Marathi Interactive Voice Response System (IVRS) using MFCC and DTW

Manasi Ram Baheti
Department of CSIT,
Dr.B.A.M. University,
Aurangabad, (M.S.), India

Bharti W. Gawali
Department of CSIT,
Dr.B.A.M.University,
Aurangabad, (M.S.), India

S.C. Mehrotra
Department of CSIT,
Dr.B.A.M. University,
Aurangabad, (M.S.), India

ABSTRACT

The need for quick and broad application of Speech enabled systems is becoming obvious. Studies have shown that there is a substantial "digital divide" that prevents many of our citizens, particularly illiterate rural people, from using the emerging technology. In order to accelerate this essential "technology transfer," this research aims to develop an Interactive Voice Response (IVR) system (IVRS) for rural agricultural purpose, using Marathi speech recognition. The automatic recognition of speech, enabling a natural and easy to use method of communication between human and machine, is an active area of research. The proposed IVR system in this research uses Speech Recognition technology in Marathi language that handles the queries of transactions at "Krushi Utpanna Bajar Samiti".

Keywords

Human Computer Interaction (HCI), ASR, IVRS, CSL MFCC, DTW, DFT

1. INTRODUCTION

Throughout the human history, speech has been the most dominant and convenient means of communication between people. With the rapid development of communication technologies, a reliable speech communication technique for human-to-machine interaction has become need. [1] Automatic speech recognition (ASR) is the core challenge towards the natural human-to-machine communication technology [2].

The study of Human Computer Interaction (HCI) currently plays vital role in Computer Science research and its importance will only deepen in the future. Understanding how to create/develop hardware and software to facilitate their use by people is a fundamental area of CS. If we look at the development of Computer Systems from its various generations, it is observed that , the main focus was to develop faster, powerful systems with cost reduction; in this common user was somehow neglected . The fastest, most powerful systems are of no use until people can adequately understand and use them.

HCI is a discipline that attracts innovation and creativity. Speech enabled IVR system will serve as bridge between people and database, by connecting the people through their phones (communication device) to access any information they need regarding specific application anywhere, anytime. [1]

The main objective of the study is to develop a Speaker independent system that will recognize the continuous sentence and also respond accordingly in Marathi language.

1.1 Introduction to Speech Recognition

Speech Recognition: It is a process by which a computer (or other type of machine) identifies spoken words. Basically, this means talking to the computer and having it correctly recognize what one is saying.

1.2 Digital Representation of Speech

The first step in processing speech is to convert the analog representations, into a digital signal. This process of analog-to-digital conversion has two steps: **sampling and quantization** (Digitization). A signal is sampled by measuring its amplitude at a particular time; the sampling rate is the number of samples taken per second. Quantization is the bit representation of the sampling. [2]

1.3 Interactive Voice Response System (IVRS)

Serving as a bridge between people and computer databases, interactive voice response systems (IVRs) connect telephone users with the information they need, from anywhere, anytime. [3]

This is one of the application areas of the automatic speech recognition (ASR) system. It is an interactive technology that allows a computer to detect voice & keypad inputs. Used extensively in telecommunications but now also used in automatic systems.

There are two ways user can give input to an IVRS:

i. **Touch tone:** in which, user is prompted to go through some navigation by pressing the option button on the keypad like, "for Marathi press1", "to check your account balance, press3" and so on.

ii. **Input** is **voice command or sentence**.

The level of complexity is illustrated in the following examples of user interaction with a speech-recognition IVR:

Touch-tone replacement

System Prompt: "For user’s current account information, press or say one."

Natural Language

System Prompt: "What transaction would user like to perform?" Caller response: "Transfer Rs.500 from current account to savings".

1.4 ASR and IVRS

The goal of speech-enabled applications has always been to allow callers to obtain information and perform transactions simply by speaking naturally rather than typing commands or navigating through some prompted menu.[4]

Speech recognition has become a foundation of self-service interactive voice response (IVR) user interfaces. The speech recognition IVR applications found to be cost-effective and the user friendly, speediest self-service alternative to speaking with a contact centre agent.

Speech technology can increase customer satisfaction and retention. Allowing the customers to use the most natural human interface, speech, to communicate with the users, instead of forcing them to navigate entering multiple digits into a key pad.[4]

2. NEED OF THE SYSTEM

Currently even though much work is done regarding HCI in Speech Recognition, it is limited up to urban area only. The development of IVRS is for handling online banking queries, customer care services, LPG Cylinder booking, etc. But actually, IVRS can also be used for solving the problems and queries of the farmers, which will develop the agricultural system on priority basis. The concept of making these advanced systems needs to reach up to rural level and solving the problem of communication gap is the main objective of the present study.

