IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

KANNADA SPEECH SEGMENTATION AND RECOGNITION FOR SPEECH TO TEXT CONVERSION

[1]Shwetha M, [2]Lavanya M, [3]Devendran B

[1]PG Student, [2]Assistant Professor, [3]Assistant Professor
Department of Electronics and Communication Engineering,
Maharaja Institute of Technology Mysore, Mandya, Karnataka, India

Abstract: Kannada is a Dravidian language spoken by 60 plus million people of Karnataka state. It is very challenging to generate acoustic model for Kannada language. For the design, implementation and performance evaluation of speech recognition system, the large amount of acoustic data is required. Hence, it's important to develop a speech database. The paper describes a simple methodology for Kannada speech recognition system and speech to text conversion. "Speech segmentation and recognition", are the two key components implemented in this work. The main objective of this work is to facilitate the user to interact with the computer in Kannada language and the uttered characters will be converted to text and display the same. To accomplish this task the user needs to train the system. System training firstly ask for recording of the spoken alphabets and words of various speakers into a specific audio file format. Secondly generating the codebooks for each speaker's uttered alphabets and words. During testing phase, each uttered speech is compared with the corresponding codebook database stored in computer memory and highly matched character is returned for the display. To accomplish the main objective of the proposed work, the following tasks such as feature extraction, template matching and other important functionalities are developed using MATLAB software.

Index Terms – Speech to Text Conversion (STT), Mel Frequency Cepstrum Coefficients (MFCC), End Point Detection, Linde Buzo and Grey (LBG) algorithm and Vector Quantization (VQ).

I. INTRODUCTION

Many technologies have arrived and vanished. But whenever a newer technology comes along, it is about to alter the way that one interacts and work with computer. Newer technologies try to simplify the existing one. The main mode of human - human interaction is speech. Hence speech signal provides a way to communicate with a computer also. Which usually solves various human-computer interaction problem. With so many different sources, speech is a signal that reflects acoustic power. The speech signal is presented as an electrical pulse, with the help of the signal envelope they can be shown by the use of amplitude modulation. Speech/voice is treated as a wave of pressure, and then transformed into numerical values if it is to be processed digitally. To convert a pressure wave to a numerical value calls for a few hardware gadgets: The change in the frequency of an electrical signal to a speech signal is accomplished by the microphone. The process of sampling at a specific rate produces a certain amount of voltage, and finally the quantization of each signal in a certain number is done using analog to the digital converter. The sampler and analog to digital converter circuits are incorporated within the preprocessing stage of the speech processing system.

Whenever a person wants to user a PC to type a document into any work processing software they have to work hard and have to spend their valuable time typing the document. Time is wasted, the main reason being that the speed of typing is not proportional to the speed that one can think, speak or feel. Secondly one has to pause in the words to find and look at the text. In the present world everyone is concerned with time management, which requires direct communication between spoken word and computer. This technology makes the job much easier. In the near future, speech may be considered as a popular input mode which is going to reduce the usage of mouse and keyboard input devices substantially. With a view to presenting a variety of mathematical concepts including audio, text, graphics, and video interface of a human laptop it may be appropriate. The goal is to provide a natural form of interplay between a human user and a pc. The powerful and popular communication mode is speech, so why is it not considered the most widely used in computer communication? Because the audio may be inappropriate in a given context. For example, very well-spoken instructions can be viewed in real time. As a result, in cases where multiple processes need to be controlled at the same time, voice communication is not an effective means. As with other forms of play, sound has certain strengths and weaknesses that need to be learned. But those difficulties were significantly suppressed, and presently sound is a good alternative input method.

Computer programs can now detect and respond to native speakers in other everyday languages within a given domain. Computational language and artificial intelligence, basic and applied research on speech were mixed to offer new opportunities in the fields of human computer communication. The primary mode of human-to-human communication is natural language and audio channels, although limited use between human and system. A Man can talk with his computer in the same way that with a friend which can be done through software tools. The presented work solely concentrated towards the development of a simple human-computer interaction system in kannada language. Speech pattern recognition in the present study is related to Kannada language. In the Kannada language, 15 (swara) +34 (sha and j ~ ja rather than Vyanjanas) varies from the letter ‘a’ to ‘ksha’. The main subject of study is to identify alphabets and some words of Kannada language. When the user speaks any letter or word to the microphone, different alphabet patterns will be recorded and compared with the similar pattern stored in the codebook database. As a result, the most matched alphabet/word of Kannada language will be displayed as text on screen. This makes human-computer interaction in Kannada language easier.

