IJESAR
International Journal of Engineering Science and Generic Research (IJESAR)
Available Online at www.ijesar.in
Volume 1; Issue 1; 2015; Page No. 14-20

Automatic digit recognition and synthesis system for marathi

Professor, Bharti Gawali

Professor, Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad.

Abstract

In this work we cohabite the automatic recognition process with the automatic synthesis process applied on the first ten Marathi digits in one system that we called Automatic Recognition and Synthesis System of Marathi Digits (ARSSAD). The system is then composed of two sub-systems; a recognizer and a synthesizer. The main task of the recognizer is the automatic recognition of the pronounced digit, so it transforms the input sound wave into a text representing the appropriate digit, the second sub-system perform the opposite process of the first sub-system; in another word, it transforms the text (digit) produced by the recognizer to a speech generated by the synthesizer. The methodology used for the system design is based on three essential stages: the creation of the acoustic database (corpus development), the recognition of the read signal and the generation of the synthetic speech. We explain the basics modules that compose it, starting from the signal acquisition and finishing to the decision taken. For the recognition sub-system we make the choice to use the Dynamic Time Warping (DTW) method for the comparison task. ARSSAD contains a Front-End and a back-end module, the front-end module convert the input sound into feature vectors that are based on Mel Frequency Cepstral Coefficients (MFCCs), to be used in the DTW method. The back-end module uses the concatenative method to perform the synthesis of the recognized digit, for this end we create a sound database that contained di-phones of the Marathi alphabets. The obtained results show that the system presents a success rate of 94.85% on the three corpuses which we recorded in a noised environment.

Keywords: analysis techniques, speech recognition, speech synthesis, synthesis by di-phones, synthesis by phonemes, PRAAT, MFCC, DTW, Standard Marathi.

I. INTRODUCTION

The automatic speech processing (ASP) is an area of research for which a significant effort has been undertaken over the past three decades. The challenges are considerable and have fundamental nature. They are also multidisciplinary: signal processing, pattern recognition, artificial intelligence, computer science, phonetics, linguistics, ergonomics and neurosciences; which behave at varying degrees in the solutions found [1].

These long-standing works nevertheless give birth at the present time to intermediate products which find their place in practical applications in the context of the Man-machine communication, as shown in [2], [3] and [4]. However, Automatic Speech Recognition/Synthesis (ASRS) systems dedicated to the Marathi language are at the moment still very modest. In this article, we will introduce our ASRS system of the first ten digits of the Standards Marathi language (SA) in mono mode speaker. We are interested exclusively to the step of analysis of the speech signal which allows us to extract the acoustic vectors characterizing it. This step is very important and primordial in the process of automatic recognition, since it produces in output a set of parameters considered pertinent and efficient for the high-quality operation of the speech signal, on this same set we will apply the algorithms of recognition and comparison [5].

In speech recognition, the step of feature extraction, commonly known as the step of analysis, can be achieved in several ways. Indeed, the acoustic vectors are usually extracted using methods such as temporal encoding predictive linear (Linear Predictive Coding LPC) or Cepstral methods as the MFCC encoding (Mel Frequency Cepstral Coding), as well as the encoding PLP (Perceptual Linear Predictive coding) which is an example of the application of knowledge of the auditory system in human speech recognition. The extraction of characteristics is a key element for the development of an ASR system. The other part of our system represent a Text To Speech (TTS) system, in which the main techniques used in it design are Articulator synthesis, Formant synthesis, and Concatenative synthesis [6] [7].

- Articulatory synthesis attempts to model the human speech production system directly.
- Formant synthesis, which models the pole frequencies of speech signal or transfer function of vocal tract based on sourcefilter-model.

Corresponding author: Professor, Bharti Gawali

➢ Concatenative synthesis, which uses different length pre-recorded samples derived from natural speech.

In our case, we have used the concatenation method for the synthesis implementation which represent, in our opinion, the method that produce a synthetic voice the most natural and intelligible compared to the others. This result came from the fact of using a set of recording units pronounced by a real speaker, priory collected and embedded within our sound database [8]. So, for the recognizer, we have to deal with two essential problems, the first one is the choice of the technique of analysis used, and the second one is the choice of parameters and their number to extract the relevant parameters of the voice signal. The purpose is to determine which gives the best recognition rate [9] [10]. Whereas, for the synthesizer, we have to face two other problems; the choice of the transcription method (rule-based method or lexicon-based method) in one hand, and the co-articulation problem to improve the quality of the generated speech, in the other hand [11].

