ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

HINDI SPEECH RECOGNITION TECHNIQUE USING HTK

Nikita Dhanvijay *, Prof. P. R. Badadapure
Department of Electronics and Telecommunication
JSPM'S imperial college of engineering
Wagholi, pune - 412207, India

DOI: 10.5281/zenodo.203964

ABSTRACT

This paper is based on SR system. In the speech recognition process computer takes a voice signal which is recorded using a microphone and converted into words in real-time. This SR system has been developed using different feature extraction techniques which include MFCC, HMM. All are used as the classifier. ASR i.e. automated speech recognition is program or we can called it as a machine, and it has ability to recognize the voice signal (speech signal or voice commands) or take dictation which involves the ability to match a voice pattern opposite to a given vocabulary. HTK i.e. The Hidden Markov model Toolkit is used to develop the SR System. HMM consist of the Acoustic word model which is used to recognize the isolated word. In this paper, we collect Hindi database, with a vocabulary size a bit extended. HMM has been implemented using the HTK Toolkit.

KEYWORDS: HMM (hidden markov model), ASR (Automatic Speech Recognition), Speech recognition (SR). MFCC (mel frequency cepestral coefficient)

INTRODUCTION

SR System is usually implemented in the form of dictation of software and intelligent assistant in personal computer, smart phones, and web browsers and may other device. The speech is cheapest source and primary mode of communication among human being and also the most natural and effective form of exchanging information among human in speech. The human speech has discriminative features which are utilized to identify speakers. Speech contains significant energy from zero frequency up to around 5 kHz. Aim of ASR is to characterize, extract and recognize the information about speaker identity. Speech signal is called as quasi-stationary.

To explain the meaning of speech, first we need to identify the components of spoken words and phonemes act as identifying markers within speech. An algorithm used to explain the speech further. The HMM is a commonly used as a mathematical model which is used to do this. Firstly, the large database of phonemes has to be create to create a SR engine. When a comparison is performed in system, most likely or closely matches are determined between the spoken phoneme and the stored one, and further computations are performed.

TYPES OF SPEECH RECOGNITION

SR Systems can be divided into the number of classes that based on their ability to recognize list of words they have. The speech recognition system can be classified into four types as:

Isolated Speech

Isolated words involve a pause between the two utterances; it doesn't mean that it only accepts a single word, but instead of it requires only one utterance at a time [4].

Connected Speech

Connected speech is another type of speech recognition system. It is similar to isolated words or speech, but it allows separate utterances with minimal pause between them.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Continuous speech
When user speak naturally it called as continuous speech, therefore, called as the computer dictation

Spontaneous Speech
Spontaneous speech has the ability to handle variety of natural speech features such as words being run together, "ahs" and "ums", and even slight stutters

Modes of Communication

Robots can communicate with human with the help of different techniques. Following are the modes of communication, shown in Fig.1.

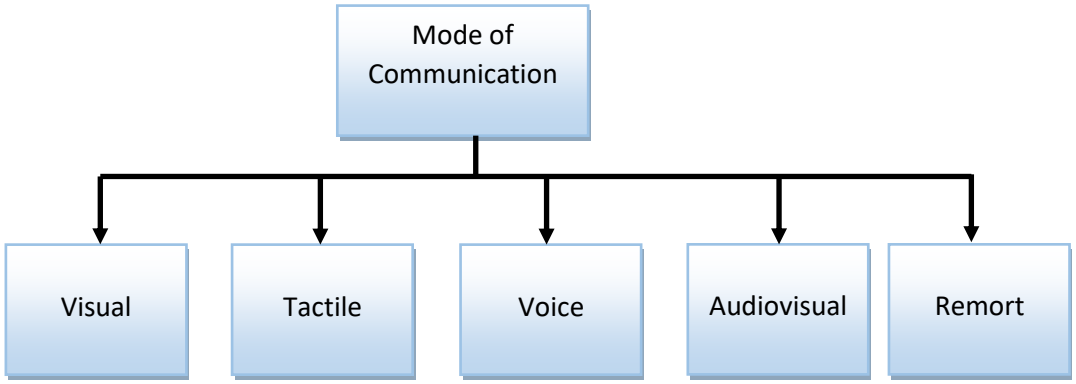

Fig. 1: Modes of Communication

Visual
Visual based communication is gesture based communication which is non-verbal and in this we use the movement of body parts to interact. Gestures can be of any type, like head gesture, hand gesture, even facial expressions etc.

