# A Review on Automatic Speech Recognition System in Indian Regional Languages

Siddharth S. More
Dept. Of Computer Science &IT
Dr.Babasaheb Ambedkar
Marathwada University,
Aurangabad (MS) India

Prashantkumar L. Borde,
PhD
Marathwada Institute of
Technology CIDCO,
Aurangabad (MS) India

Sunil S. Nimbhore, PhD
Dept of Computer Science & IT
Dr. Babasaheb Ambedkar
Marathwada University,
Aurangabad (MS), India

## ABSTRACT
Speech Recognition is the system whose allows a user to use their voice in the form of input data. It may be used to command text to the computer and give order to the computer system. Speech technologies are commonly used available for an unlimited but most important range of tasks. Older speech recognition application needs to identify each single world by the different phases. This process allow to the machine to conclude where one word begins and the next word stops. This type of speech recognition application are still used to direct to the computer's system. And operate applications like web browser and spread sheets. New speech recognition system allow a user to order text fluently into the system. The system that allow continuous speech are generally designed to recognize text and format it, rather than controlling the computer system itself.in This research paper studied various types of techniques which is mostly used in automatic speech recognition.

## Keywords
Automatic Speech Recognition, Speech Processing, Signal Processing, Pattern Recognition

## 1. INTRODUCTION
Speech is a prime medium of communication between two people for express their thoughts, ideology, behaviors, feelings, by the mean of communication i.e. speech or any other complex movement. In this emerging era of machine and technology automatic speech recognition has attracted a great deal of attention over the past five decades. Speech recognition can be defined as the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. Speech recognition by machine means understanding voice of a person by the computer and performing any required task or the ability to match a voice against a provided or acquired vocabulary. The task is to design a system for the computer to understand spoken language, i.e. to react appropriately and convert the input speech into the form of desired output as commended. Now a day's most appliances have some sort of electronic or computer control, this feature prepares the way to realize the goal of use of speech recognition system in such appliances. Due to the rapid development in this field all over the world speech recognition systems have been implemented in a variety of applications, most eminent automated caller systems, automated information systems, speech recognition systems converting speech to text etc. are extensively used technology of today's world. These devices perform various tasks from simple user voice command. A speech recognition system consists of a microphone, for the person to speak into; speech recognition software; a computer to take and interpret the speech; a good quality soundcard for input and/or output; a proper and good pronunciation. There is Projected number of languages in the world varies between 6,000 and 7,000. In India there are 22 official languages out of which the work on speech recognition system is done only on 14 languages so far. In This day's Speech is best way to express everything. There are 15 million human beings stutter, most of those who's come fast in the adult age. This is similar age of 12 years old children. They are commonly trouble with this problem. Most important things is that the speech disorders also may occur in the physical disabilities children. The speech recognition system is the major aspect of the computer system.

In a speech recognition system used microphone for person to speak or getting data from the speaker; a computer to take and interpret the speech; a good quality soundcard for input and/or output; a proper and good pronunciation [1]. There is Projected number of languages in the world varies between 6,000 and 7,000. In India there are 22 official languages out of which the work on speech recognition system is done only on 14 languages so far [2]. The first step in any automatic speech recognition system is to capture data from the speaker then extract features i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion etc. [3]. The main point to understand about speech is that the sounds generated by a human are filtered by the shape of the vocal tract including tongue, teeth etc. This shape determines what sound comes out. If determine the shape accurately, this should give an accurate representation of the phoneme being produced [4]. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope. Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since [5]. This paper provides an overview of speech recognition system and the review of techniques available at various stages of speech recognition in Indian Languages.

The paper is organized as follows. Section 2 presents the type of speech with speaker model, section 3 explains about the framework of ASR with feature extraction and classification techniques. Section 4 investigates the Review of Speech Recognition Works done in Indian Languages. Finally, the conclusion is summarized in section 5 with acknowledgement

## 2. TYPE OF SPEECH
Speech recognition system can be separated in different classes by describing what type of utterances they can recognize.