II. SYSTEM DESIGN

When designing a system, two broad ways could be taken into account, the first one is to design the whole system using the known theories, and use it as it is designed, in the real conditions. An alternative way would be to subdivide the system into modules that can be independently created and tested, to eventually be used in other systems to perform several functionalities.

To facilitate the implementation and improvement of our system, we have used the modular approach; this concept makes the program understandable on one part and decreases the cost of development of each module in another part. We have also used the concept of the object-oriented programming which is particularly suitable with the modular technique. We must therefore make out different modules which structure the system as shown in the following diagram:

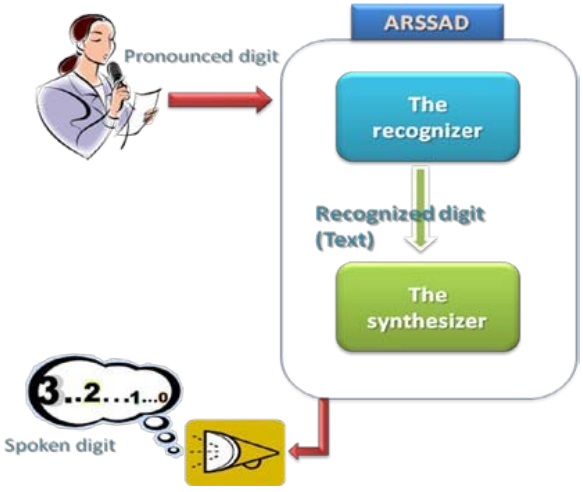

Figure 1: General architecture of our system ARSSAD.

The objective here is to describe the role of each module, explaining in the same time the interest of links which provide the cooperation between them [9][10].

III. THE RECOGNIZER

This module represent the front-end of the whole system, it is also composed of a set of sub-modules that can be shown in the following diagram

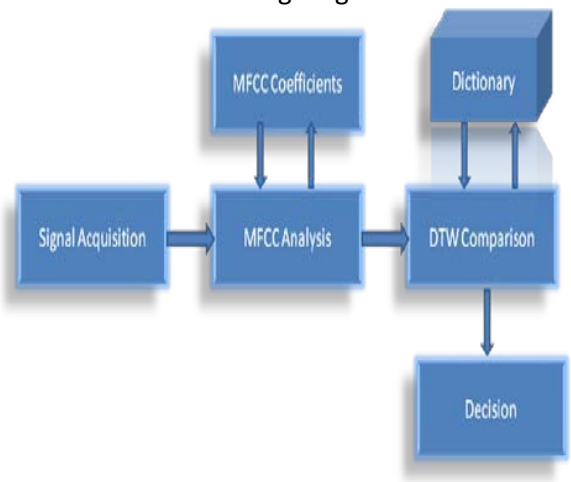

Fig. 2 general schema of the recognizer

We give now the principal functionalities of each sub-module one after another.

A. *Signal Acquisition*

This module carries out the acquisition of the acoustic signal recorded by a microphone and converts it into a digital form that can be used directly by a machine. There are many types of microphones but all of them provide the same function: transform the pressure fluctuations caused by the acoustic wave of speech into an electrical signal. This signal will be converted from the analogical form to the digital one, i.e. it will be discrete both in time (sampling) and value (quantification) [2]. As a result we obtain a digital signal in the form of a sequence of samples which measure the amplitude of the microphone's signal in regular spaced moments, and the amplitude of each sample is represented in its digital form. The choice of the sampling frequency is usually determined by the application and referred to the platform used [3]

B. *Sampling Frequency*

Some thoughts on the frequency of sampling are required in first. According to the theorem of Shannon [4]: « a bandlimited function can be perfectly reconstructed from a countable sequence of samples if the bandlimit, B, is no greater than half the sampling rate (samples per second)".

