# Pitch Frequency Based Speech Recognition System for Isolated Hindi Words

| Prof.Ujwala Patil | Suraj Ingole | Akshay Dhakne | Supriya Shedge | Yogeshwari Nikam |
|---|---|---|---|---|
| Professor(E&TC) | Student (E&TC) | Student(E&TC) | Student (E&TC) | Student (E&TC) |
| RSCOE | RSCOE | RSCOE | RSCOE | RSCOE |
| Pune, India | Pune, India | Pune, India | Pune, India | Pune, India |

***Abstract* :-** The concept of Speech Recogntion is to take an utterance of speech signal in the form of input, captured by microphone and convert it into a sequence of text which is as close as possible to the input which was represented by the acoustic data.In India, we have 22 distinct languages and 419 local languages. Majority of population is unaware of English language either in reading or writing but they can take advantage from resources of Information Technology sector if they support native idiom. Therefore, there is huge scope to develop such systems in Hindi language.The national language of OThe literacy rate of India is 74.04% amongst which only 12.18% of Indians are familiar with English language.Speech recognition is the inter-disciplinary sub-field of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers. It is also known as "automatic speech recognition"(ASR). The procedure includes creating of database of various speech signals , convert it to computer friendly language and display on a notice board.

## I. INTRODUCTION

Speech recognition was originated in 1940s and the first program for speech recognition was developed in 1952. This program was written for recogniting digits in noise less environment.The foundation of speech recognition which is automation and information models were worked on in this perios.Later in 1960’s, small set of isolated words were recognized using phoenetic properties of audios.Time normalization and filter banks were developed after this. In 1970’s, vocabulary of medium set of isolated words were recognized using pattern recognition.Then, in 1980’s large vocabulary of words were recognized where some speech recognition problems were faced.The most importatnt developed of this period is HMM.HMM stands for Hidden Markov Model which introduced facilities for recognition of continuos speech efficiently.

Speech Recognition system has now entered in market and is benifiting various users. Further challenge is to develop a machine which functions like a intelligent human.

## II. LITERATURE SURVEY

Suman K. Saksamudre et al. [12] proposed “Comparative Study of Isolated Word Recognition System for Hindi Language”.Author presented Isolated Word Recognition System for Hindi Language using MFCC as feature extraction and KNN as pattern classification technique. The system is trained for 10 different Hindi words. The experimental result of this system is that it gives 89% accuracy rate.

R .R. Deshmukh [8] developed isolated word recognition system for Hindi language. Here MFCC is used as feature extraction technique and KNN as pattern classifier. MFCC and KNN have given us 89% of recognition rate for 300 vocabulary data. Further ANN classifier can be used.

Neema Mishra et al. [18] in 2010 proposed overview of Hindi speech recognition. This paper describes how speech is produced and the properties of Hindi phoneme. This paper gives the basic idea of speech recognition system and basic information about Hindi acoustic –phonetics.

U. G. Patil et al. proposed " Automatic Speech Recognition of isolated words in Hindi language using MFCC".Author presented Isolated Hindi words recognition system which is a part of Automatic Speech Recognition (ASR) system. The main goal of ASR system is to understand a voice by computer or microphone and converts it into the text to perform required task. In this paper,MFCC as feature extraction technique, Vector Quantization (VQ) with GMM (Gaussian Mixture Model) for recognition of Hindi isolated words is used. For analysis, Hindi words speech dataset of different males and females speakers were also prepared.

## III. METHODOLOGY

The recording of the audios is a huge task. After the recording is completed , 3 processes take place.These are pre-processing , feature extraction and feature matching.

Pre-processing is the first step in which the audios are processed according to our requirements.Feature Extraction is extracting certain features such as MFCC,covariance,energy are extracted from the audios and stored.The features which are extracted in step 2 and matched against features which are stored in database. Generally pattern recognition algorithm is used for this process.The block diagram explaining the processes is shown in the figure below.