Tactile
It has two types [11] first is a tactile screen sensing and the second is tactile skin sensing. Tactile skin sensing is skin sensing, tactile screen sensing is used in the field of Robotics. And in the tactile screen examine when the force is being given, as well as in what amount.

Voice
Voice is the easiest and simplest medium which is used for communication. The Human vocal track is the source of vocal voice signal.

Audiovisual
Audiovisual is a single or multiple way of communication. It includes Speech based and Gesture based communications are the most important among them, which is the concentration of the researchers these days.

LITERATURE REVIEW

Tarun Pruthi (2000) is represented a speaker-dependent, and isolated word recognizer for Hindi. Feature extraction has been done using MFCC and this system was implemented using HMM (hidden markov model). He recorded 2 male speakers sound and the vocabulary consists of Hindi digits like 0, which can be pronounced in Hindi as “shoonya” and to 9 is “nau”. So system showing good result, but the design of the system is speaker dependent and size of the vocabulary is very small.

Gupta (2006) worked on the isolated voice commands, SR tool for Hindi language. They use continuous hidden markov model. An acoustic word model was also used for recognition. The word vocabulary consists of digits of Hindi language. The recognition Results were satisfactory when it is tested for speaker dependent model.

Similarly the results are satisfactory for other sounds too. Main disadvantage is that vocabulary size is very small.

Anup Kumar Paul, Dipankar Das (2009) developed SR System for BANGLA Language by using LPC and ANN. This paper represented recognition system of the Bangla speech. It has two major parts of Bangla SR system. Signal processing is the 1st part and the 2nd part is speech pattern recognition technique. Beginning point and end point are detected during the speech processing stage. Speech Pattern recognition is done in the 2nd part Artificial Neural Network. Voice signals, i.e. Speech signals recorded with the help of wave recorder and room environment are normal**.**

Al-Qatab (2010) developed Arabic words, ASR machine using HTK. The machine recognized both continuous speech and isolated speech. A Machine used Arabic dictionary which is built manually and used 13 speakers, and vocabulary size of thirty three words.

R.L.K. Venkateswarlu, R. Ravi Teja et al (2010) Developed Efficient speech recognition System for Telugu language. In this research, they used both MLP (Multilayer Perceptron) and TLRN (Time Lagged Recurrent Neural Network) models were trained and tested on a dataset, it consists of Four different speakers to Male and two Females are allowed to utter the letters for 10 times. This is Speaker dependent mode, which is used for recognition of the Telugu words.

R. Kumar (2010) Present a real-time experimental, speaker- dependent, isolated speech recognition system for Punjabi language. The performance of a recognition system for the small vocabulary size of words using the HMM and DTW (Dynamic Time Warp) technique. This work highlighted template-based recognizer approach with the help of LPC with dynamic programming computation and VQ (vector quantization) with HMM based recognizers in isolated SPEECH recognition tasks**.**

Ahmad A. M. Abushariah (2010) worked for English speech recognition. This paper focused on to design and implement English digits on the SR system by using MATLAB (GUI). This work was based on the HMM, MFCC technique was used for feature extraction. The paper focused on all the English digits from (Zero to Nine).
M Singh et al. (2011) described a SI (speaker independent), real time, an isolated Speech ASR system for the Punjabi language. It was developed by the Vector Quantization and Dynamic Time Warping (DTW) approaches were used for the recognition system. The database of the features (LPC Coefficients) of the trained data was produced for testing and training the system, the test pattern was compared with each reference pattern using DTW alignment. This system developed for small isolated word vocabulary**.**

Bharti W. Gawali1, Santosh Gaikwad et al (2011) describe a Marathi database and isolated SR system. MFCC and DTW are used for features extraction. The vocabulary included Marathi vowels and isolated words; it started with each vowels and simple sentences in Marathi. And voice of 35 speakers was recorded and each word with 3 repetitions.

Kuldeep kumar (2011) worked for ASR for Hindi language. This paper aims to build a SR System for Hindi language. The System is developed using Hidden Markov Model Toolkit (HTK). He made an Acoustic word model which recognizes the isolated words. The system is trained for 30 Hindi words and recorded data from eight speakers. The overall accuracy is noted as 94.63%**.**

K. Kumar [2012] worked on connected-words SR system for Hindi. HTK tool kit was used to develop the system which used for recognized the trained words or sentences.