Sounds that are made by the human voice normally contain relatively insignificant frequency components at

or above 10 kHz [1]. Sampling such an audio signal at 20k sam/sec, or more, provides an excellent approximation to ensure that the Shannon criterion is met. But often the sample-rate is pre-determined by other considerations, such as an industry standard format (e.g. 8k sam/sec). In this situation, the human voice should be filtered, to remove frequency components above 4 kHz, before being sampled. So we consider in our work that the acoustic signal is located mainly in the bandwidth (50 Hz -8 kHz), the frequency of sampling should therefore be at least equal to 16 kHz, according to the theorem of Shannon. For the case of our application, we have used a sampling frequency of the order of 22050Hz, the default value taken by the software used in this operation PRAAT [12].

C. *The corpus preparation*

Most of the works carried out in the field of ManMachine communication often require the registration, and the manipulation of corpus of continuous speech, and this to carry out the studies on the contextual effects, on the phonetic indices, and on the variability intra and inter speakers. There were three recorded corpuses each one containing ten sounds of ten prime numbers of Standards Marathiin a noisy environment and we have changed the speed of elocution from a corpus to another without changing the speaker. The step of analysis may therefore begin.

D. *The Cepstral Analysis (MFCC)*

The aim of the analysis of the voice signal is to extracts the acoustic vectors which will be used in the stage of recognition follows. In this step the voice signal is transformed into a sequence of acoustic vectors on the way to decrease redundancy and the amount of data to be processed. And then a spectral analysis by the discrete Fourier transform is performed on a signal frame (typically of size 20 or 30 ms) [1]. In this frame, the voice signal is considered to be sufficiently stable and we extract a vector of parameters considered to be enough for the good operation of the voice signal, in our work we choose to use MFCCs coefficients resulting from a Cepstral analysis of the read signal.The method of extraction of the MFCCs coefficients is one of most popular of calculation of the acoustic vectors in the field of automatic speech recognition. We have also decided to use it in our context of application and we chose a set of 12 coefficients. We expose the different steps leading to Cepstral analysis using the tool of speech analysis, PRAAT, we show the different parameters required for the analysis that we have chosen, and in the end the exploitation of the MFCCs coefficients resulting [17].

Step1: reading the file to analyze and the choice of MFCC method

- Start PRAAT
- Open the sound file:
- Read > Read from file (open a sound file)
- Edit (for the view)
- File > Extract Selection (for "cut" the sound)
- Write > Write to .wav file (to save a sound file)
- Select the file to analyses
- Choose the Cepstral method:
- Formants&LPC > To MFCC

Step 2: determination of the parameters required for the analysis

- Number of coefficients: 12
- Duration of windows: 30 ms
- Duration between the windows: 10 ms

Step 3: analysis Results

It remains now to save the results in a text file format with the extension .MFCC (Write > Write to txt file), to be used in the following stage.

E. The use of DTW method

Our speech recognition sub-system is based on the algorithm of DTW (Dynamic Time Warping), this method is based on an evaluation of the distance between an observation and a list of references (dictionary). As well the reference for which this distance is minimal allows us to decide what word is it. The evaluation of the distance between two signals is not performed with the signals themselves. This would lead to lot of calculations. It is therefore in a prime time to find a better representation of the signals. Here MFCC analysis shines. So we have programmed the DTW method using, for the comparison, the MFCC coefficients. The training part concerns the recording of the sounds corpuses in order to design our dictionary which will be used as reference in the comparison of the signals tested. Problems of recognition may appear depending on the conditions in which the signal to test is recorded. If the word is pronounced more or less close to the microphone recognition rates can vary greatly. However if the user says the word always at the same distance and with the same intensity, the rate of recognition is very acceptable. We judge, however, that the representation using the MFCC coefficients provides better results, and it supports better the limitations related to the problem of the capture of the signal. The common skeleton of the DTW algorithm has three steps illustrated as follow:

- Acquisition of the sound file to test
- Extraction of the MFCC coefficients
- Comparison with the dictionary of references *F. The decision*

This last module of our recognizer plays two essential roles; it represents the interface in which the user

interacts with the system. After the user has entered his voice signal, he starts the search and awaits the results. The system displays the recognized digit written in both Marathi and French language. The second role is that this same decision (the displayed digit) represents the input (text) of the second module of our system, which is the synthesizer [18][19][20].