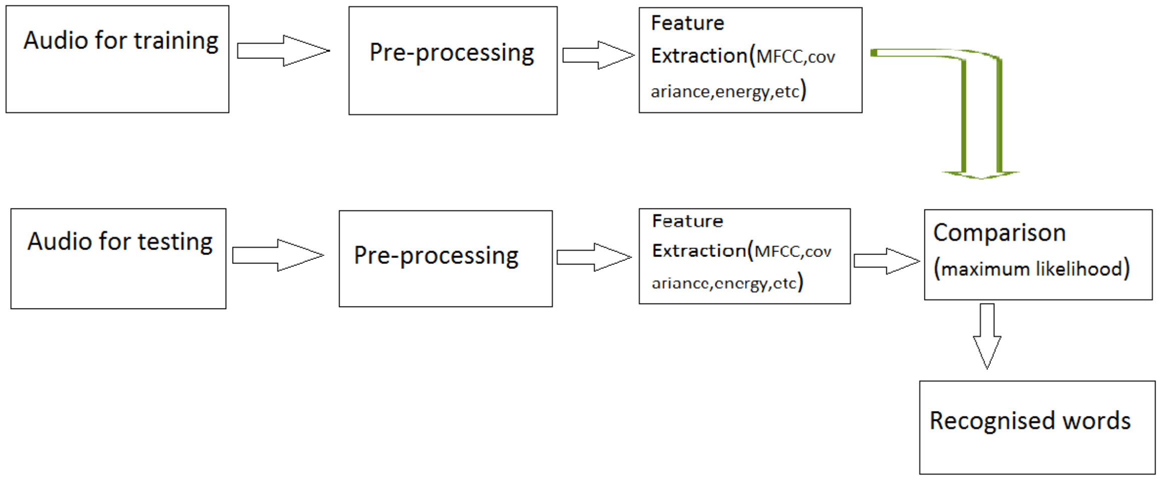
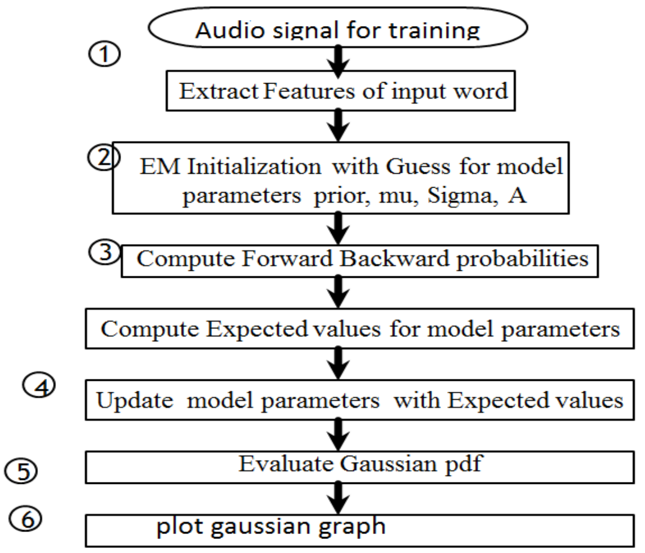

Audio for training
Pre-processing
Feature Extraction(MFCC,covariance,energy,etc)
Audio for testing
Pre-processing
Feature Extraction(MFCC,covariance,energy,etc)
Comparison (maximum likelihood)
Recognised words

Fig 1:Block Diagram of Speech Recognition

***Pre-processing*** :- For the production of speech,it is necessary that we boost the lower frequencies and suppress the higher frequency.Because of this loss of information is caused.To avoid this loss and retain the features of signal,an high pass FIR filter is used. This is done before speaker verification and speech recognition system.

***Feature Extraction*** :- The performance of the ASR system is dependent on the feature extraction. Here we are using pitch frequency as parameter for feature extraction. In this step, peak frequencies of each audio sample are selected and used for comparisons. These frequencies are stored in the database along with the word labels respectively. This step is necessary both while training and testing of samples.

***Comparison :-***The extracted features of both training and testing data are compared using the log likelihood values of the samples. The Maximum Log Likelihood is used and the samples having nearly same maximum likelihood values are considered as similar and the word labels are assigned accordingly.