II. Related Work

Kannada Speech to Text Conversion Using CMU Sphinx by Shivakumar K.M, Aravind K.G, Anoop T.V Deepa Gupta. This paper presents the complicated issues of speech to textual conversion of Kannada language. They presented a unique Kannada Speech to Textual content conversion System (ASTC). They have used CMUsphinx framework for speech processing. CMU sphinx is dynamic in nature with help for different languages together with English. they trained the acoustic version for Kannada speech with a thousand widespread spoken sentences and examined 150 sentences. In this paper, Kannada sentence with 4 to 10 phrase length is researched. The speech conversion system lets in everyday human beings to talk to the laptop so one can retrieve statistics in the form of text [1].

Development of Automatic Kannada Speech Recognition System by Akshata K Shinde, Anjali H R, Deepika N Karanth, Gouthami K, Vijetha T S Automatic Speech Recognition (ASR) is the technology that allow individual to use their voices to speak with a computer/ laptop interface in a manner that, in its most sophisticated variations, resembles regular human conversation. Automatic speech recognition is a place of studies which offers with the recognition of speech by using machine in several conditions. The paper presents a short survey on Automated Kannada Speech Recognition structures. Many studies and trends are made to improve the performance of ASR to work efficient with the aid of the researchers. The criteria for designing speech popularity system are facts training, pre-processing filter, end-point detection, feature extraction strategies, speech classifiers, and overall performance assessment. Speech recognition device for Kannada language has been applied using the Hidden Markov Tool kit [2].

Automatic identification of Silence, Unvoiced and Voice Chunks in Speech by Poonam Sharma and Abha Kiran Rajpoot. The main theme of this is to segment the speech signal into silence, voiced and unvoiced parts automatically which in turn contributes to the improvement of accuracy and performance the system. Certain rules were proposed based on 3 crucial traits of speech signal namely zero crossing price, short time energy and fundamental frequency. The overall performance of the proposed set of rules is evaluated using the statistics accumulated from four speakers and accuracy of 96.61% is achieved [3].

A Review on Speech to text Conversion Methods by Miss. Prachi Khilari and Prof. Bhope V. P. This work gives a top-level view of main technological perspective and appreciation of the essential development of speech to text conversion and additionally offers evaluation technique evolved in every stage of classification of speech to text conversion. In this system, they developed an on line speech-to-textual content system. But, the switch of speech into written language in actual time requires special strategies as it need to be very fast and nearly 100% correct to be comprehensible. The goal of this assessment paper is to recapitulate and match up to different speech recognition systems in addition to procedures for the speech to textual content conversion [4].

III. Methodology

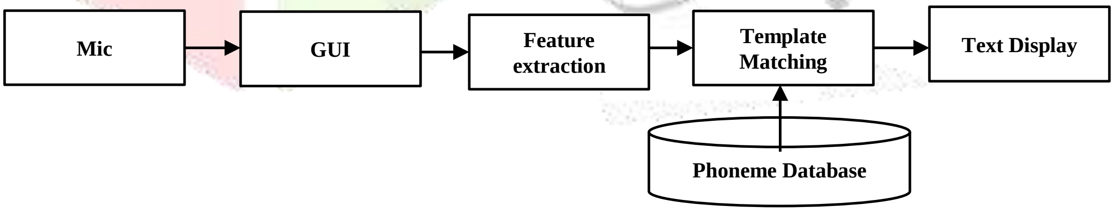

Figure-1 Conceptual diagram of proposed system

The proposed system is divided into 5 exclusive steps (figure-1). Speech acquisition (Mic), GUI (Graphical User Interface), Feature extraction, Template matching and Text Display.