IV. THE SYNTHESIZER

It is based essentially on two principal parts; a frontend and a back-end. The front-end is composed of two modules, the first is for the sound database creation and the second is for the conversion text-to-phoneme or grapheme-to-phoneme. The back-end part represents the speech generation module or in other words the synthesizer itself. So the different modules that compose the system are as follow:

- The sound database creation (segmentation): we have recorded a set of pieces of speech and store it in our database, this set is composed of phonemes and di-phones which are the basic units utilized within the back-end module in order to generate voice using the concatenation method.
- The grapheme/phoneme conversion: before achieving this process, a text normalization or preprocessing operation has to be done. After that the module assigns to each word in entry it phonetic transcription, and then divides and marks the text into prosodic units like syllables. This process of assigning phonetic transcription to words is called text-tophoneme or grapheme-to-phoneme conversion. The output of the front-end module is a symbolic linguistic representation resulting from the phonetic transcription and prosody information together, which represents the input of the back-end module.
- The synthesizer: the back-end module uses information provided by the front-end to converts the symbolic linguistic representation to speech using a specific method. In literature, there are two kind of synthesis method; rulebased method and concatenative corpus-based method.

Like we have mentioned before, we have used the concatenative method of phonemes and graphemes previously stored in our sound database [13][14][15]. The general architecture of this module could be shown in figure 1 as follow:

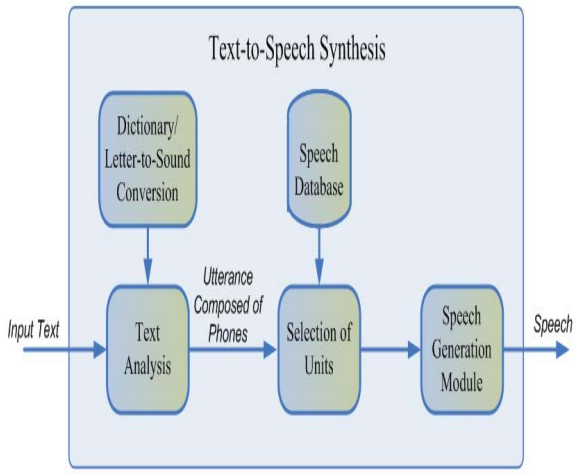

Fig. 3 the general architecture of the synthesizer.

A. The corpus description

We have created two corpuses; The first contains phonemes: It is composed of a set of basic sounds (which consists of the phonemes corresponding to the 28 consonants and 6 vowels, and other additional character. To improve the quality of the words synthesized by the method of concatenation of phonemes, and to reduce the effects of co-articulation, the solution is to record the transition that exists between phonemes instead of recording the phonemes themselves; we talk about di-phones which are an adjacent pair of phones. Indeed, the transition (di-phones) is the bearer of a significant quantity of acoustic information in relation to the phoneme itself. Each transition or di-phone also varies from the stable part of a phoneme up to the stable part of phoneme that follows.

B. The phonetic and orthographical transcription "POT":

Transcription provides a phonetic text from the alphabetic text. To accomplish this, it must apply to many pronunciation rules. French language has a few thousands of basic rules; English language has tens of thousands of rules. Therefore, during the passage from the written form to the spoken form two approaches can be used which are: the lexicon-based approach and theRule-based approach [21].

•The use of rules

In this approach each grapheme is converted to phoneme depending on the context and this is thanks to the use of a set of rewriting rules [12]. The main advantage of this approach is the ability to model the linguistic knowledge of human beings by a set of rules that can be incorporated in expert systems. Each of these rules has the following form:

[Phoneme] = {LC (Left Context)} +{ C (Character)} + {RC context)}

Our transcription module grapheme-phoneme is based on a set of rules;

The rule of tanwin, al madd, etc… Prioritized, and organized in the form of a tree list. Each rule is written in the graphics context in which it is applied.

Here is a concrete example of transcription rule

" The rule of Tanwin " *If (grapheme[char]== 'Tanwin')*

{ If (API[position][0][= = '') Phoneme = phoneme + "an";

Else

{ If (API[position][0][= = '') Phoneme = phoneme + "in";

Else

Phoneme=phoneme+ "a"; }

}

•The use of the lexicon

In this case we must assign to each word in entry the pronunciation which corresponds to it without taking into account its context. The speed, flexibility and simplicity are the main advantages of this approach.