Audio signal for training
①
Extract Features of input word
②
EM Initialization with Guess for model parameters prior, mu, Sigma, A
③
Compute Forward Backward probabilities
Compute Expected values for model parameters
④
Update model parameters with Expected values
⑤
Evaluate Gaussian pdf
⑥
plot gaussian graph

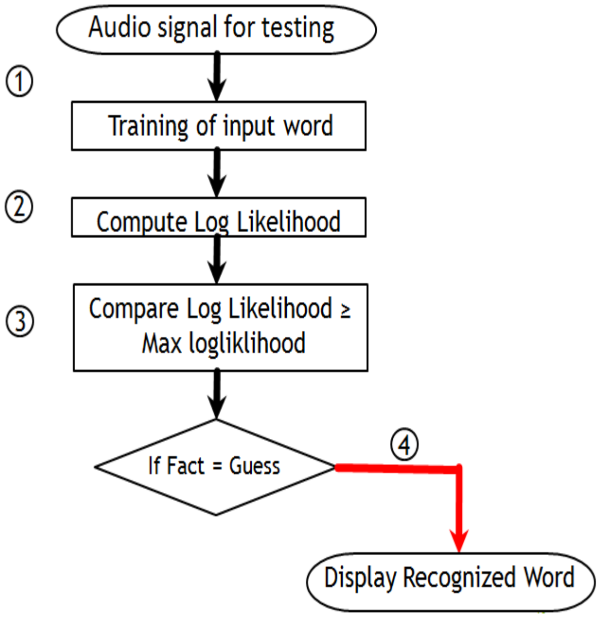


Fig 2.a :flowchart for training

Fig 2 : flowchart for recognition

A . ***Pitch frequency :-*** Pitch detection is also referred as fundamental frequency estimation is a popular topic for research today.While current recognition engines predict that pitch information is not an useful feature and phones and recognizable even if pitch is not known, we cannot assume that pitch infomtation is not necessary.To estimate pitch, the bandpass filter outputs from the auditory model are divided into two components; a high frequency part, where center frequencies are greater than 800Hz, and a low frequency part. An energy envelope is then extracted from the high frequency part using a Teager energy operator (TEO) [6]. An auto-correlogram is obtained from the energy envelope of the high frequency component and the remaining low frequency signals.

**B. *Modelling Pitch Frequency in Hybrid HMM/ASR***

Standard HMM based ASR models [4], the evolution of the observed space $X = \{x_1, \cdots, x_n, \cdots, x_N\}$ and the hidden state space $Q = \{q_1, \cdots, q_n, \cdots, q_N\}$ for time n = 1,….,N as:

$$p(Q, X) \approx \prod_{n=1}^{N} p(x_n|q_n) \cdot P(q_n|q_{n-1}) \qquad (1)$$

In hybrid HMM/ANN ASR $p(x_n|q_n)$ is replaced by the scaled likelihood $p_{sl}(x_n|q_n)$, which is estimated as [1]:

$$p_{sl}(x_n|q_n) = \frac{p(x_n|q_n)}{p(x_n)} = \frac{P(q_n|x_n)}{P(q_n)} \qquad (2)$$

For incorporating pitch frequency information Fo={Fo1,……,Fon,…….,FoN}, p(Q,X,Fo) is to be modeled.The pitch frequency can be any discrete value i.e. Fon € {1,…..,l,…..,L} or continuous valued. The easiest and most usual practice is to augment the feature vector xn with Fon and model the evolution of the augmented feature vector over the hidden state space Q similar to (4), resulting in:

$$P(Q,X,Fo)= \pi \ p(xn|qn,Fon) . p(Fon|qn) . P(qn|qn-1)$$

The implementation of such a system is straightforward, irrespective of whether the pitch frequency is discrete or continuous valued. As it can be observed from (6), thisfundamental approach also models the dependency between the state (qn) and the pitch frequency (Fon), which may be noisy. For example , pitch frequency does not describe anything about the state qn . In such a case, it would be good to relax the joint distribution in (2) by assuming independence between Fon and qn, yielding

$$P(Q,X,Fo)= \pi\ p(xn|qn,Fon).\ P(Fon)\ .\ P(qn|qn-1)$$

If the pitch frequency has a discrete value then, a system could be easily realized by training an ANN corresponding to respective discrete value. This is nearly same as gender modelling, in which acoustic models for male and female speaker are trained separately. When pitch frequency is continuous valued , it is not evident how to implement a HMM/ANN system.