Rural customers lack real-time access to critical information such as commodity pricing, weather reports, local news, entertainment, agro machinery, agro products etc. Hence, we need to create "voice" based services on Interactive Voice Response (IVR) platform to reach out to such users. How voice recognition technology can be used to impact rural India was the main issue behind this. [5]

Although many interactive software applications are available, the uses of these applications are limited due to language barriers. Hence development of speech recognition systems in local languages will help anyone to make use of this technological advancement. In India, speech recognition systems have been developed for many languages.

Much work done in the area of ASR also IVRS is one of the application areas of Automatic Speech Recognition. On the other hand, speech recognition is yet to make its mark on regional languages like Marathi. Although recognition of Phonemes and small vocabulary words in some of the Indian languages has been attempted, yet the recognition of continuous speech in Indian languages is still awaiting serious attention.

In India, where 70% of the country's population is involved in the agriculture industry, speech technology has begun to play a critical role through user friendly speech solutions to rural farmers.[6] It was observed that, that the markets are highly skewed against producers with restrictive practices and non transparency due to their non-linkage with markets and the absolutely no access to information (literacy rates in Rural India is 58.7% with almost 71% males being literate while just 46% females are literate)[6] It is aimed to build a technology driven agriculture and allied services market place and information exchange accessible by the rural farming community in India.

It is aimed to make the farmer an equal trading partner with his buyers and his suppliers during the entire agricultural cycle thereby connecting the gaps in the information flow of the agriculture cycle. We also aim to introduce women centered appropriate technology and challenge the status quo of lower educational and literacy levels in women in rural India and the cultural barriers that discourage female ownership of productive assets and technical literacy issues.[7]

3. CREATEING DATABASE

The database creation was the important and first step in this research; as no database of such queries is available and that too on Marathi language. To create the database, we determined the most common inquiries in the "Mandi" i.e. "KRUSHI UTPPANNA BAJAR SAMITI"'s daily FAQs, by talking to agents, listening to customers.

For this study, the limited data base was created with ten questions and their related possible answers. Database creation, includes recording selected ten sentences related to agricultural queries and their answers in Marathi language. For this, experimental facility used was Computerized Speech Lab (CSL) system and Praat. The questions were related with the daily transactions in the Mandi which is at Taluka level where all the villages in that taluka are included. Farmers need information for selling their goods so they contact some intermediate person /agent or they have to traval to Mandi at taluka place. These were the questions related to the commodity price, storage place on rent, which crop is sold at what price and in what amount in tons, kwitnals etc. The Fig. 1 shows the general flow of the proposed system.

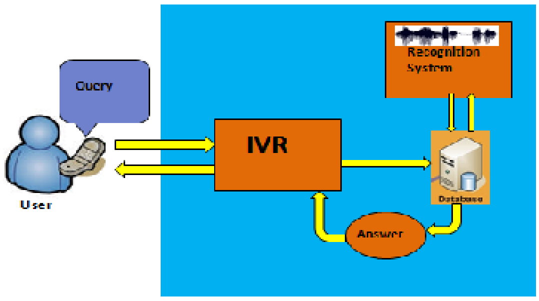


Figure1: IVR System flow

The database was created using microphone to record sound in the closed room using Praat and CSL. Ten speakers were selected from age group of 21years to 50 years old. It was combination of male and female speakers. Both, literate and partially literate persons were speakers for this database. Sampling rate of 16 KHz was selected in all cases.

Thus 1000 samples were acquired related to ten questions and corresponding 10 answers for ten different subjects.

It should be noted that for small size of this data base, large number of samples are required for reliable results. After creation of the database, the sentences recorded were stored as .wav files, and these wav file were processed to extract the features. The Fig. 2 shows the waveform, filter bank energies and the MFCC features for one sample (say Question Q1 spoken by the subject one , say S1) for the sentence " आज कोणत्या मालाची सर्वात जास्त   वक्री झाली ?"