V. RESULTS

The main interface of our system, with an example of the recognition of the digit ten. We have applied the recognition on a corpus containing Marathi digits from one to ten pronounced by a male sex speaker in the Standard Marathi language. To evaluate the performance of our system, we have illustrate two formula for each module; a recognition rate (RR) for the recognizer, and a success rate (SR) for the synthesizer. We have fixed the number of tests performed to recognize a digit to twelve times. The recognition rate (RR) for each digit is calculated by the following formula:

$$RR = \frac{Nb\_recognized\_digit}{Nb\_tested\_digit} * 100\%$$

In the other hand, to calculate the success rate (SR) associated with each digit tested; we got the following formula:

$$SR = \frac{Nb\_well\_prononced\_digit}{Nb\_tested\_digits} * 100\%$$

The results obtained for each digit are summarized in the following table:

Table 1: Recognition/ Success rate for the ten digits

The Word in Marathi	Transcript	The word in Hindi	The word in English	RR	SR
१	एक	एक	One	85%	99%
२	दोन	दो	Two	99%	99%
३	तीन	तीन	Three	99%	88%
४	चार	चार	Four	99%	99%
५	पाँच	पाँच	Five	99%	99%
६	सहा	छहः	Six	89%	99%
७	सात	सात	Seven	99%	88%
८	आठ	आठ	Eight	99%	80%
९	नऊ	नौ	Nine	99%	82%
१०	दहा	दस	Ten	92%	98%

When investigating across the natural language processing field, we haven't found a lot of works dealing with the automatic recognition and speech synthesis in a same work, especially for the Marathi language. Therefore, in the comparison with previous works, we take into account just the success accuracy of the automatic recognition. The comparison results obtained are summarized in the following table:

Table 2: Comparison with previous work

ASR using CMUSphinx	85.55 %

DTW-Based ArSR	86%
Heuristic Method	86.45 %
Monophone-Based ArSR	90.75 %
VQ and HMM Rrna	91%
MCCF-based FPGA Rrna	95 % -98%
ARSSAD	**95%**

The recognition sub-system achieved 95% correct digit recognition in the case of mono-speaker mode. On the other hand, the speech synthesis sub-system achieved 89% correct well synthesized digit. So the system present in general 90, 82 % of success rate.

VI. CONCLUSIONS

We set several objectives for this research: that of discover the definitional character of the human voice, to describe the various stages and components used in the production of the voice and to dissect an ASRS system in its main floors. To that end, we have detailed our system of recognition and synthesis of Marathi digit as well as the results obtained. The system presents, using isolated words and in the absence of noise, a success rate quite honorable and acceptable. The acoustic variability of the voice signal, and in particular that due to the effects of coarticulation, is better apprehended by the modeling of its production. In fact, the voice signal is not an ordinary acoustic signal and the Anatomical constraints may explain the effects of coarticulation, for example, in the framework of the articulatory phonology. At the end of this rapid assessment on the voice recognition and synthesis, it has been noted that this area is particularly broad and that there is no miracle product capable of responding to all applications. The noise, for example, remains a brake to the generalization of recognition systems. The voice recognition is still a compromise between the size of the vocabulary, its possibilities multi-speaker, its rapidity, training time, etc... The power of the current calculating tools and the integration capabilities of systems have caused a resurgence of interest in the recent years among the industrials. In fact, they see in the voice recognition or synthesis, "the more commercial ", allowing making the difference with the competition. A quick tour of horizon on the very numerous publications allows us to set the ideas on the nature of the work in progress. Apart from the products dedicated to the voice recognition, the systems with analytical approach (HMM and ANN) give today the best results and currently have the wind in their sails. As regards the future prospects, the optimism is more measured than in the past. Without risk, we can say that the general problem of the automatic processing of the voice signal will probably not rule before the middle of the next century. We can as even quote a few perspectives to our work in the following points:

- Enlargement of the vocabulary for all digits;
- Recognition of continuous speech;
- Recognition in speaker independent mode;
- Use of the HMM, neural networks and hybrid methods.