## A. Type of Speech Utterance

### 2.1 Isolated Word

An isolated-word system operates on single words at a time - requiring a pause between saying each word. This is the simplest form of recognition to perform because the end points are easier to find and the pronunciation of a word tends not affect others.

### 2.2 Connected Word

It is similar to isolated word recognition, this mode allows several words to be run together with minimal pausing between them. Longer phrases are therefore possible to recognize, and the necessary computation increases as a result

### 2.3 Continuous Speech

A continuous speech system operates on speech in which words are connected together, i.e. not separated by pauses. Continuous speech is more difficult to handle because of a variety of effects. First, it is difficult to find the start and end points of words.

### 2.4 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 feature such as words being run together.

## B. Type of Speaker Model

All speakers have their special voices, due to their unique physical body and personality. Speech recognition system is broadly classified into two main categories based on speaker models namely speaker dependent and speaker independent.

### 2.5 Speaker Dependent Models

Speaker dependent systems are trained by the individual who will be using the system. These systems are capable of achieving a high command count and better accuracy for word recognition. The drawback to this approach is that the system only responds accurately only to the individual who trained the system.

### 2.6 Speaker Independent Models

Speaker independent is a system trained to respond to a word regardless of who speaks. Therefore the system must respond to a large variety of speech patterns, inflections and enunciation's of the target word. The command word count is usually lower than the speaker dependent however high accuracy can still be maintain within processing limits.

## 3. FRAMEWORK OF ASR

The typical ASR system accepts the audio input as shown in figure 1. The audio input is captured with the help of standard audio mic. Once the input is acquired, it will be preprocessed for acoustic feature extraction and further used for recognition of utterance.

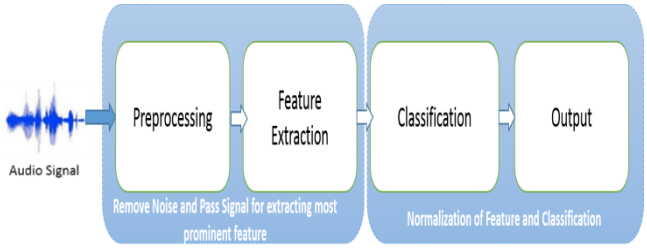


**Figure 1 show the framework of Speech Recognition System.**

The above figure shows the framework for developing robust speech recognition system. First step to collect the speech database, then pass it to preprocessing. The preprocessing part includes cleaning of signal and removal of silence which exists in the signal before utterance and after utterance. This procedure return the absolute signal representing only user utterance corresponding to isolated word. Then preprocessed acoustic signal will pass for feature extraction. MFCC, DTW are mostly used for acoustic features extraction. Mel-Frequency Cepstral Coefficients (MFCC) approach is the most popular and widely accepted, due to its spectral base as parameters for recognition. Then classification purpose Hidden Markov Model (HMM) and ANN approached was used [6] [7].

### 3.1 Speech Feature Extraction Techniques

Feature Extraction is the most important part of speech recognition since it plays an important role to separate one speech from other. Because every speech has different individual characteristics embedded in utterances. These characteristics can be extracted from a wide range of feature extraction techniques proposed and successfully exploited for speech recognition task. The most widely used feature extraction techniques are shown in below table 1.

**Table 1. Table show some feature extraction techniques in ASR**

| ***Sr. No.*** | ***Method*** | ***Database*** | ***Result*** |
|---|---|---|---|
| 1. | Hidden Markov Modeling (HMM), Linear Predictive Coding (LPC), Genetic Algorithm approach, Vector Quantization (VQ) [8]. | 5 Words Uttered by Females. | 81.82 % |
| 2. | Mel Frequency Cepstral Coefficient (MFCC), Hidden Markov Model Toolkit (HTK) [9]. | 3 Male and 2 Female Speaker with 4 Utterances | 80% |
| 3. | MFCC, Dynamic Time Warping (DTW)[10] | 10 Isolated Words | 94.85 % |
| 4. | Mel-Frequency Cepstral Coefficients(MFCC) , (MFCC) and Dynamic Time Warping (DTW) [11]. | 100 Isolated Words, 100 Native Speakers with 3 Utterances | 89% |
| 5. | Mel frequency cepstral coefficients (MFCC), Dynamic time warping (DTW) [12]. | 72 Marathi Words spoken by age group between 20-25 years | 72.22 % |