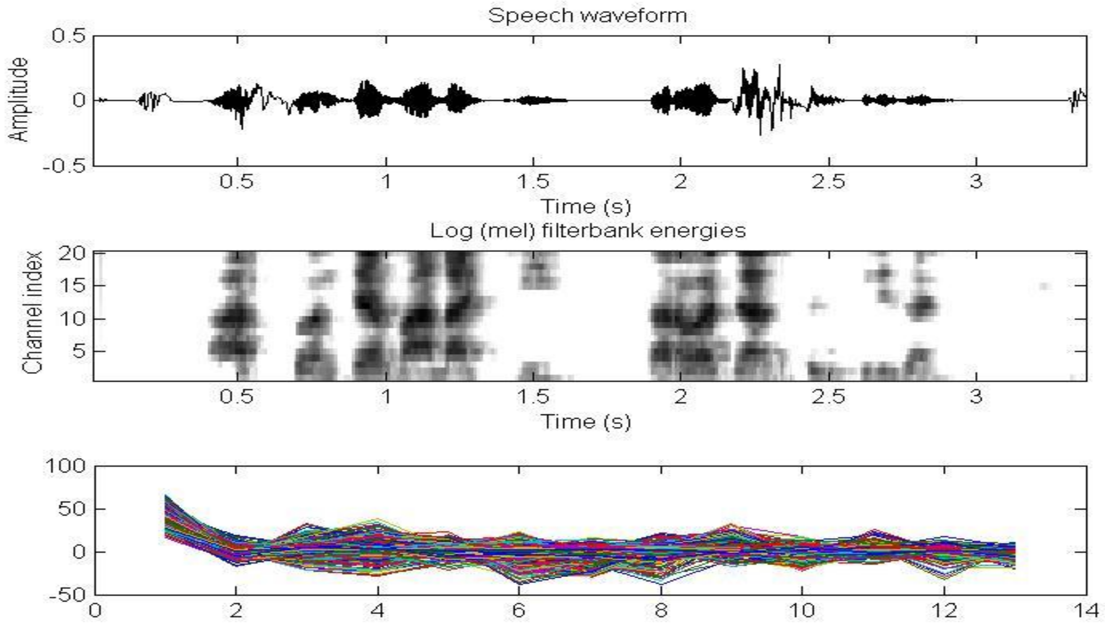


Figure 2: Waveform, Channel Index and MFCC plot of sentence " आज कोणत्या मालाची सर्वात जास्त वक्री झाली ?"

4. FEATURE EXTRACTION

A feature is a parameter that can be computed or estimated through processing signal waveform. There are three basic steps in the ASR, viz. (1) parameter estimation (in which the test pattern is created) (2) parameter comparison and (3) decision making. The function of the parameter measurement is to represent the relevant acoustic events in the speech signal in terms compact efficient parameters. The basic and most widely used parametric representation are linear predictive coding (LPC), Perceptual Linear Predictive coding (PLP),RASTA(RelAtive SpecTrA) and Melfrequency cepstrum Coefficient (MFCC).[6] The role of the speech recognition system is to create a mapping between the speech vectors and the corresponding symbols.[8]

4.1 Mel Frequency Cepstral Coefficient (MFCC)

MFCC is based on human hearing perceptions which cannot perceive frequencies over 2 KHz. In other words, MFCC is based on known variation of the human ear's critical bandwidth with frequency. Method used for extracting the features of the voice signal is to find the Mel frequency cepstral coefficients. Mel-frequency cepstral coefficients (MFCCs) are the coefficients that collectively represent the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. The Table 1 shows one example of MFCC. For each tone with an actual frequency f measured in Hz, a subjective pitch is measured on a scale called the 'Mel Scale'. The Fig. 3 shows the steps involved in MFCC feature extraction of the given wav file.[9] The fig. 4 shows the windowing process on the speech data.

