the pronunciation of the user or speaker. In the future, the speech recognition systems need to be free of these limitations to give hundred percent results. The detailed information about their advantages, disadvantages are listed which helps in choosing the best feature extraction technique useful in different situation. Also the paper consists of process flow of different techniques.

Namrata Dave et al [7], the automatic recognition of speech, enabling a natural and easy to use method of communication between human and machine, is an active area of research. Speech features are extracted from recorded speech of a male or female speaker and compared with templates available in database. In this paper, some feature extraction methods are discussed with their pros and cons. LPC parameter is not so acceptable because of its linear computation nature. It was seen that LPC, PLP and MFCC are the most frequently used features extraction techniques in the fields of speech recognition and speaker verification applications. As human voice is nonlinear in nature, Linear Predictive Codes are not a good choice for speech estimation. PLP and MFCC are derived on the concept of logarithmically spaced filter bank, clubbed with the concept of human auditory system and hence had the better response compare to LPC parameters.

III. METHODOLOGY

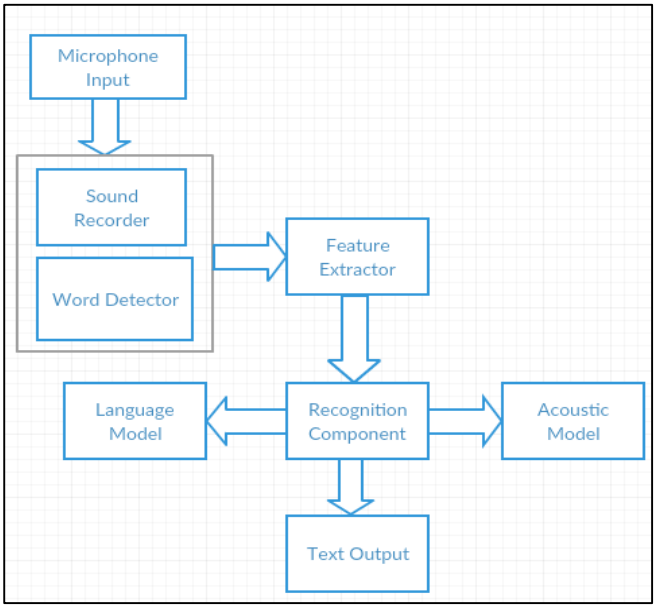

Fig. 2: Block Diagram of Automatic Speech Recognition System

Input to this project will be speech in Hindi language which will be recognized by microphone. Sound recorder will record the speech and produce the electrical signal related to the speech, and then the input goes to the word detector which approximately detects the spoken word. The sound file in ".wav" format further is passed to the feature extraction model which basically converts the sound signal[4]. Feature extraction model basically includes the MFCC technique, which will recognize the phoneme in the spoken language. For Hindi, phonemes are different than English. The word boundary module detects the start and end of the sentence. It also detects energy & the zero crossing rate of the signal. The knowledge models are trained using HMM (Hidden Markov Model).

A. Feature Extraction

Only Voiced segment of the speech signal are processed for MFCC extraction. The procedure to determine MFCC is describe below:

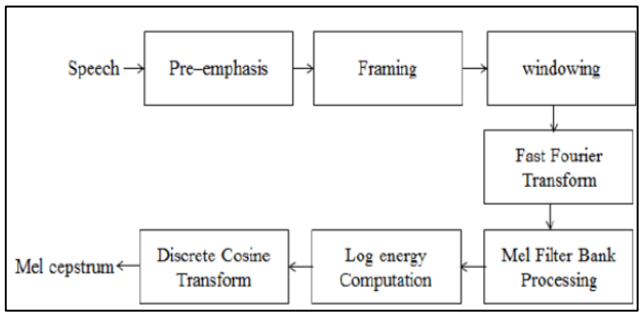

Fig. 3: Workflow of Feature Extraction using MFCC

- Firstly, signal is passed through a first-order high-pass filter. This process will increase the energy of signal at higher frequency.
- After this the signal is framed with the frame size of 10ms to 30ms because in this time interval speech can be considered as periodic.
- The next step is step is to reduce signal discontinuities, and make the ends smooth enough to connect with the beginnings. In order to reduce the discontinuities of the speech signal at the edges of each frame, a tapered window is applied to each frame.
- The next step is to transform each frame of N samples from time domain into frequency domain using the Fast Fourier Transform (FFT).
- The final step of computing filter banks is applying triangular filters, typically 40 filters, on a Mel-scale to the power spectrum to extract frequency bands.
- Log energy Computation is done by computing logarithm of the square magnitude of the output of mel filter bank. Thus the input to Mel filter bank is the power spectrum of each frame, such that for each frame a log spectral energy vector is obtained as output of the filter bank analysis.
- The last process converts the log Mel Spectrum into time domain using discrete cosine transform or inverse DFT. The result the conversion is called Mel Frequency Cepstrum Coefficient.