| | | | |
|---|---|---|---|
| 6. | Mel Frequency Cepstral Coefficients (MFCC), KNN [13]. | 80 Hindi words, 10 speaker and 3 utterances of each words | 89% |
| 7. | Mel Frequency Cepstral Coefficients, Hidden Markov Modeling Tool Kit −3.4[14] | 1806 Hindi utterances taken from18 Males and 12 Females Speaker | 79.11 % |
| 8. | Hidden Markov Model(HMMs), Hidden Markov Model, Toolkit (HTK), HMM, Mel Frequency Cepstral Coefficient (MFCC)[15] | Total 13th Isolated Hindi words used. | 95.49 %. |
| 9. | Mel Frequency Cepstral Coefficient (MFCC), HMM, HTK. [16] | 115 distinct words<br>Total of 2760 Samples got from 115 spaker (115*3*8) | 94.08% |
| 10. | HMM; HTK; Mel Frequency Cepstral Coefficient (MFCC).[17] | 30 Hindi words | 94.63%. |

# 4. REVIEW OF SPEECH RECOGNITION WORK

In this section present the review of speech recognition systems for different Indian languages.

## 4.1 Hindi Language

Sharmila et al (2012) [18] describe isolated Hindi Digit recognition system using Hidden Markov Model. The developer train 10 Hindi Digit from twenty four speaker with sample rate 16KHZ, so the total volume to database was 2400 samples. LPC, PLP and wavelet base feature extraction techniques used for feature extraction and pass these feature for classification. 200 samples of each user was used for training set and 40 samples used for testing set, HMM was used for classification purpose and achieved 66%, 77.3% and 89.85% recognition rate using LPC, PLB, and db10 wavelet features,

Preeti saini et al (2013) [19] researcher have built a speech recognition system for Hindi vowel and consonants. Data captured using Sony F-V120 Microphone and distance between speaker mouth and microphone was 5-10cm. Total 113 samples used for training set. For feature extraction 39 order MFCC techniques used 12 melcepstrum plus log energy and their first and second order derivatives feature pass for classification. HMM used for classification and got 95.49% result.

Ankit kumar et al (2014) [20] compared the performance of continuous Hindi speech recognition system with different vocabulary sizes and feature extraction techniques. For better feature extraction both Mel- frequency Mel Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) this both feature extraction are used. Hidden Markov Model (HMM) is used at backend of an Automatic Speech Recognition system for mono phone based acoustic modelling. This System was implemented using HTK 3.4.1 toolkit and got 95.08% accuracy.

Babita Saxena et al (2015) [21] studies they stated a standard digits recognizer for Hindi language. Total 10 speaker from 20 to 40 year age group selected for data accusation. Data captured using Sony Xperia L headset, out of 10 speaker 8 sperker data used for traning set and 2 speaker data used for testing set. Feature are extracted using MFCC techniques, HMM used for classification purpose and achieved 86.17% accuracy.

## 4.2 Sanskrit Language

Jitendra Singh Pokhariya et al. (2014) [22] describe a work done for building a speech recognition system for Sanskrit language. The system was trained using 50 Sanskrit words, total 10 speaker used for data acquisition. 5 state HMM topology used for classification and got 95.2% result and for 10 states achieved 97.2% result.

## 4.3 Ahirani Language

Patil A. S. (2014) [23] describe the implementation of HMM based speaker independent Ahirani Speech recognition System. 20 Ahirani words collected by 10 speaker, MFCC technique were used for feature extraction and HMM used for classification. The experimental result show the 94% performance of system.