1. **Speech acquisition (Mic):** Speech samples are recorded in real time for each speaker and stored in computer memory for further processing. It requires a microphone integrated with an analog-to-digital converter (ADC) with proper amplification to obtain the good voiced speech signal, then sample it and converts it into digital form of speech. In the designed system 25 different speakers (Both male and female) are considered. The directories are named after the speaker names. The recorded wave files and corresponding codebooks of each speaker is saved in their respective directories. All speech data is stored in one directory named “User database” which includes Swaras, Vyanjanas and few words of all the twenty-five speakers. Once the recording is over and codebooks were generated for each user, then the codebooks were segregated based on Swaras, Vyanjanas, words and gender (i.e. Male and Female). Final step involves combining all different Male and Female speaker codebook into single master codebook according to their gender (i.e. Male codebook and Female codebook).

2. **GUI:** A Graphical User Interface (GUI) is a graphical show in a single or many window containing controls, referred to as component, which allow a user to carry out interactive obligations. The user of the GUI does no longer have to create a script or type instruction at the command line to perform the required operations. In order to make it simpler and flexible platform for a user, GUI is required. In this work the GUI is developed using MATLAB software, where one can create his own GUI, which includes a figure window containing menu bars,

push buttons, textual content, graphics, and many others. There are two main steps involved in creating a GUI: one is designing its outlook, and the other is writing call back capabilities that carry out the required operations while the user selects specific features. [8]

3. Feature extraction [4][5]:

i. **Pitch period:** In this context a 10 ms interval is treated as a pitch period and it is also expressed in terms of pitch frequency f, where f = 1/ (10 ms) = 100 Hz. For male speakers, pitch frequency varies from 40 to 120 Hz, whereas for female speakers it could be as high as 200 to 400 Hz.[3]

ii. **End point detection:** The prime challenge in speech signal processing is detection of voiced speech in a noisy environment. This problem is known as the end point detection problem. The accurate detection of a words starts and end point allows the subsequent processing of the desired data, which reduces the memory requirement and processing time. Consider the speech recognition approach based totally on template matching. The exact timing of an utterance will generally not be similar to that of the template. They may additionally have specific intervals. In lots of cases the accuracy of alignment depends at the accuracy of the endpoint detection. Which will perform well, the algorithm must to take some of unique situations into account inclusive of:

- characters start or end up with low-energy phonemes
- characters ends with an unvoiced plosive.
- phrases with nasal sound at the end of utterance.[3]
- speakers ends phrases with a low intensity

iii. **Vector Quantization (VQ):** Feature extraction techniques used in speaker recognition system involves Dynamic Time Warping (DTW), Hidden Markov Modelling (HMM), and Vector Quantization (VQ). The presented work uses VQ approach. It is a method of mapping vectors from a large vector space to a finite amount of areas in the same region. Each area is considered as a group and represented as a codebook. The figure-2 suggests a conceptual diagram to illustrate the VQ process, In the figure, the circles compare the acoustic vectors from the speaker 1 wherein the triangles from the speaker 2. For each speaker, a speaker-specific VQ codebook is generated for each and every uttered speech character represented as a training acoustic vector. The resultant codewords (centroids) are decided with the aid of way of using black coloured circles and triangles for speaker 1 and speaker 2, respectively. The minimum distance from a vector sample to the codeword (Centroid)of a codebook is called a VQ-distortion. In real time testing, VQ distortion is computed for unknown utterance as stated in the above procedure and resulting codewords are used for template matching process. [10]

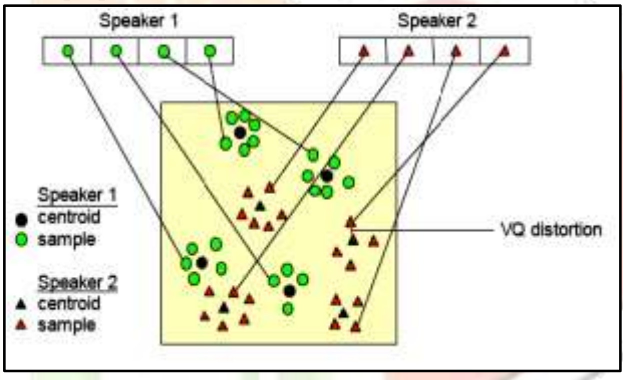

Figure-2 Illustrating the formation of VQ codebook

iv. **LBG (**Linde, Buzo and Gray**) algorithm:** It is used for clustering of a fixed of L training vectors into a fixed M codebook. The algorithm is formally carried out by using way of the following recursive manner:

(1) Design a single-vector codebook; i.e., the centroid of the whole set of training vectors.