METHODOLOGY

In this paper SR System has been implemented by the following steps:

We start with a data corpus (data collection) with a size of vocabulary is 45 Hindi words have been prepared. The voice samples are recorded from 20 speakers. Audacity tool is an audio voice recorder which used for Data Corpus preparation. After preparing a data corpus, the corpus is then used to train the system as well as to test it.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Finally, using the HTK Toolkit and the data corpuses prepared, the SR system is developed by applying MFCC, LPC as feature extraction techniques and HMM as a classifier. For parameterization purposes from wav files to MFCC or LPC as well as to implement HMM efficiently HTK Toolkit has been used.

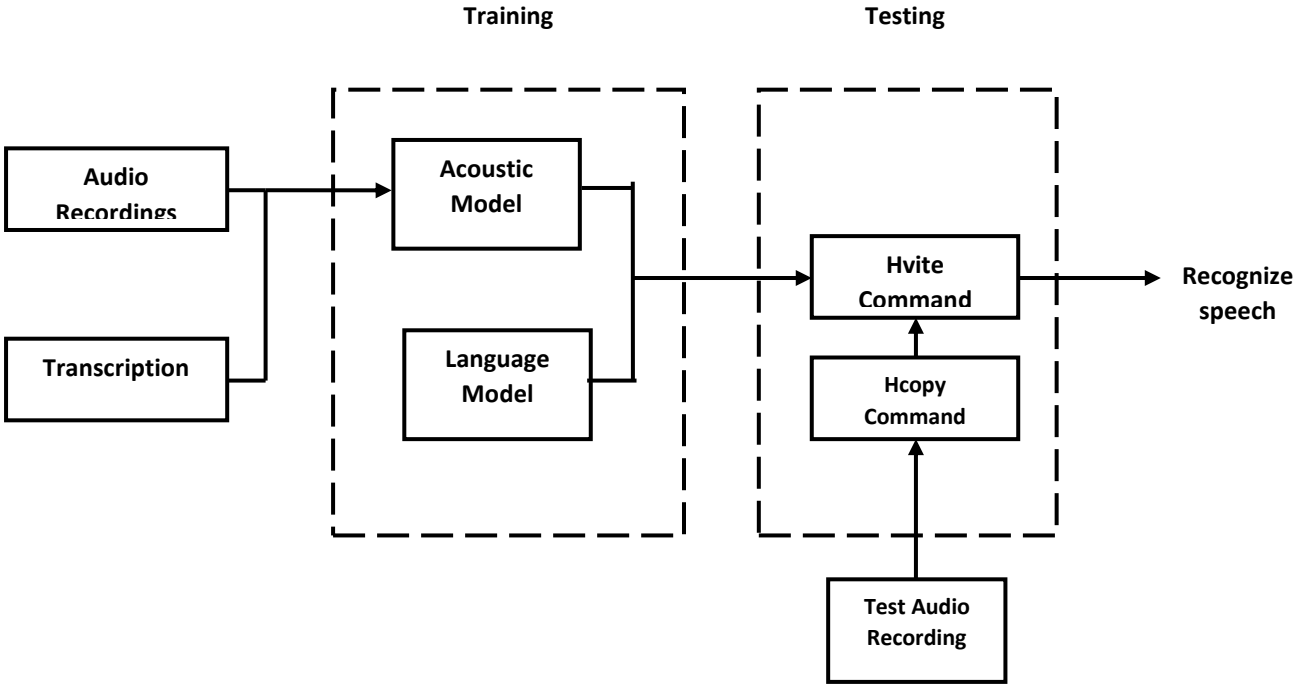

Fig. 2 Framework of proposed work

SPEAKER RECOGNITION TECHNIQUES:

Speaker recognition concentrates on the recognition i.e. identification task. The aim in speaker identification is to recognize the unknown speaker from a set of known speakers (closest Speaker identification). A speaker recognition system consists of below modules.

Front-end processing

In this module, front end system converts speech signal into feature vectors. Every speaker has its unique feature vector. Both training and testing phases are performed in the Front end processing

Speaker modeling

In this part, speaker modeling performs a reduction of feature data by modeling the distributions of the feature vectors.

Speaker database

The speaker database stored the feature vectors of each speaker. The number of speakers increases the size of the database increases.

Decision logic

It is also called as testing. In this module, the feature vector of unknown speaker is provided to the system. System recognized the speaker by best matching of unknown feature vector with database feature vector. .