## 4.4 Assamese Language

Himanshu sarma et al (2014) [24] described automatic transcription of Assamese speech using HTK toolkit. For database creation, SONY ICD-UX533F microphone was used. Total 27 speaker contributed to develop the Assamese database, out of 27 speaker, 14 are male and 13 are female speaker with an age group 20 to 40 years. 527 samples and 127 samples were used for training and testing set and pass these data for feature extraction using MFCC techniques. Using HTK toolkit they achieved 65.65% accuracy.

## 4.5 Punjabi Language

Kumar Ravinder (2010) [25] has worked for development of isolated Punjabi speaker dependent system. He has extended his work up to comparison of speech recognition system for small vocabulary of Punjabi Language. Total 500 Isolated Punjabi words vocabulary used for design the ASR system. As Punjabi language gave us the changes between consecutive phonemes detection of end point with high difficultly. So LPC with Dynamic Programming computation and Vector quantization techniques used for feature extraction and HMM and DTW used for classification. Using HMM and DTW techniques he has achieved 91.3% and 94.0% accuracy respectively.

Mohit Dua et al (2012) [26] implement an ASR system using Isolated words of Punjabi Language. For database collection fourteen speaker were contributed. Out of fourteen speaker eight speaker used for training set and six speaker data used for testing set. For testing module GUI was implemented using JAVA platform to make the system faster. MFCC and HTK were used for feature extraction and classification of

data respectively. Using HTK toolkit for classification he has achieved 9563% accuracy.

## 4.6 Tamil Language

K. Murali Krishna et al (2014) [27] they had state that speech feature vector which was generated by projecting an observed vector on to an Independent Component Analysis (ICA) & Principal Component Analysis (PCA).for isolated Tamil words speech recognition has evaluated by the new feature. The method was used is provide higher accuracy of speech recognition than conventional method of clean environment. They used HTK to build it. The performance of system using MFCC feature was in range of 87 % to 88% with word error rate 12 % to 13%,

## 4.7 Gujarati Language

Jinal H. Tailor et. al. (2016) [28] in their research paper state the architecture of ASR for the Gujarati language. And the database used for training purpose to collect from & 4 male and 2 female. They belongs to 18 to 36 age group. To measure performance & error parameter the authors used Hidden Markov Model Toolkit. The implemented analyzes WER (word recognition rate) 95.9% & WER (word recognition rate) as 5.85% in lab environment in noisy environment calculated WR was 95.1% & WER from 7.40%.

## 4.8 Telugu Language

Surabhi Sreekanth et al (2005) [29] describe the text dependent system for Telugu language which design for low security access the control system. These system was used to recognize the spoken password and conform the identity of user. For database corpus creation, total 7 speaker were contributed with 20 utterances of each speaker. To design these system MFCC feature extraction techniques were used. For classification purpose Mahalanobis distance technique used and achieved 98.85% correct classification.

## 4.9 Manipuri Language

Rahul. L. et al (2013) [30] have discussed in their research paper about implementation of phoneme. They have used HMM tool kit (HTK), version3.4 the better implementation the system. A five state Hidden Markov Model (HMM) left to right with 32 mixture continuous density diagonal covariance Gaussian Mixture Model (GMM) per state was used to build a model for single phonetic unit. For the developing also the database data of around 5 hr read was collected from 4 male and 6 female speakers. Also for analyzing the system performance continuous speech data it was collected from 5 males and 8 females. Total 69 words were chosen for the database. Using those chosen words sentences were framed for the purpose of recognition those keywords by the system. For transcription of the data the symbols of International Phonetic Alphabet (IPA) were used. An overall performance the system has shown after analysis was 65.24%.

## 4.10 Kannada Language

G. Hemakumar et al (2013) [31] designed the Kannada speaker independent Isolated word recognition system. For signal normalization and feature extraction Linear-Predictive Coding (LPC) techniques were used and LPC coefficient were selected for classification.

