Speech Recognition for Hindi using Zero Crossing Rate Method

Tulsi Meghwal[1] Prof. Ajaykumar T. Shah[2]
[1,2]Department of Computer Engineering
[1,2]Alpha College of Engineering & Technology, Khatraj, India

Abstract— Speech Recognition is the most promising field of research and technology. Speech to text conversion is the process of converting input acoustic speech signal into the text similar to information being conveyed by the speaker. This paper is to build speech to text conversion system for Hindi language to reduce the gap between computer and people in rural areas. Although there are many interfaces are already available, but need is to build more vocabulary and accuracy in it. The system is trained for 100 words, collected from different speakers of different age groups. In this system features including zero crossing rate and short term energy is studied. In this paper mainly three phases are there, training phase, testing phase and recognition phase. In training phase, training database is created with Hindi speech samples and trained using feature extracted using algorithms of zero crossing rate and energy calculation. In testing phase extracted features are matched and then created testing database with minimum and maximum range of zero crossing rate and energy from the speech samples. In recognition phase same techniques applied to speech and compared with the database and fetch word using data of training and testing phase.

Key words: Zero Crossing Rate (ZCR), Speech to text (STT), MFCC (Mel Frequency Cepstral Coefficient), LPC (Linear Predictive Coding)

I. INTRODUCTION

In today's world where technologies increasing day by day and become indispensable especially for urban people, but for the development of country where the most of the people lived in rural areas as whole, so the technology has to be reach them as well. Also physically impaired people or blind people can't experience the world that we do. They face a lot of difficulties while communicating with surrounding and this is where speech recognition becomes useful. Hindi is widely used language in India. People are not that much fluent in English especially rural people so they need system like speech to text conversion in Hindi language and can make their work possible in fast way with the help of technology. Speech recognition refers to the process where machine understand the input given by speaker in speech waveform and then identify the features and the information presents in phoneme, words and display that word to speaker in the form of text. So computer convert speech in audio format to text format.

II. SPEECH TO TEXT CONVERSION

The goal of speech recognition is for a machine to be able to "hear," understand," and "act upon" spoken information. The earliest speech recognition systems were first attempted in the early 1950s at Bell Laboratories, Davis, Biddulph and Balashek developed an isolated digit Recognition system for a single speaker [1].

A. Types of Speaker Model

Every speaker have different voice, some speak loud some speak slowly and some speak spontaneous so all have different style of speaking therefor speech recognition is divided in two models, namely speaker independent and speaker dependent.

1) Speaker Independent Model

Speaker independent model is more complex model then speaker dependent model. It is design for variety of speakers. It works for all different speakers. It gives less accuracy and is more expensive than speaker dependent model. But this model is more flexible. As it is trained using large numbers of vocabularies to get the accurate result.

2) Speaker Dependent Model

Speaker dependent model is designed for specific speaker. This type of model is more accurate for particular speaker but not for other speaker. It is easy to develop, cheaper and more accurate than but not as flexible as speaker independent model.

B. Flow of Speech to Text Conversion

Speech is the important means of communication between people. When machine understand human language it is called as man-machine communication. Recognising speaker's continuous speech with large vocabulary training data is very hard. It has too much complexity however with the help of modern process, diagrams, algorithms, flow chart we can process speech signal and can recognised speech by the speaker.

C. Stages of Speech to Text Conversion

There are four various stages involved in speech recognition system:

1) Speech Analysis
2) Feature extraction
3) Modelling
4) Testing

1) Speech Analysis

In this technique of speech analysis, speech data contain different types of information because of different style of speaking, vocal tract, behaviour, emotion etc. Because of different physical structure and vocal tract as well as excitation speech data varies and it is useful while extracting features and for signal processing in speech recognition [6].

2) Feature Extraction

Feature extraction is the main part in the process of speech recognition. It is called as heart of speech recognition. Feature extraction is the process through which we can differ one speech from the other. Every speech has different feature in it. This process extract the features from the speech and then classify speech. There are various feature extraction techniques are available in signal processing. Most important technique is MFCC (Mel Frequency Cepstral Coefficient) and another is LPC (Linear Predictive Coding).