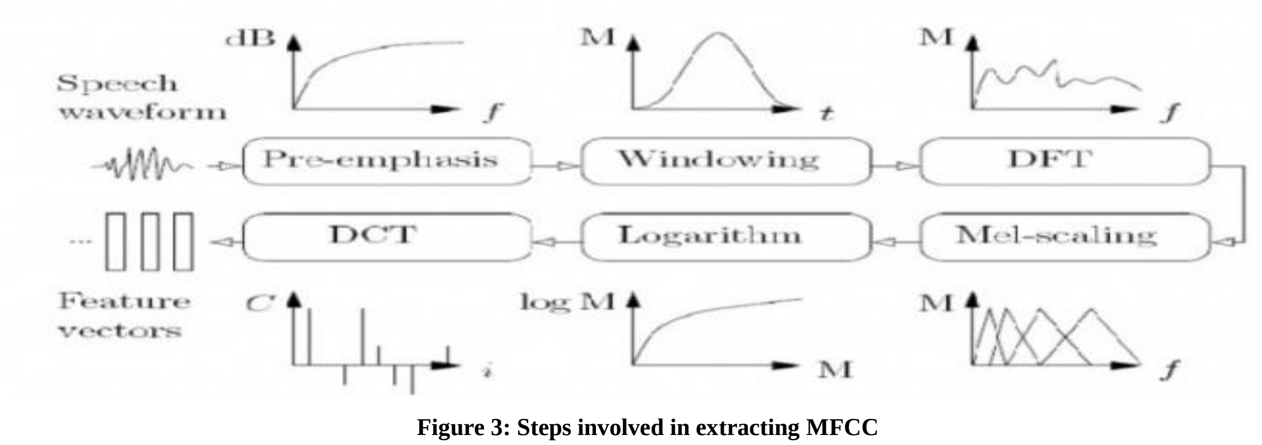


Figure 3: Steps involved in extracting MFCC

Frames and Windows

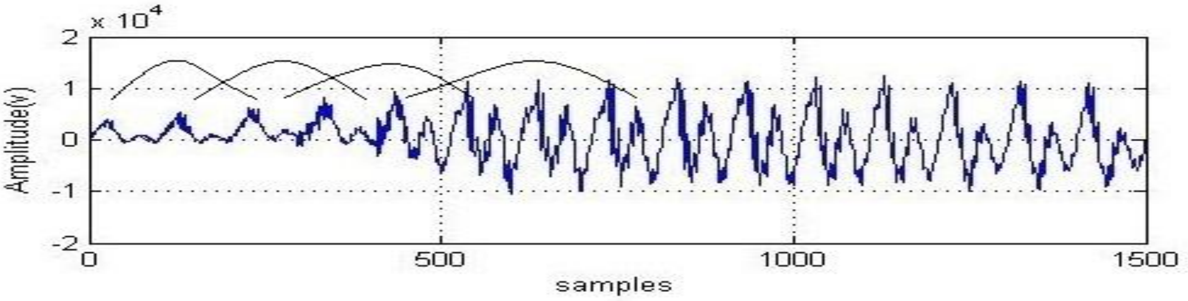


Fig. 4: Frame and Windows having 4000 Sample Points

Speech is a non-stationary signal; means that its statistical properties are not constant across time. So it is necessary to extract feature from a very small part . A window function is then applied, as practically every technique to extract features depends on the analysis of the spectrum.

We call the speech extracted from each window a **frame**, and we call the number of milliseconds in the frame the **frame size** and the number of milliseconds between the left edges of successive windows the **frame shift**. The extraction of the signal takes place by multiplying the value of the signal at time. Following table1 shows the mfcc features for the Sentence Q1 of Subject 1.We have selected 13 features with signal length 2 sec. (for table alignment) but practically we have taken the total length of the signal. Column shows frames and row shows the thirteen features. Speech is divided into sequence of 10- msec frames for faster

processing.[10]

Each frame will then lead to a series of features that are extracted. This is normally called a feature vector. A whole signal will thus lead to a series of feature vectors that can then be classified in the matching stage [11] Thus from figure 4 , we can have : Fs = 16,000 samples/second , Frame rate (overlap percentage) = 10 ms, Window length (Frame length) = 25 ms, (25ms * 16,000 = 4000 sample/frame).