# 5. CONCLUSION

In this research paper has discussed some of the papers which are related to automatic speech recognition systems. The most researcher use HTK toolkit for different Indian regional Language. This research paper has also compared all those system on the basis of their language, year type of utterance, number of speakers, utterance of each word recording environment, number of words / sentences, feature extraction technique used, word accuracy, number of state in HMM, Word Error Rate (WER). Indian official language the work done is less as compared to other foreign languages. The HTK toolkit have high accuracy rate as compared to other techniques. This study will motivate people for developing automatic speech recognition (ASR) system using HTK toolkit for different languages.

# 6. ACKNOWLEDGEMENT

Authors would like to thank the Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University authorities for providing infrastructure to carry out the experiments. This work is supported by BARTI.

# 7. REFERENCES

[1] Saini, Preeti, and Parneet Kaur. "Automatic speech recognition: A review." International Journal of Engineering Trends and Technology 4, no. 2 (2013): 1-5.

[2] P. P. Shrishrimal, R. R. Deshmukh, Vishal B. Waghmare, “Development Of Isolated Words Speech Database Of Marathi Words For Agriculture Purpose”, Asian Journal of Computer Science And Information Technology 2: 7 (2012) 217 – 218.

[3] Yu, Dong, and Li Deng. AUTOMATIC SPEECH RECOGNITION. SPRINGER LONDON Limited, 2016

[4] Nisha V. S, M. Jayasheela, Speaker Identification Using Combined MFCC and Phase Information, IJARCCE, Feb. 2013, Vol.2.

[5] Louis-Marie Aubert, Roger Woods, Scott Fischaber, and Richard Veitch “Optimization of Weighted Finite State Transducer for Speech Recognition”, IEEE Transactions on Computers, Vol. 62, No. 8, August 2013.

[6] M. Wiśniewski, W. Kuniszyk-Jóźkowiak, E. Smołka, and W. Suszyński, "Automatic Detection of Disorders in a Continuous Speech with the Hidden Markov Models Approach", in Computer Recognition Systems 2, vol. 45, pp. 447-453, 2007.

[7] M. Wiśniewski, W. Kuniszyk-Jóźkowiak, E. Smołka, and W. Suszyński, "Automatic Detection of Prolonged Fricative Phonemes with the Hidden Markov Models Approach", Journal of Medical Informatics & Technologies, vol. 11,pp.293-298, 2007.

[8] Shinde MB, Gandhe DS. Speech processing for isolated Marathi word recognition using MFCC and DTW features. International Journal of Innovations in Engineering and Technology. 2013 Oct; 3(1).

[9] Saksamudre SK, Deshmukh RR. Comparative Study of Isolated Word Recognition System for Hindi Language. In International Journal of Engineering Research and Technology 2015 Jul 20 (Vol. 4, No. 07, July-2015). ESRSA Publications.

[10] Gaurav, Gaurav, Devanesamoni Shakina Deiv, Gopal Krishna Sharma, and Mahua Bhattacharya. "Development of application specific continuous speech recognition system in Hindi." (2012).

[11] Saini P, Kaur P, Dua M. Hindi Automatic Speech Recognition Using HTK. International Journal of Engineering Trends and Technology. 2013 Jun; 4.

[12] Khetri GP, Padme SL, Jain DC, Fadewar DH, Sontakke DB, Pawar DV. Automatic Speech Recognition for Marathi Isolated Words. International Journal of Application or Innovation in Engineering & Management (IJAIEM). 2012 Nov; 1(3):69-74.

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

[14] Bansod, Nagsen S., et al. "Speaker Recognition Using Marathi (Varhadi) Language." Intelligent Computing Applications (ICICA), 2014 International Conference on. IEEE, 2014.

[15] Suhas Mache1, C. Namrata Mahender "Development of Text-to-Speech Synthesizer for Pali Language", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I (May-Jun. 2016), PP 35-42.