B. ***Maximum Likelihood Estimators***

Based on the past values of the variables,their likelihood values of being correct is calculated and compared with different values of variables. The purpose is to recognize the slight unstability of naturally occurring frequency partials in a audio signal.

The procedure followed by the model is:

It consists of an observation set O consists of a subset of frequency partials and Fourier transform representation of an audio. We assume that each observation has been produced by an audio with a particular pitch, and also provides other information including inharmonic and non-sinusoidal partials such as noise about the spectrum. This model is a simplified version of generalised sound model.It assumes that a sound consists mainly of harmonic partials present at integer multiples of f0, with some inharmonic partials and noise.

For a set of pitch frequencies, using the algorithm we compute the likelihood that a given observation was generated from each f0 present in the set and then we find the maximum value. The selection of the set of pitch frequencies is essential because there is a chance that the observation could originate from any f0 present in the set.

## IV. BAUM-WELCH ALGORITHM :

The EM algorithm is also known as the Baum-Welch algorithm.This algorithm is used with HMMs to calculate maximum likelihood , it estimates the model parameters. The EM algorithm has two steps: E-step and M-step which stands for expectation step and maximization step. The expectation-step predicts the expected value of the complete log-likelihood with reference to the state considering the data and the available model parameters. The maximization-step maximizes the computed expectation in the E-step in order to find next state parameters. The Q function which is expectation of complete log likelihood given by:

$$Q\ \Delta,\Delta' = \in [\log p\ x,y\ \Delta) \mid y,\Delta'](8)$$

In (8), Δ= set of model parameters for current Iteration Δ′ = set of model parameters from preveous Iteration Probabilities for E-step are calculated by using forward and backward algorithms. These probabilities are used in the M-step. These two steps are repeated until convergence expectation.

## V . DATABASE

The database is created by recording 10 speech samples by single speaker at sampling rate of 16 KHz. Recorded speech files were saved in .wave file format. There are few isolated words such as „APP‟, „BACCHON‟, „MERA ‟, „HAI‟, „DESH‟, „BHARAT‟, „PUNE‟ „SHEHAR‟, „MEIN‟, „CHORO ‟, „SAVDHAN‟, „RAHE‟, „PATARI‟,

„SE" ", „DUR ", „RAHO", „KA", „DHYAN", „RAKHE" . Hence there are total 10 samples in database. This database is used for training and for testing purpose. Hence there are total 190 Samples of isolated words. The collected speech samples are then going to pass through the features extraction, features training and features testing stages.

## VI . RESULT

The approximate time required for training of all the samples is 6-7 minutes. The error eclipse graph is obtained in figure no. 3. As we increase the number of states the density of the graph goes on increasing as well as the accuracy increases. We obtain the log likelihood values of all the samples.The log likelihood values are mentioned in figure no. 4.