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

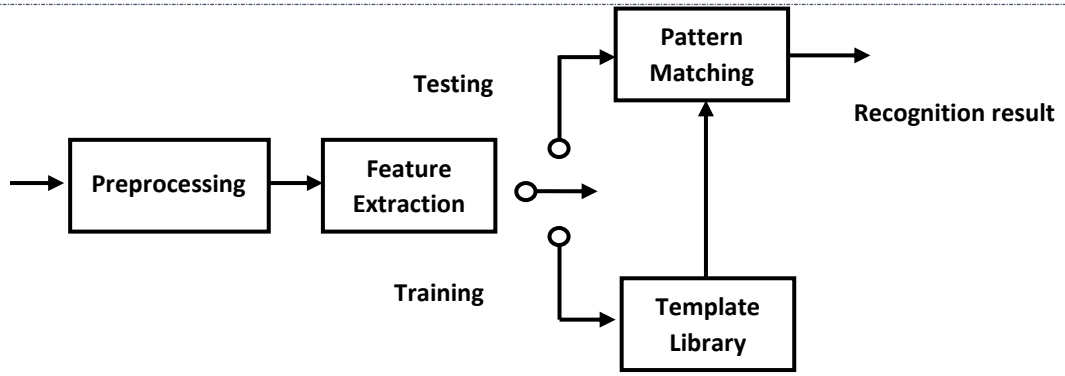

Fig. 3 Basic configuration of SR system

Before processing of speech signal it must be digitized, sampling of speech should be subordinated to the Nyquist sampling theorem i.e. the sampling frequency of speech signal should be greater than twice the highest frequency of the speech signal, and the collecting speech signal without loss of the original signal. After sampling of the speech signal, we can get the speech signal which is discrete in time domain and serial in amplitude, after that amplitude value of a signal into a finite interval, this process is called quantization, and then encoded the quantized values in the interval.

MFCC (Mel Frequency Cepstral Coefficients)

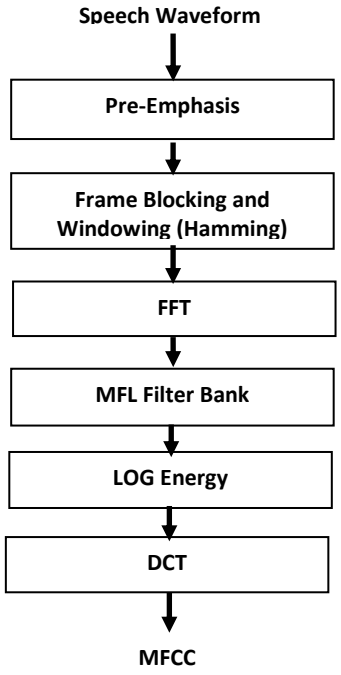

Fig 4 MFCC feature extraction

Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in speaker recognition and automatic speech recognition (ASR). Which are introduced by Davis and Mermelstein in the 1980's. In this paper Mel-Frequency Cepstrum Coefficients (MFCC) which we extract the cepstral envelop in which the formants shows the MFCC coefficients. The formants of a speech signal show the unique properties of a speech, using this speaker can be recognized the speech. For such reasons, a SR as well as speaker identification (SI) system here uses the concept of formants to identify the speakers, as well as recognize the speech.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Above Figure shows the overall process to extract the MFFC vectors from the speech signal. It gives special importance to the process of MFCC extraction is applied over each frame of speech signal independently. The MFCC vectors will be obtained from each speech frame after the pre-emphasis and the frame blocking and windowing stage. The 1st step of the MFCC extraction process is to compute the Fast Fourier Transform (FFT) of each frame and obtain its magnitude. At the end of the extraction process of MFCC is to apply the modified DCT to the log-spectral-energy vector, obtained as input by the mail filter bank, resulting in the desired set of coefficients called MFCC

HMM:

Markov property is important to understand hidden Markov model. Some of the systems exhibit the property in which the future states of the system is dependent on the present state of the system. This property is known as Markov property and the systems which exhibit such a property are called Markov Process and model is called Hidden Markov Model. The future state of the system is predicted by Viterbi Algorithm. In other hand Baum-Welch Algorithm sets the initial parameters such as initial states, initial transition probabilities etc. of hidden markov model. The performance of the system tested in two different types of environments, Speaker Dependent environment and Speaker Independent environments. Firstly, system performance using MFCC and LPC in speaker dependent as well as speaker independent has been computed. And finally, all of them have been compared with each other.