[16] Dua, Mohit, R. K. Aggarwal, Virender Kadyan, and Shelza Dua. "Punjabi automatic speech recognition using HTK." IJCSI International Journal of Computer Science Issues 9, no. 4 (2012): 1694-0814.

[17] Dua, Mohit, R. K. Aggarwal, Virender Kadyan, and Shelza Dua. "Punjabi automatic speech recognition using HTK." IJCSI International Journal of Computer Science Issues 9, no. 4 (2012): 1694-0814.

[18] Sharmila, Dr. Neeta Awasthy, Dr. R. K. Singh, "Performance of Hindi Speech Isolated Digits In HTK Environment", IOSR Journal of Engineering May. 2012, Vol. 2(5) pp: 1020-1023.

[19] Preeti Saini, Parneet Kaur, Mohit Dua, "Hindi Automatic Speech Recognition Using HTK", International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue6- June 2013.

[20] Ankit Kumar, Mohit Dua, Tripti Choudhary,"Continuous Hindi Speech Recognition Using Monophone based Acoustic Modeling", International Journal of Computer Applications® (IJCA) (0975 – 8887) International Conference on Advances in Computer Engineering & Applications (ICACEA-2014) at IMSEC,GZB.

[21] Babita Saxena, Charu Wahi,"Hindi Digits Recognition System on Speech Data Collected In Different Natural Noise Environments", International Conference on Computer Science, Engineering and Information Technology (CSITY 2015) February 14~15, 2015, Bangalore, India. Volume Editors: David C. Wyld, Jan Zizka ISBN : 978-1-921987-31-1.

[22] Jitendra Singh Pokhariya, D r. Sanjay Mathu r, "Sanskrit Speech Recognition using Hidden Markov Model Toolkit", International Journal of Engineering Research & Technology (IJERT) Vol. 3 Issue 10, October- 2014.

[23] Ajay S. Patil, "Automatic Speech Recognition for Ahirani Language Using Hidden Markov Model Toolkit (HTK)", International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 3, May-Jun 2014.

[24] Sarma, Himangshu, Navanath Saharia, and Utpal Sharma. "Development of Assamese Speech Corpus and Automatic Transcription Using HTK." Advances in Signal Processing and Intelligent Recognition Systems. Springer International Publishing, 2014. 119-132.

[25] Kumar Ravinder, "Comparison of HMM and DTW for Isolated Word Recognition System of Punjabi Language", Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications ,Volume 6419 of the series Lecture Notes in Computer Science pp 244-252, 2010.

[26] Mohit Dua, R. K. Aggarwal, Virender Kadyan, Shelza Dua, "Punjabi Automatic Speech Recognition Using HTK",IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012.

[27] K.Murali Krishna, M.Vanitha Lakshmi, "Speaker Independent Isolated Tamil Words for Speech Recognition using MFCC, IPS and HMM", International Journal of Scientific & Engineering Research, Volume 5, Issue 4, April-2014.

[28] Jinal H. Tailor, Dipti B. Shah, "Speech Recognition System Architecture for Gujarati Language", International Journal of Computer Applications (0975 – 8887) Volume 138 – No.12, March 2016.

[29] Surabhi Sreekanth, Kavi Narayana Murthy, "TextDependent Speaker Recognition System for Telugu" Osmania Papers in Linguistics, Vol.31, 2005, 84-99.

[30] Rahul, L.; Nandakishor, S.; Singh, L.J.; Dutta, S.K., "Design of Manipuri Keywords Spotting System using HMM," in Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013 Fourth National Conference on , vol., no., pp.1-3, 18-21 Dec. 2013.

[31] G. Hemakumar, P. Punitha, "Speaker Independent Isolated Kannada Word Recognizer", Multimedia Processing, Communication and Computing Applications Volume 213 of the series Lecture Notes in Electrical Engineering pp 333-345, Date: 26 May 2013.