(2) Double the scale of the codebook by way of splitting every contemporary codebook in step with
$r^+ = r\,(1+ \varepsilon)$, $r^- = r\,(1- \varepsilon)$ where ε is a splitting parameter ($\varepsilon = Zero$).

(3) Nearest-Neighbour search using Euclidian distance measure (disteu function): for every training vector, locate the codeword inside the cutting-edge codebook this is closest, and assign that vector to the corresponding cellular.

(4) Centroid replace: update the code word in each cell the use of the centroid of the training vectors assigned to that cell.

(5) Iteration 1: repeat steps three and four till the common distance falls underneath a present threshold.

(6) Iteration 2: repeat steps 2, 3 and 4 until a codebook of length m is generated.

v. **Mel Frequency Cepstrum Coefficient (MFCC)** [6] [7]**:**

In MFCC, the sequence of processing for speech data are (1) Windowing of data with a hamming window (2) Transform the data into frequency domain by taking FFT (3) Determining the magnitude of FFT (4) Obtaining filter bank output using FFT data. (5) calculate the log base 10 and (6) calculate the cosine transform to lessen dimensionality. A block diagram representation of MFCC method is shown in figure-3.

Speech Segmentation: In this step the uttered continuous speech signal is segmented into N frames, with adjoining frame being differentiated by means of M samples(M< N). The first frame includes the primary N sample. The second frame starts with M samples after the primary frame, and overlaps it by (N-M) samples. In addition, the third frame starts 2M samples after the first frame and overlap it with the aid of N - 2M samples. This process continues until the entire speech signal is accommodated within one or more frames. Typically, N and M values are assumed to be 256 and 100 respectively.

Windowing: The hamming window is used. In MATLAB software an inbuilt function called hamming(N) is used to locate the hamming window, W = hamming(N) returns the N-point symmetric Hamming window in a column vector w, where N should be a finite positive integer. The hamming window coefficients are calculated as:

$$w[k+1] = 0.54 - 0.46\cos\left(\frac{2\pi k}{n-1}\right), k = 0, \ldots\ldots n-1$$

Fast Fourier Transform: The fast Fourier Transform converts each frame of N samples from time domain to frequency domain.

Mel-Frequency Warping: The spectrum acquired from the above step is Mel Frequency Warping. The Mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above a 1000 Hz. As a reference component, the pitch of a 1 kHz tone, 40 dB above the perceptual attention to threshold, is defined as 1000 Mels. Consequently, the subsequent approximate system can be used to compute the Mels for a given frequency f in Hz:

Mel (f) =2595*log10 (1+f / 700)

This step helps to transform the frequency spectrum to Mel spectrum.

Cepstrum: Here in this step, the log Mel spectrum is converted back to time. The result is known as Mel-Frequency Cepstral Coefficient (MFCC). The cepstral representation of the speech spectrum gives a first-rate representation of the nearby spectral properties of the signal for the given frame analysis, Because the Mel spectrum coefficients are real numbers, which can be converted to time domain using Discrete Cosine Transform (DCT).

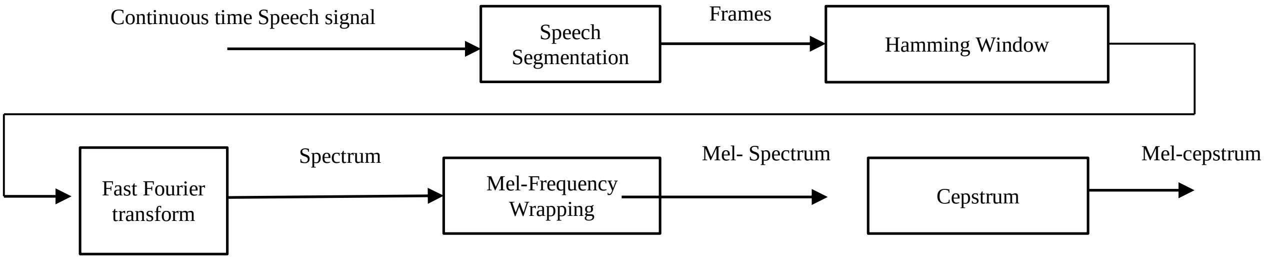

Figure-3 MFCC block diagram for feature extraction

4. **Template Matching:**

Euclidean Distance [10]:

In speech recognition system, an unknown speakers voice is represented as a set of characteristic vectors {x1, x2,. . xi} after which it is as compared with the codebook database stored in computer memory. The distance between current vectors and codebook vectors were calculated as Euclidean distance using the following mathematical equation. So that it will identify the unknown speaker. Euclidean Distance between two points can be calculated as follows:

$$P = (P_1, P_2 \dots \dots P_n) \text{ and } Q = (Q_1, Q_2 \dots \dots Q_n)$$
$$= \sqrt{(P_1 - Q_1)^2 + (P_2 - Q_2)^2 \dots \dots (P_n - Q_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (P_i - Q_i)^2}$$

5. **Text Display**: Upon executing all the above described processes, a text character corresponding to the result of template matching step is displayed on the screen.

IV. RESULTS AND DISCUSSION

The developed GUI is capable of performing Kannada speech segmentation, recognition and speech-to-text conversion. The push buttons and displays are coded in such a way that they can perform the above said five steps of the proposed work. The record push button records the speech, performs endpoint detection operation and saves the speech signal into the required folder in wave format. Dedicated pushbuttons are designed to test and generate codebooks of Swaras, vyanjanas and words individually. Four auxiliary displays are provided to plot original speech signal, endpoint removed speech signal (silence removed signal), corresponding text and spectrogram of the uttered speech which can be seen in clockwise direction in the developed GUI shown in figure 4. After recording the speech samples, each speaker codebooks can be updated by using dedicated pushbuttons. While testing for speaker independent and speaker dependent performance of the developed system, a (select speaker) pushbutton is given to select the required database for matching with the real time uttered speech. The developed system is also capable of identifying the gender of the speaker based on the pitch period of uttered speech. With the developed system, the average efficiency of 85% and 65% is achieved for speaker dependent and speaker independent database respectively for both male and female speakers are taken into consideration. Sample results are tabulated in the following tables.