Hidden markov model training

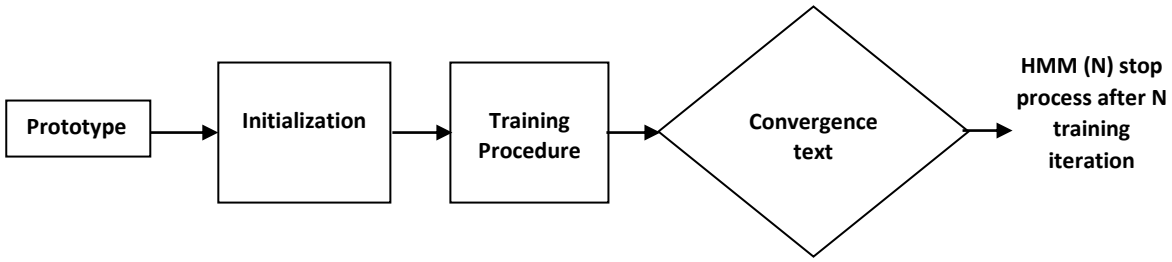

Fig 5 Hidden Markov Model training process

Initialization

Before the training procedure, we should make the training algorithm is accurate and fast. Parameters of HMM must be properly initialized. Hidden markov tool kit (HTK) provides two different initialization tool: first is H_{Init} and another one is H_{Compv}. HInit is used to reads all initial training data and cut out all the phonemes; if initialization data is none, then "unified start (flat start)" method is used. Therefore, all of phoneme models will be initialized to the same parameter values, all variance and state mean are equal to the mean and variance of the global voice signal. We use H_{compv} to find out the variance and mean of the global. In these experiments, H_{init} model is used to initialize the model parameters.

Training

In this H_{Rest}, tool of HTK It is function to estimate, it estimates that HMM parameter's optimum value. Every time, it displayed HRest iteration, when it has converged we can use it, once the measure of value i.e. absolute value is no longer decreasing, this process should be stopped.

HTK

It is a portable toolkit for building and manipulating HMM. Basically used for speech recognition. HTK is working on windows and Linux. We can create acoustic models with the help of this toolkit.

AUDACITY

Audacity is used for audio recording. It is free, easy-to-use, multi-track audio recorder. Audacity supports Windows, OS X, Linux and other operating systems.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

EXPERIMENTAL RESULTS AND ANALYSIS

Audacity is used for data corpus (collection) preparation. For system development, HTK Toolkit is used to implement HMM efficiently. A train database with sound samples of 10 words with 3 repetition samples per word has been created. The voice samples have been taken from 5 speakers. Thus, the train database consists a total of (10 x 3 x 5 = 150) sound samples.

Result

_================ HTK Results Analysis ==

Ref: words.mlf

Rec : recout.mlf

---------- Overall Results ----------

SENT: %Correct=36.00 [H=18, S=32, N=50]

WORD: %Correct=78.67, Acc=78.67 [H=118, D=0, S=32, I=0, N=150]

==========

REFERENCES

In this proposed system, the speech recognition system has been implemented using the HTK toolkit. The Hindi fruit name is taken as a query for the system. The experimentation is carried for two types such as sentence and word. The proposed system gives 36% correct rate for sentence level and 78.67% for word level.

REFERENCES

1. A. N. Kandpal and M. Rao, "Implementation of PCA and ICA for Voice Recognition and Seperation of Speech," in proc. of IEEE International Conference on Advanced Management Science (ICAMS), vol. 3, pp. 536-538, 2010.
2. M. A. Anusuya and S. K. Katti, "Mel Frequency Discrete Wavelet Coefficients for Kannada Speech Recognition using PCA," in Proc. of Int. Conf. on Advances in Computer Science, 2010.
3. Tarun Pruthi, Sameer Saksena and Pradip K Das, "Swaranjali: Isolated word recognition for hindi language using VQ and HMM," in proc. Of International conference on multimedia processing and systems, Aug. 2000.
4. D. Spiliotopoulos, I. Androutsopoulos, and C. D. Spyropoulos, "Human- Robot Interaction based on Spoken Natural Language Dialogue", in proc. Of the european workshop on service and humanoid robots.
5. A. A M Abushariah, T. S. Gunawan, O. O. Khalifa, and M. A. M. Abushariah, "English Digits SR System based on Hidden Markov Models," in Proc. IEEE International conference on computer and communication engineering, pp. 1-5, May 2010.
6. K. Kumar and R. K. Agarawal, "Hindi speech recognition using HTK," in International Journal of Computing and Business Research, vol. 2, May, 2011.
7. S. Young, et al., The HTK Book. December, 1995.
8. Gaikwad, S.K. and Gawali, B.A. A review on speech recognition technique. In International Journal of Computer Applications, volume 10, November, 2010.
9. H.-J. Bohme, T. Wilhelm, and J. Key. An approach to multi-modal human-machine interaction for intelligent service robots. Robotics and Autonomous Systems, elsevier science, 44:83–96 December2004.