REFERENCES

1. Sangramsing N.kayte "Marathi Isolated-Word Automatic Speech Recognition System based on Vector Quantization (VQ) approach" 101th Indian Science Congress Jammu University 03th Feb to 07 Feb 2014.
2. Monica Mundada, Bharti Gawali, Sangramsing Kayte "Recognition and classification of speech and its related fluency disorders" International Journal of Computer Science and Information Technologies (IJCSIT)
3. Monica Mundada, Sangramsing Kayte "Classification of speech and its related fluency disorders Using KNN" ISSN2231-0096 Volume-4 Number-3 Sept 2014
4. Monica Mundada, Sangramsing Kayte, Dr. Bharti Gawali "Classification of Fluent and Dysfluent Speech Using KNN Classifier" International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 9, September 2014
5. Sangramsing Kayte, Monica Mundada, Dr. CharansingKayte "Marathi Text-To-Speech Synthesis using Natural Language Processing "IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 63-67e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
6. Sangramsing Kayte, Dr. Bharti Gawali "Marathi Speech Synthesis: A review" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 3 Issue: 6 3708 – 3711

7. Sangramsing Kayte, Monica Mundada, JayeshGujrathi, "Hidden Markov Model based Speech Synthesis: A Review" International Journal of Computer Applications (0975 – 8887) Volume 130 – No.3, November 2015
8. Sangramsing Kayte, Monica Mundada, Dr. CharansingKayte "Implementation of Marathi Language Speech Databases for Large Dictionary" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 40-45e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197
9. Sangramsing Kayte "Transformation of feelings using pitch parameter for Marathi speech" Sangramsing Kayte Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part - 4) November 2015, pp.120-124
10. Sangramsing Kayte, Monica Mundada "Study of Marathi Phones for Synthesis of Marathi Speech from Text" International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-10) October 2015
11. Sangramsing N. Kayte, Monica Mundada, Dr. Charansing N. Kayte, Dr.BhartiGawali "Automatic Generation of Compound Word Lexicon for Marathi Speech Synthesis" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)Volume 5, Issue 6, Ver. II (Nov -Dec. 2015), PP 25-30e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
12. 12) Paul Boersma and David Weenink, "PRAAT: doing phonetics by computer" Phonetic Sciences, University of Amsterdam Spuistraat 210, 1012 VT Amsterdam the Netherlands.
13. Sangramsing Kayte, Monica Mundada, Dr. CharansingKayte "Di-phone-Based Concatenative Speech Synthesis System for Hindi" International Journal of Advanced Research in Computer Science and Software Engineering -Volume 5, Issue 10, October-2015
14. Sangramsing Kayte, Monica Mundada, Dr. CharansingKayte "Di-phone-Based Concatenative Speech Synthesis Systems for Marathi Language" OSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 5, Ver. I (Sep –Oct. 2015), PP 76-81e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197
15. Sangramsing Kayte, Monica Mundada, Dr. CharansingKayte "A Corpus-Based Concatenative Speech Synthesis System for Marathi" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 20-26e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197
16. Sangramsing Kayte, Monica Mundada, Dr. CharansingKayte "A Marathi Hidden-Markov Model Based Speech Synthesis System" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 34-39e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197
17. Sangramsing Kayte "Duration for Classification and Regression Tree for Marathi Text-to-Speech Synthesis System" Sangramsing Kayte Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part-4) November 2015
18. Sangramsing Kayte, Monica Mundada, Santosh Gaikwad, Bharti Gawali "PERFORMANCE EVALUATION OF SPEECH SYNTHESIS TECHNIQUES FOR ENGLISH LANGUAGE " International Congress on Information and Communication Technology 9-10 October, 2015
19. Sangramsing Kayte, Monica Mundada,Dr. CharansingKayte " Performance Evaluation of Speech Synthesis Techniques for Marathi Language " International Journal of Computer Applications (0975 – 8887) Volume 130 – No.3, November 2015
20. Sangramsing Kayte, Monica Mundada, Dr. CharansingKayte " Performance Calculation of Speech Synthesis Methods for Hindi language IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 13-19e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197
21. Sangramsing N. Kayte ,Monica Mundada,Dr. Charansing N. Kayte, Dr.BhartiGawali "Approach To Build A Marathi Text-To-Speech System Using Concatenative Synthesis Method With The Syllable" Sangramsing Kayte et al.Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part-4) November 2015, pp.93-97