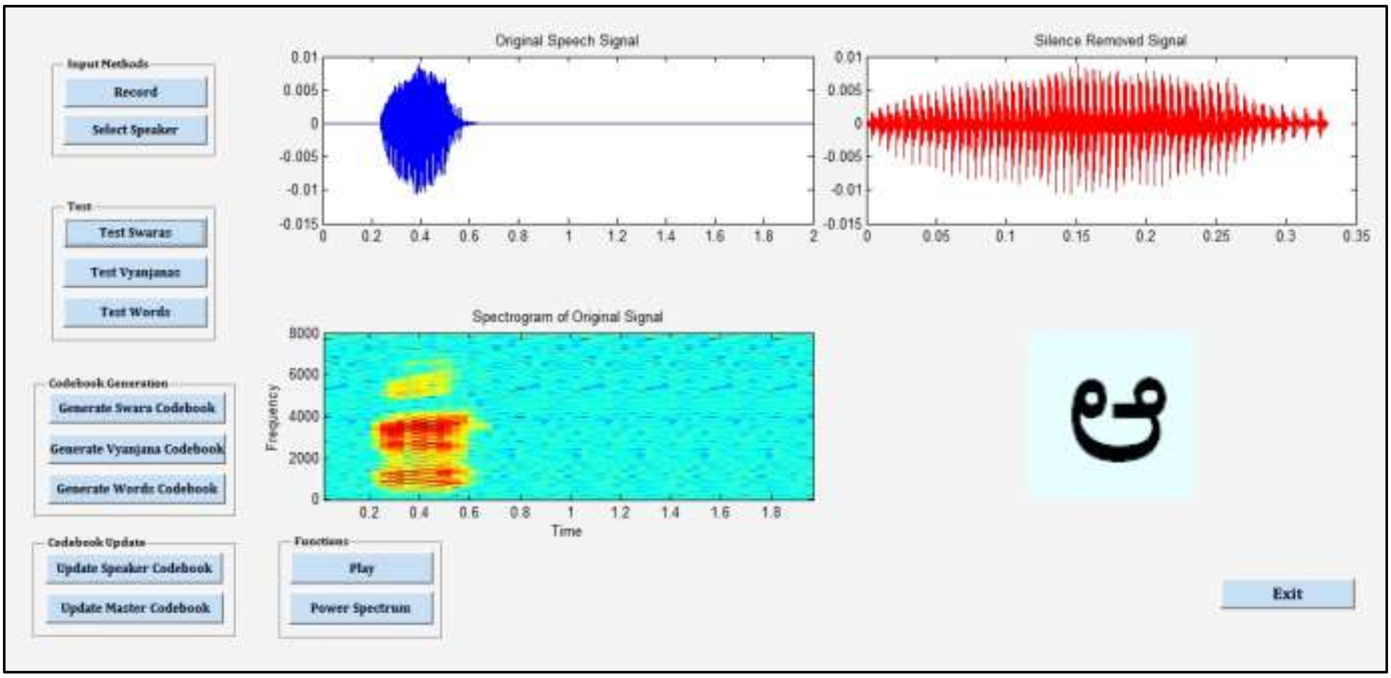

Figure-4 Developed GUI (Front end)

Table 1: Speakers dependent recognition of swara characters

SL. No	Speaker	Male Speaker Trials						Female Speaker Trials					
	Swaras	1	2	3	4	5	Utterance Success Rate	1	2	3	4	5	Utterance Success Rate
1.	***a***	a	a	a	a	a	100%	a	a	a	a	a	100%
2.	***aa***	aa	aa	aa	aa	aa	100%	aa	aa	a	aa	aa	80%
3.	***i***	i	ii	i	i	ii	80%	i	i	i	ai	i	80%

Table 2: Speakers independent recognition of swara characters

SL. No	Speaker	Male Speaker Trials						Female Speaker Trials					
	Swaras	1	2	3	4	5	Utterance Success Rate	1	2	3	4	5	Utterance Success Rate
1.	***a***	ai	a	a	a	a	80%	aa	a	a	a	a	80%
2.	***aa***	aH	o	ai	aa	a	20%	aa	aa	a	aa	aa	80%
3.	***i***	ii	e	i	i	ii	60%	i	i	i	ai	i	80%

Table 3: Speakers dependent recognition of vyanjana characters

SL. No	Speaker	Male Speaker Trials						Female Speaker Trials					
	Vyanjanas	1	2	3	4	5	Utterance Success Rate	1	2	3	4	5	Utterance Success Rate
1.	***ka***	ka	ka	gna	ka	ka	80%	ga	ka	ka	ka	ka	80%
2.	***cha***	cha	cha	cha	cha	cha	100%	cha	chha	cha	cha	cha	80%
3.	***dda***	dda	dda	dda	dda	dda	100%	dda	dda	ddha	dda	dda	80%
4.	***ba***	ba	ba	ba	ba	ba	100%	ba	ba	ba	ba	ba	100%
5.	***sha***	sha	sha	sha	sha	sha	100%	sha	sha	ksha	sha	sha	80%

Table 4: Speakers independent recognition of vyanjana characters

SL. No	Speaker	Male Speaker Trials						Female Speaker Trials					
	Vyanjanas	1	2	3	4	5	Utterance Success Rate	1	2	3	4	5	Utterance Success Rate
1.	***ka***	ka	ka	ka	ka	ka	100%	ga	ka	ka	ka	ka	80%
2.	***cha***	cha	cha	chha	cha	cha	80%	cha	cha	cha	cha	cha	100%
3.	***dda***	ddha	ddha	tha	ta	dda	20%	dda	dda	ddha	dda	dda	80%
4.	***da***	da	da	da	da	da	100%	dha	dha	da	na	da	40%
5.	***sha***	sha	sha	sha	sha	sha	100%	sa	sa	ksha	ksha	sha	20%