Table1: MFCC Features the Sentence Q1 of Subject1

Frame 1 >> Q1	Frame 2	Frame 3	Frame 4	Frame 45	Frame 6	Frame 7	Frame 8	Frame 9	Frame 10	Frame 11	Frame 12	Frame 13
39.50737	47.27597	35.56057	49.04477	51.42742	42.2486	38.2634	34.77708	31.24231	27.77913	23.80565	21.26535	21.20368
-4.41982	-4.4146	-1.13364	9.895247	10.0015	2.169704	2.348698	3.067804	2.915738	0.072006	-2.95831	-7.48792	-6.81056
-0.11037	-0.10305	1.207346	5.804082	6.003898	0.629288	-1.44873	-2.51802	0.431667	-1.5925	-0.53127	-0.8451	-2.49311
-1.01202	-0.98071	-4.59241	-18.7437	-19.6236	-21.1884	-20.2181	-18.3936	-19.0269	-14.5762	-6.28929	-0.69788	-1.35163
0.07345	0.159369	2.523557	8.056113	9.435053	10.92527	10.67868	9.304574	4.700345	7.280687	3.96903	7.926042	10.55164
-0.43247	-0.3613	5.413068	0.341966	0.011036	1.296927	0.617655	1.801624	0.645949	1.628452	-6.13435	-7.05088	-1.75574
-0.09844	-0.07374	5.952951	2.031762	1.215276	1.444086	6.758804	2.701601	-0.26462	-6.44408	-8.46792	-12.184	-12.0323
-0.45607	-0.45407	-2.38951	-0.52559	1.279827	7.883485	8.589263	-4.36664	-6.31892	-0.58366	-2.30299	-9.99376	-6.77976
0.034796	0.12418	-4.00386	-2.77169	-3.98529	-4.94906	6.058572	4.299339	16.85351	14.48744	7.822887	17.6475	1.494007
0.185187	0.294888	-1.6873	-1.53951	-1.71725	-4.04995	-0.07904	-4.18928	-1.48661	-1.62741	-4.45347	-3.489	-5.55003
0.309434	0.350951	3.402604	-2.33902	-3.05206	1.816885	0.311169	-9.37959	-5.42303	-3.52365	-3.63843	-6.66381	-3.65583
0.096093	-0.01385	2.120751	4.608487	4.580226	4.63639	5.150503	5.361192	1.245736	3.35396	1.266714	-0.22235	-3.40084
0.479335	0.297649	3.422805	-2.43284	-1.42488	1.284808	-2.80436	-2.76813	1.642407	-0.9921	10.87455	6.09885	9.286741

4.2 Dynamic Time Warping (DTW)

DTW algorithm is based on Dynamic Programming. This algorithm is used for measuring similarity between two time series which may vary in time or speed.

The digitized speech samples were processed using MFCC to produce voice features. After that, the coefficient of voice features can go through DTW, to select the pattern

To recognize proper sentence in the database. [12]

DTW algorithm is based on Dynamic Programming. This algorithm is used for measuring similarity between two time series which may vary in time or speed.

The technique also used to find the optimal alignment between two times series if one time series may be “warped” non-linearly by stretching or shrinking it along its time axis. This warping between two time series can then be used to find corresponding regions between the two time series or to determine the similarity between the two time series.[13]

4.2.1 Dynamic Time Warping

In fig. 5 one can see the matching of the template with input using the algorithm of DTW.

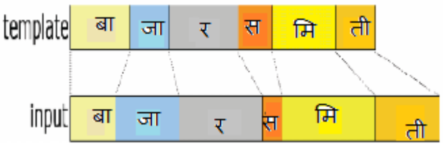


Figure 5: Template for Dynamic Time warping

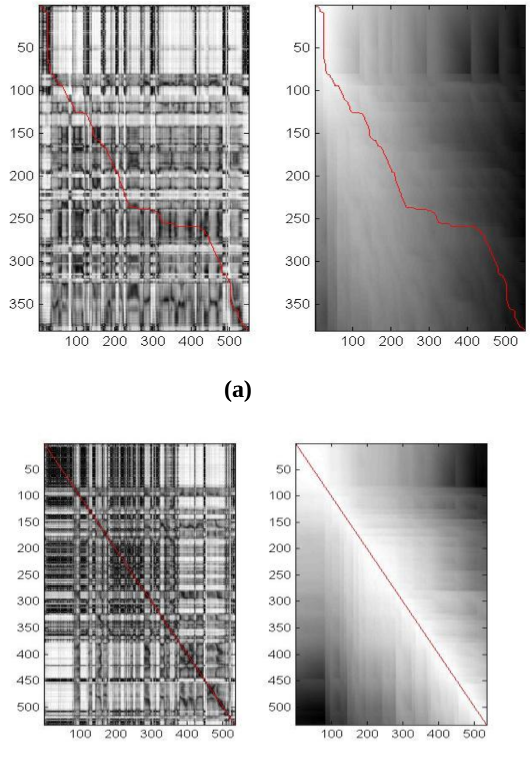
Speech recognition systems distinguish between two kinds of variability: acoustic and temporal.