B. Feature Classification

Classification is another important part of a speech recognition system since the patterns are classified into different classes during this stage[8]. During classification stage, decisions are made based on the similarity measures

from training patterns using information relating to known patterns. Then they are tested using the unknown patterns. Since there are a number of different classes in a speech recognition problem, a multiclass classification technique is needed. In almost all classification methods, the data is separated into training and test sets. Each instance in the training set contains a target value which represents the corresponding class and a set of attributes. The test data do not contain a target value. The objective of the classifier is to produce a model from the training data which predicts the target values of the test data.

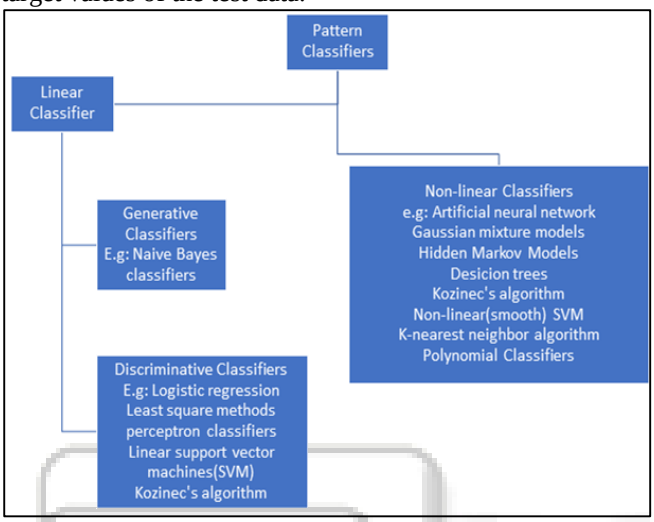

Fig. 4: Types of Classifiers

Hidden Markov Model (HMM) is a statistical Markov demonstrate in which the framework being displayed is thought to be a Markov procedure with imperceptibly (for example concealed) states. The Hidden Markov model can be spoken to as the least complex powerful Bayesian system[9].

In less difficult Markov models (like a Markov chain), the state is specifically obvious to the eyewitness, and along these lines the state progress probabilities are the main parameters, while in the concealed Markov demonstrate, the state isn't straightforwardly unmistakable, however the yield (as information or "token" in the accompanying), reliant on the state, is noticeable. Each state has a likelihood appropriation over the conceivable yield tokens[10]. Hence, the succession of tokens created by a HMM gives some data about the grouping of states; this is otherwise called example hypothesis, a subject of language structure acceptance.

A Hidden Markov model can be viewed as a speculation of a blend display where the shrouded factors (or inactive factors), [11] which control the blend part to be chosen for every perception, are connected through a Markov procedure as opposed to autonomous of one another. As of late, concealed Markov models have been summed up to pairwise Markov models and triplet Markov models which permit thought of progressively complex information structures and the demonstrating of nonstationary information.

IV. Proposed Work

Speech is the simplest mode of communication. If we use speech in exchanging information with the computers it would be advantageous for the society, as getting information would be a quick thing then. Adding more to the fact, that recognition systems are only present in English language, here a system supporting local Hindi language will be developed.

V. Conclusion

An efficient, fast and user friendly ASR System is need of an hour. Although many speech interfaces are already available, the need is for speech interfaces in local Indian language; hence this attempt to build a speech recognition system in Hindi is made. With the help of this project, the people who are living in rural areas can easily communicate with computer without the help of keyboard.

Reference

[1] I. Francisco, C. Cerón, A. Graciela and G. Badillo, "A Keyword Based Interactive Speech Recognition System for Embedded Applications," Master's Thesis, June 2011. (Unpublished)
[2] Abhisek Paul and Satyabrata Chayani,"Speech Recognition in Hindi," 2010 (Unpublished)
[3] J. Rajnoha and P. Pollák, "ASR systems in noisy environment: Analysis and solutions for increasing noise robustness," Radioengineering, vol. 20, no. 1, April 2011.
[4] P. Saini, P. Kaur and M. Dua, "Hindi Automatic Speech Recognition Using HTK," Int. J. Eng. Trends Technol., vol. 4, no. 6, pp. 2223–2229, June 2013.
[5] H. Kaur and R. Bhatia, "Speech Recognition System for Punjabi Language," vol. 5, no. 8, pp. 566–573, 2015.
[6] S. Narang and M. Divya Gupta, "International Journal of Computer Science and Mobile Computing Speech Feature Extraction Techniques: A Review," Vol.4 Issue.3, March- 2015, pg. 107-114.
[7] Namrata Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition," International Journal for Advanced Research in Engineering and Technology Volume 1, Issue VI, July 2013.
[8] C. Shekhar, "Speech Recognition System for English Language," Int. J. Adv. Res. Comput. Commun. Eng., vol. 2, no. 1, pp. 919–922, 2013.
[9] G. Sivakumar, "Speech Recognition for Hindi." (Unpublished)
[10] U. Shrawankar and V. M. Thakare, "Techniques for Feature Extraction In Speech Recognition System : A Comparative Study," 2013. (Unpublished)
[11] Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk, "Speech Recognition using MFCC", international Conference on Computer Graphics, Simulation and Modeling (ICGSM' 2012) July 28-29, 2012 Pattaya (Thailand) III.