3) Modelling

After the feature extraction process, all features are computed statistically and then compare to find difference between all speeches. The objectives of modelling techniques is to generate speaker model using speaker specific feature vector. The speaker modelling technique divided into two classification speaker recognition and speaker identification. The speaker identification automatically identify who is speaking on basis of individual information integrated in speech signal. The main aim of speaker identification is comparing a speech signal from an unknown speaker to database of known speaker [1]

4) Testing

Based on above all steps system is tested to get the accuracy level. By using high quality microphone, speech sample is recorded and then that speech undergo the matching of features range available in database and then map word and display on desktop.

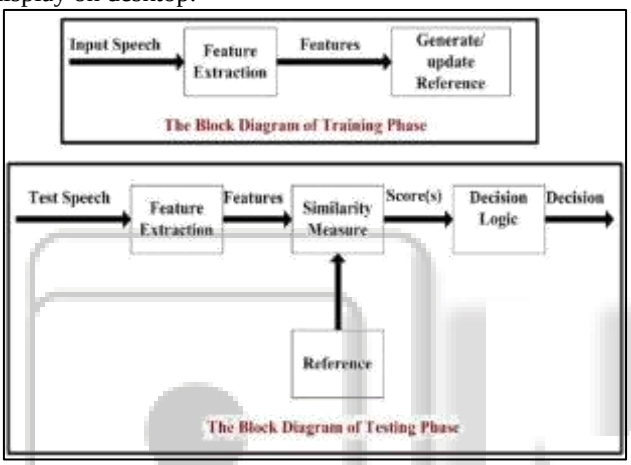

Fig: 1: Block Diagram of Testing and Training Phase

III. CLASSIFICATION OF VOICED AND UNVOICED DATA USING ZERO CROSSING RATE

Speech can be divided into numerous voiced and unvoiced regions. The classification of speech signal into voiced, unvoiced provides a preliminary acoustic segmentation for speech processing applications, such as speech synthesis, speech enhancement, and speech recognition. Unvoiced speech is non–periodic like random sound it is because of noise in background and/or air passing through constriction of vocal tract. Voiced speech is periodic in nature and do not contain noise or air in background and hence can be identified easily and get extracted. Qi and Hunt classified voiced and unvoiced speech using nonparametric methods based on multi-layer feed forward network. Acoustical features and pattern recognition techniques were used to separate the speech segments into voiced/unvoiced [11].

The method we used here to classify voiced and unvoiced speech is the calculation of zero crossing rate and energy of speech signal. Here, we used Hindi word/phoneme to classify them as voiced and unvoiced speech. The objective is to determine the ZCR and energy of each word/phoneme and then determine whether it is voiced or unvoiced. When zero crossing rate is lower than energy of speech signal it is said that speech is voice and if it is higher than energy then we called it as unvoiced data.

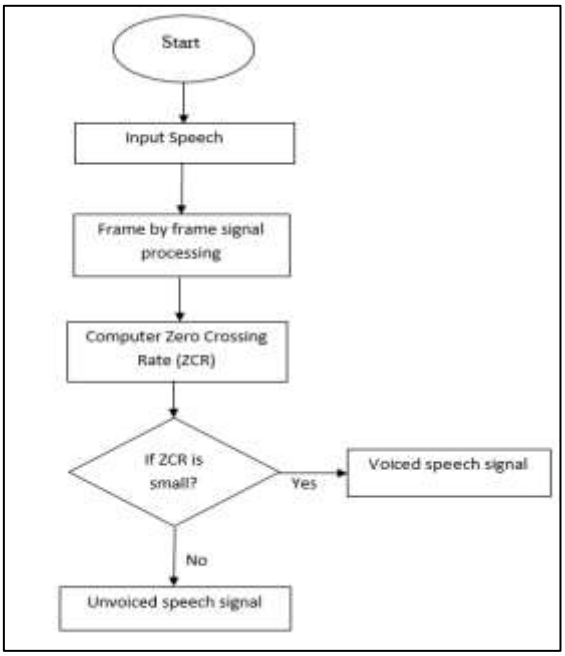

Fig: 2: Flow Chart to Classify Voiced and Unvoiced Speech Signal

A. Zero Crossing Rate

Zero crossing rate gives the information about number of zero crossing presents in given signal. Zero crossing is to calculate how many times the signal waveform crosses the zero amplitude line by transition from positive to negative and vice-versa in specific time. In mathematical term it is a point where sign of function changes represented by crossing of zero axis in the graph of function. If zero crossing are more in signal then signal will change rapidly and implies that it contain high frequency information. Similarly, if zero crossing are less then signal will change slowly denoting less frequency information [11].