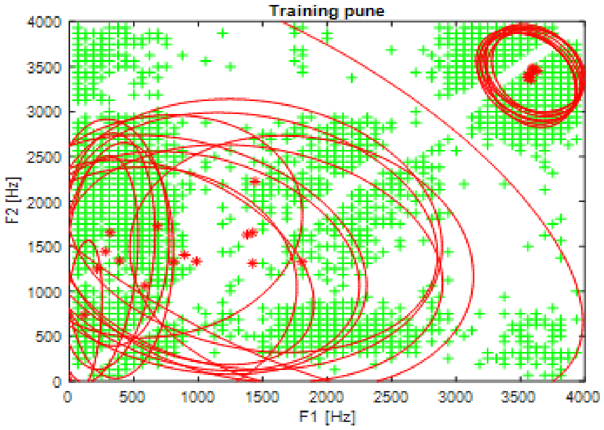


Fig 3: Training of word pune using error ellipse

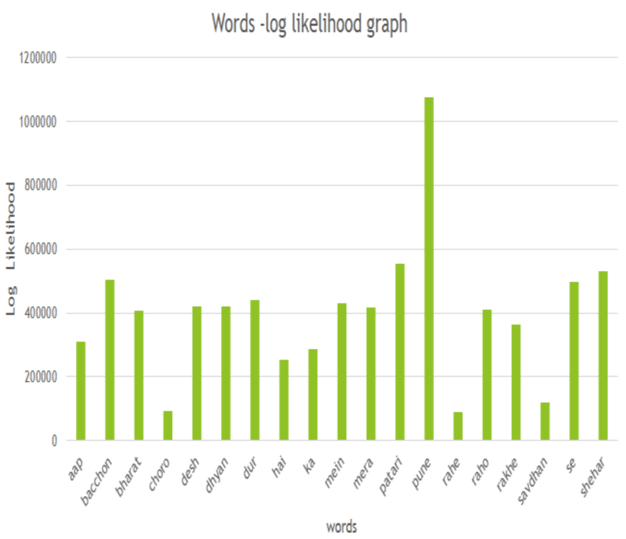


Fig 4 :Log-Likelihood graph of word samples

## VII. CONCLUSION

In this way, we developed a Automatic speech recognition system for hindi language. This work is carried for isolated words and will be later displayed on a LCD notice board. The approximate accuracy obtained is 75%. Further accuracy can increased by increasing the number of states. We can extend this project in future for large vocabulary and for various regional languages.

## VIII .REFERENCES

[1] Ms. Priyanka Shinde,Prof. P. M. Ghate 'Feature Selection for Speech Recognition using Hidden Markov Model' International Journal of Scientific Development and Research (IJSDR)www.ijsdr.org ,october 2017 IJSDR volume 2,issue 10 ISSN :2455-2631

[2] U. G. Patil, Dr. S. D. Shirbahadurkar and A. N. Paithane 'Automatic speech recognition of isolated words in Hindi Language using MFCC'2016 Internationalconference on computing,Analytics and security Trends.CAST-2016, 19-21 Dec 2016.http://ieeexplore.ieee.org/document/7915008/

[3] Hidden_Markov_Model_based_isolated_Hindi_word_recognition Ishan Bhardwaj; Narendra D Londhe 2012 2nd International Conference on Power, Control and Embedded Systems Year: 2012

[4] U. G. Patil, S. D. Shirbahadurkar, A. N. Paithane 'Isolated word recognition in Hindi Language',2nd International conference on Computing, communication,controlandAutomation(ICCUBEA2016),Aug2016.http://ieeexplore.ieee.org/document/7860101/

[5] Mathew Magimai.-Doss, Todd A. Stephenson, herve Bourlard 'Using pitch frequency information in speech recognition', IDIAP ,Eurospeech 2003-GENEVA

[6] David Gerhard, 'Pitch Extraction and Fundamental Frequency: History and current techniques' Department of Computer Science University of Regina Regina, Saskatchewan, CANADA S4S 0A2

ISSN 0828-3494 ISBN 0 7731 0455 0

[7] Priyanka Wani;U.G. Patil;D.S.Bormane , "Automatic Speech Recognition in Hindi Language", International conference on computing,Analytics and security Trends,2017.

[8] K.Kumar and R.K,Agarawal,"Hindi Speech Recognition using HTK",International Journal of Computing and Business Research,May2011 .

[9] _MFCC_based_Hindi_speech_recognition_technique_using_HTK_Toolkit Shweta Tripathy; Neha Baranwal; G. C. Nandi 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013) 2013

[10] J. A. Bilmes, „„A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models,"" Int. Comput. Sci. Inst., vol. 4, no. 510, p. 126, 1998.

[11] Aggarwal, R. K., and M. Dave. "Using Gaussian mixtures for Hindi speech recognition system." International Journal of Signal Processing, Image Processing and Pattern Recognition 4.4 (2011): 157-170.

[12] Saksamudre, Suman K., and R. R. Deshmukh. "Isolated Word Recognition System for Hindi Language." (2015).