Acoustic variability covers different accents, pronunciations, pitches, volumes, and so on, while ***temporal variability*** covers different speaking rates. These two dimensions are not completely independent — when a person speaks quickly, his acoustical patterns become distorted as well — but it's a useful simplification to treat them independently. Of these two dimensions, temporal variability is easier to handle. An early approach to temporal variability was to linearly stretch or shrink (*"warp"*) an unknown utterance to the duration of a known template. Table 2 shows the values obtained for the question 1 for the subject1.[6]

Table 2: DTW value, Comparison of speech signals

		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
	Question	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00
Q1	1.00	0.00	168.79	207.28	200.33	264.80	264.80	209.49	161.97
Q2	2.00	168.79	0.00	200.56	135.29	130.77	142.20	148.85	143.47
Q3	3.00	207.28	200.56	0.00	165.94	209.38	193.94	203.93	210.45
Q4	4.00	200.33	135.29	165.94	0.00	152.84	157.68	150.11	152.80
Q5	5.00	264.80	130.77	209.38	152.84	0.00	139.00	144.98	115.48
Q6	6.00	264.80	142.20	193.94	157.68	139.00	0.00	152.70	131.89
Q7	7.00	209.49	148.85	203.93	150.11	144.98	152.70	0.00	150.35
Q8	8.00	161.97	143.47	210.45	152.80	115.48	131.89	150.35	0.00

The Table 2 shows the comparison of time using DTW in which eight training sentences with eight testing sentences of extracted MFCC features of the given subject 1. Figure 6 (a),(b) shows the pictorial representation of the values obtained.



(a)

(b)

Figure 6 (a), (b): The image showing the training and testing sentences comparison

5. CONCLUSION

The experimental work carried out in this research is performed on the Speech Database in rural Marathi, which was created especially for this work using FAQs from "Krishi Utpanna Bajar Samiti"'s daily transactions at taluka place. The work is useful to develop an IVR system in Marathi language that will serve the need of rural agricultural queries regarding selling and buying of Agri goods.

After creating the database, features of every sentence were extracted and processed. Graphical representation of each sentence was also done. The statistical features related to MFCC were extracted.

The MFCC features were fed to DTW, for the matching purpose. It was to able to recognize the given sentence with satisfactory recognition rate. These both algorithms, MFCC and DTW, have been found to be efficient for this type of application.

6. ACKNOWLEDGEMENT

We are thankful to Dr.R.R.Deshmukh, Head of the department, for his support and encouragement.

7. REFERENCES

[1] https://www.hcii.cmu.edu/courses/speech-recognition-and-understanding.

[2] "Speech Recognition on DSP: Algorithm Optimization and Performance Analysis" , YUAN Meng, The Chinese University of Hong Kong, July 2004.

[3] Intervoice Speech -enabled IVR Systems", *Volume* 50, India HCI '14 Proceedings of the India HCI 2014 Conference on Human Computer Interaction.

[4] P.Laxminarayana,A.V.Ramana,Mythilisharana,A.Srikanth, B.Sandeep kumar and J.Mounika "Automatic Speech Recognition, Tutorial & Lab Manual", Prepared by,

Research and Training Unit for Navigational Electronics Osmania University, Hyderabad, INDIA - 500 007.

[5] Jay E. Coop , "IVR: The History and Future Of Speech Recognition", January 21, 2011 http://ezinearticles.com/?expert=Jay_E._Coop

[6] "A comparative study of speech and dialed input voice interfaces in rural India", DOI: 10.1145/1518701.1518709 Conference: Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Boston, MA, USA, April 4-9, 2009.

[7] "Villgro Innovations Foundation Case Study Series" – Uniphore Software Systems, 2012.

[8] "Evaluation and error recovery methods of an IVR based real time speech recognition application", Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference.

[9] Santosh A. Kulkarni1, Dr. A.R.Karwankar,"IVRS FOR COLLEGE AUTOMATION", , International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 6, August 2012, ISSN : 2278 – 1021.

[10] Bharti Gawali, Ganesh Janawale, S.C. Mehrotra ,"Marathi Isolated Word Recognition System Using MFCC Features", Department of Computer Science & IT, Dr.B.A.M.University, Aurangabad, International Journal on Information Technology , Mar2011, Vol. 1 Issue 1, p21.

[11] Mugdha Parande, Prof. Shanthi Therese , Prof. Vinayak Shinde, " Government Policies Search using Marathi Speech Recognition System– Based on MFCC with Gammatone Filter" ,International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Web Site: www.ijettcs.org Email: editor@ijettcs.org , Volume 4, Issue 3, May-June 2015

[12] Hemakumar G., Punitha P., "Speech Recognition Technology: A Survey on Indian Languages", International Journal of Information Science and Intelligent System, Vol. 2, No.4, 2013.

[13] Saurabh Chatterjee ,Project guides: Harish Karnick, Srinivasan Umesh, Speech Recognition in Indian Languages Btp term1 report