A definition for zero-crossings rate is:

$$Zn = \sum_{m=-\infty}^{\infty} | sgn[x(m)] - sgn[x(m-1)] | w(n-m) \quad (1)$$

Where,

Sgn[x (n)] = 1 x (n) ≥0

= -1 x (n) <0

And w (n) is the windowing function with window size of N samples

w = 1/2N 0≤n≤N-1

= 0 otherwise

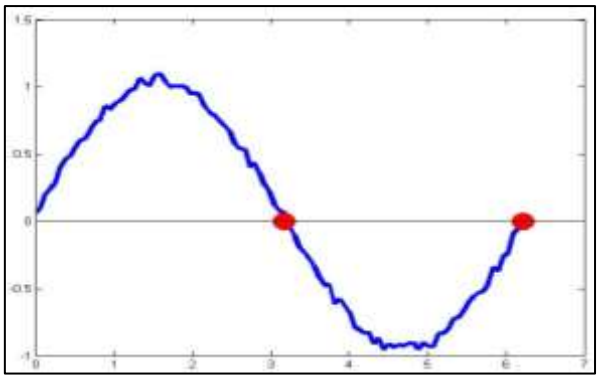

Fig. 3: Zero Crossing Rate

B. Energy of Speech Signal

The short time energy reflects the variation in amplitude. In particular amplitude of the unvoiced segment is generally much lower than the amplitude of voiced segment. The short time energy provide convenient representation of amplitude variation. For voiced speech, energy is higher than zero crossing rate and for unvoiced it is lower than ZCR [10].
The definition of short time energy can be given as:

$$En = \sum_{m=-\infty}^{\infty} [x(m)\, W(n-m)]^2 \quad (2)$$

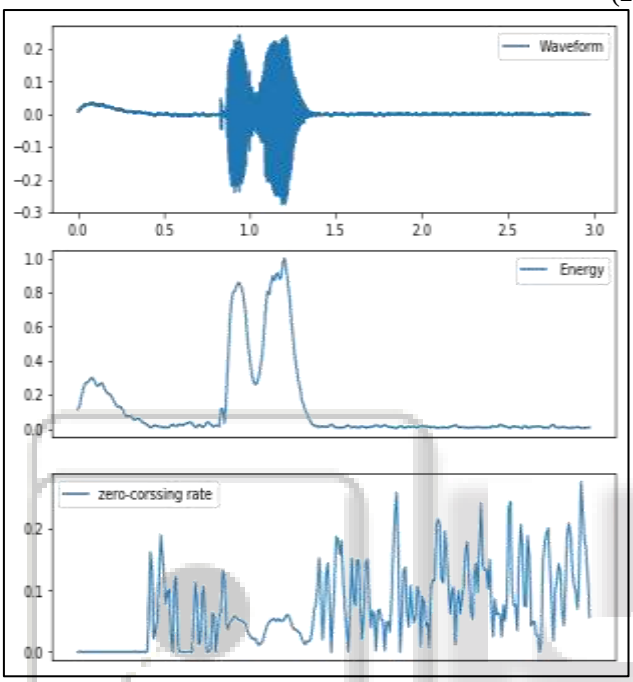

Fig. 4: Plotting of Waveform, ZCR, RSM energy

IV. RESULTS

Python programming language and PostgreSQL database is used for our speech recognition system. It contains a variety of signal processing and statistical tools, which help users in generating a variety of signals and plotting them. Hardware used are personal laptop/desktop and high quality microphone. There are two database used training and testing database. Training database consist of zero crossing rate and energy rate of every sample of single phoneme. There are 10 words each word have 10 samples recorded by speakers of different age groups. So total 100 samples are there to train. Testing database consist of minimum and maximum range of zero crossing and energy rate of all 10 words. So we can fetch word using min and max range. When speaker speaks any word from these 10 words it will first finds ZCR and energy rate using all the algorithms and classify voiced and unvoiced data and then compare with our training database template and finally fetch word whichever comes in that range. Table 1.Proposed classification of voiced and unvoiced decision using algorithms of zero crossing rate and energy rate. Table 2. Proposed minimum and maximum range value of 5 frame word.