Table 5: Speaker dependent recognition of words

SL. No	Speaker	Male Speaker Trials						Female Speaker Trials					
	Words	1	2	3	4	5	Utterance Success Rate	1	2	3	4	5	Utterance Success Rate
4.	***Nayana***	Nayana	Nayana	Nayana	Nayana	Nayana	100%	Nayana	Nayana	Nayana	Nayana	Nayana	100%
5.	***Kamala***	Kamala	Kamala	Kamala	Kamala	Kamala	100%	Kamala	Kamala	Kamala	Jana	Kamala	80%

6.	***Pala***	Pala	Pala	Nala	Pala	Pala	80%	Nala	Pala	Pala	Pala	Pala	80%

Table 6: Speaker independent recognition of words

SL. No	Speaker	Male Speaker Trials						Female Speaker Trials					
	Words	1	2	3	4	5	Utterance Success Rate	1	2	3	4	5	Utterance Success Rate
7.	***Nayana***	Nayana	Samaya	Nayana	Nayana	Nayana	80%	Samaya	Nayana	Yama	Nayana	Nayana	60%
8.	***Kamala***	Tabala	Kamala	Kamala	Kamala	Kamala	80%	Kamala	Kamala	Kamala	Jana	Kamala	80%
9.	***Pala***	Pala	Pala	Nala	Pala	Pata	60%	Nala	Pata	Pala	Pala	Pala	60%

V. ACKNOWLEDGMENT

I bear my heartfelt gratitude to the Principal **Dr. B.G. Naresh Kumar**, for giving an opportunity to pursue this project work. I bear my grateful thanks to **Dr. Mahesh Rao**, Head of the Department of E&C Eng., for his encouragement and guidance in this project. I would like to extend my thanks to **Mrs. Lavanya M**, Assistant Professor, Department of E&C Eng., the guide in this project for her invaluable suggestions and involvement throughout the pursuit of this project work. I extend my deepest gratitude to **Mr. Devendran B**, Assistant professor, department of E&C Eng. for the guidance and cooperation provided.

References

[1] Shivakumar K.M, Aravind K.G, Anoop T.V and DeepaGupta Kannada Speech to Text Conversion Using CMU Sphinx

[2] Akshata K Shinde, Anjali H R, Deepika N Karanth, Gouthami K, Vijetha T S Development of Automatic Kannada Speech Recognition System Vol-5 Issue-3 2019 IJARIIE-ISSN(O)-2395-4396

[3] Poonam Sharma And Abha Kiran Rajpoot Automatic Identification of Silence, Unvoiced and Voiced Chunks in Speech

[4] Miss.Prachi Khilari, Prof. Bhope V. P. A Review on Speech to Text Conversion Methods International Journal of Advanced Research in Computer Engineering & Technology (Ijarcet) Volume 4 Issue 7, July 2015

[5] Ayushi Trivedi, Navya Pant, Pinal Shah, Simran Sonik and Supriya Agrawal, Speech to text and text to speech recognition systems-A review IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 2, Ver. I (Mar.-Apr. 2018), PP 36-43 www.iosrjournals.org

[6] Su Myat Mon, Hla Myo Tun, Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM).

[7] Shreya Narang, Ms. Divya Gupta, Speech Feature Extraction Techniques: A Review IJCSMC, Vol. 4, Issue. 3, March 2015 **[8]** Nuzhat Atiqua Nafis and Md. Safaet Hossain Speech to Text Conversion in Real-time International Journal of Innovation and Scientific Research ISSN 2351-8014 Vol. 17 No. 2 Aug. 2015, pp. 271-277 © 2015 Innovative Space of Scientific Research Journals http://www.ijisr.issr-journals.org/

[9] Sheela Deshmukh, Neha Jundare, Mohini Gore, Harsha Sarode Speech to Text Recognition System International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 03 Issue: 10 | April 2017 www.irjet.net p-ISSN: 2395-0072.

[10] Akanksha Singh Thakur1, Namrata Sahayam Speech Recognition Using Euclidean Distance International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 3, March 2013)**.**