Database for classification of voiced and unvoiced speech signal and min and max range of ZCR and energy rate is as follows:

Id	Frame ('Kaha')	Zero Crossing Rate(ZCR)	Energy Rate	Decision
1	Kaha1	21.28	38.26	Voiced
2	Kaha2	34.85	39.56	Voiced
3	Kaha3	16.43	32.59	Voiced
4	Kaha4	20.87	61.81	Voiced
5	Kaha5	25.69	50.49	Voiced
6	Kaha6	20.36	24.72	Voiced
7	Kaha7	59.93	22.31	Unvoiced
8	Kaha8	51.49	60.67	Voiced
9	Kaha9	17.02	19.57	Voiced
10	Kaha10	28.21	39.31	Voiced

Table 1: Classification of Voiced and Unvoiced Data

Id	Frame Word	ZCR _Min	ZCR _Max	En_ Min	En_ Max
1	आज	13	45	18	42
2	कब	9	36	10	37
3	केसे	15	37	9	50
4	कोन	10	50	11	40
5	कहा	15	60	22	62

Table: 2 Min & Max Range of Frames

V. APPLICATION OF SPEECH TO TEXT CONVERSION SYSTEM

Application of STT system is increasing day by day and also there is increasing improvement in the quality steadily. There are numbers of application of this system, some are as given below.

1) Aid to Vocally Handicapped
2) Source of Learning for Visually Impaired
3) Games and Education
4) Telecommunication and Multimedia
5) Voice Enabled E-mail
6) Home Automation System
7) Call Routing
8) Transcription of Speech
9) Vehicle Navigation system
10) Command and Control System

VI. CONCLUSION

In this paper, we discussed the topics relevant to development of speech to text conversion system. It seems effective and efficient to its user if its produce natural speech and by making several modification we can make it more useful. It is useful to deaf and dumb people to interact with their surroundings. This system is useful to rural people who have problem in English language, they can use it in Hindi language. In this paper, we have presented approach of classifying voiced and unvoiced speech signal in simple and efficient way by using algorithms of zero crossing rate and short term energy. There is a strong correlation between zero crossing rate and energy with frequency. Low frequency implies low ZCR and high frequency implies high ZCR. We have used just 100 samples in our system but for better results and accuracy we should include more and more vocabularies to get more accuracy. Speech recognition is the most

challenging and promising field to work with. It is natural communication between human and machines by applying knowledge from various areas of machine learning, Artificial intelligence and Neural network etc.

REFERENCES

[1] Santosh K. Gaikwad, Bharti W. Gawali and Pravin Yannawar, "A Review on Speech Recognition Technique", Volume 10– No.3, November 2010.
[2] Shaikh Naziya S and R.R. Deshmukh, "Speech Recognition System – A Review", Volume 18, Issue 4, Ver. II (Jul.-Aug. 2016), PP 01-09.
[3] Vimala.C and Dr.V.Radha "A Review on Speech Recognition Challenges and Approaches", Vol. 2, No. 1, 1-7, 2012.
[4] Bhoomika Dave, Prof. D. S. Pipalia "Speech Recognition: A Review", Volume 1, Issue 12, December -2014.
[5] Bhoomika Dave and Prof. D. S. Pipalia,"Speech Recognition: A Review", Volume 1, Issue 12, December -2014.
[6] Miss.Prachi Khilari and Prof. Bhope V. P, "A Review on Speech to Text Conversion Methods", Volume 4 Issue 7, July 2015.
[7] Sanjib Das, "Speech Recognition Technique: A Review", Vol. 2, Issue 3, May-Jun 2012, pp.2071-2087.
[8] Preeti Saini and Parneet Kaur, "Automatic Speech Recognition: A Review", Volume 4 Issue 2- 2013.
[9] POONAM.S.SHETAKE, S.A.PATIL and P. M JADHAV, "Review Of Text to Speech Conversion Methods", Volume-2, Issue-8, Aug.-2014.
[10] Bachu R.G, Kopparthi S, Adapa B. and Barkana B.D, "Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal", unpublished.
[11] D.S.Shete, Prof. S.B. Patil and Prof. S.B. Patil "Zero crossing rate and Energy of the Speech Signal of Devanagari Script", Volume 4, Issue 1, Ver. I (Jan. 2014), PP 01-05.
[12] Bitopi Sharma and Prof. P.H. Talukdar "Zero Crossing Rate of the Voice and Unvoiced Speech Signal of Assamese Words", Volume 7, Issue 12, December-2016 402 ISSN 2229-5518.

