International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 02 Issue: 08 | Nov-2015 www.irjet.net p-ISSN: 2395-0072

The Marathi Text-To-Speech Synthesizer Based On Artificial Neural Networks

Sangramsing N. Kayte[1], Dr.Bharti Gawali[1]

[1]*Department of Computer Science and Information Technology Dr. Babasaheb Ambedkar Marathwada University, Aurangabad*

Abstract - ***The research paper rapid advancement in information technology and communications, computer systems increasingly offer the users the opportunity to interact with information through speech. The interest in speech synthesis and in building voices is increasing. Worldwide, speech synthesizers have been developed for many popular languages English, Spanish and French and many researches and developments have been applied to those languages. Marathi on the other hand, has been given little attention compared to other languages of similar importance and the research in Marathi is still in its infancy. Based on these ideas, we introduced a system to transform Marathi text that was retrieved from a search engine into spoken words. We designed a text-to-speech system in which we used concatenative speech synthesis approach to synthesize Marathi text. The synthesizer was based on artificial neural networks, specifically the unsupervised learning paradigm. Different sizes of speech units had been used to produce spoken utterances, which are words, di-phones and tri-phones. We also built a dictionary of 1000 common words of Marathi. The smaller speech unit's di-phones and tri-phones used for synthesis were chosen to achieve unlimited vocabulary of speech, while the word units were used for synthesizing limited set of sentences.***

Key Words: *Artificial neural networks, di-phones, tri-phones, text-to-speech synthesis, concatenative synthesis, signal processing.*

1. INTRODUCTION

A Text-To-Speech synthesizer is a computer-based program in which the system processes through the text and reads it aloud. For most applications, there is a demand on the technology to deliver good and acceptable quality of speech. The quality of a speech synthesizer is judged by its similarity to the human voice naturalness and by its ability to be understood intelligibility. High quality speech synthesis finds a wide range of applications in many fields, to mention a few [1]: Telecommunications services, Language education, Multimedia applications and Aid to handicapped persons [1]. The speech synthesizer consists of two main components, namely: the text processing component and the Digital Signal Processing module [6-9]. The text processing component has two major tasks [2]. First, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words, this process is often called text normalization. Then it converts the text into some other representation and output it to the DSP module or synthesizer, which transforms the symbolic information it receives into speech [2] [3].

The primary technologies for generating synthetic speech waveforms are formant synthesis and concatenative synthesis [1]. Each technology has strengths and weaknesses and the intended uses of a synthesis system will typically determine which approach is used. The speech synthesizer that we built in this work depends on the concatenative synthesis approach. In concatenative synthesis the waveforms are created by concatenating parts of natural speech recorded by humans. The easiest way to produce intelligible and natural synthetic speech is to concatenate prerecorded utterances. But, this method is limited to one speaker and one voice and the recorded utterances require a larger storing capacity compared to the other methods of speech synthesis. In present systems, the recorded utterances are divided into smaller speech units, such as: words, syllables, phonemes, di-phones and sometimes tri-phones. Word is the most natural unit for the written text and suitable for systems with very limited vocabulary. Di-phones are two adjacent half-phone context-dependent phoneme realizations, cut in the middle and joined into one unit. Tri-phones are like di-phones, but contain one phoneme between steady-state points half phoneme-phoneme-half phoneme [18]. In other words, a tri-phone is a phoneme with a specific left and right context [4] [5].

2. TECHNIQUES

The general architecture of the Text-To-Speech system is shown in Fig. 1. The input to the system is the result of queering an existing search engine which is capable of retrieving Marathi textual data. The text-to-speech synthesis procedure consists of two main phases. The first phase is text analysis. In this phase the input text is pre-processed and then classified using artificial neural networks, we used unsupervised learning paradigm, specifically the kohonen learning rule. Such network can learn to detect the features of the input vector. The second phase is the generation of speech waveforms. Here, we use concatenative speech synthesis approach for this purpose. The post processing is used to smooth the transitions between the concatenated di-phones [10].

Text pre-processing: Before the words enter the neural network, a series of preliminary processing has to be fulfilled. At first, the punctuation marks are removed, then the numbers are identified and the abbreviations are expanded into full words. The next step is to fully diacritise the retrieved text to eliminate any ambiguity RAM-RAM the word's pronunciation. The final step is to prepare the words as input vectors for the neural network. However, neural networks only recognize numerical inputs, therefore, the ASCII code of each character is taken and replaced with its corresponding binary representation. Next the 0's were replaced with (-1)'s to discriminate them from trailing zeros that will be added later. Now the text is ready to be processed and classified by the neural network [2][11].

Text to speech conversion: When building a speech synthesizer, one has to decide which synthesis unit to choose. There are different unit sizes and each choice has its own advantages and disadvantages. The longer the unit the more accuracy you get, but at the expense of the number of data needed [4][5].

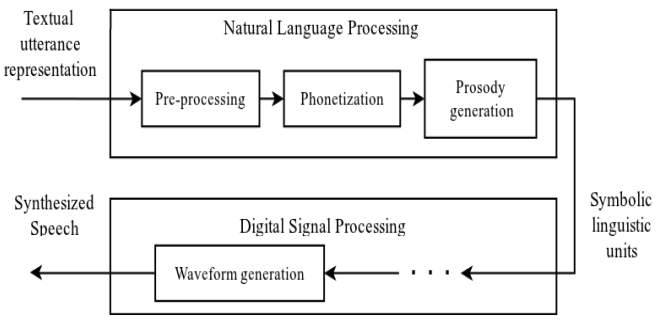


Fig. 1: The basic building of the Marathi TTS

The word model: Systems that simply concatenate isolated words or parts of sentences, are only applicable when a limited vocabulary is required typically a few hundreds of words and the sentences to be pronounced respect a very restricted structure. In this model, a dictionary containing 1000 words that are commonly used in Marathi is built [12] [17].

Training the words: The goal of this procedure is to generate the corresponding speech of each word in the dictionary. Since our database of speech doesn't contain complete words, we constructed each word out of its di-phone sequence. To train the words of the dictionary, each word is converted into its di-phone sequence then passed to the pre-processing unit as explained previously. Neural networks require that all inputs are of the same length, so we chose a vector length of 154 in regard to the longest word in the dictionary. Thus, words producing a vector shorter than 154 are padded with trailing zeros. Figure 2 shows the functional diagram of the training process, the input feature vector is passed to the network at the beginning. The neural network in turn produces a cluster representing the input. Then each cluster is passed to the converter module and is converted into a pattern of 1's and 0's for comparison purposes to be performed later. Now, the pattern is mapped to its corresponding speech signals and saved in a look-up table. This process is performed for all the words of the dictionary [12].

Synthesizing words: In this process the input text is tokenized into single words and each word is processed individually. Each word goes through the same training process to produce the feature vector and the output pattern. This pattern is then compared with the patterns in look-up table and classified by the Euclidean distance metric. At last, the recognized word is mapped to the corresponding sound and output as a speech [13] [14]. The synthesis procedure is shown in Fig. 3.

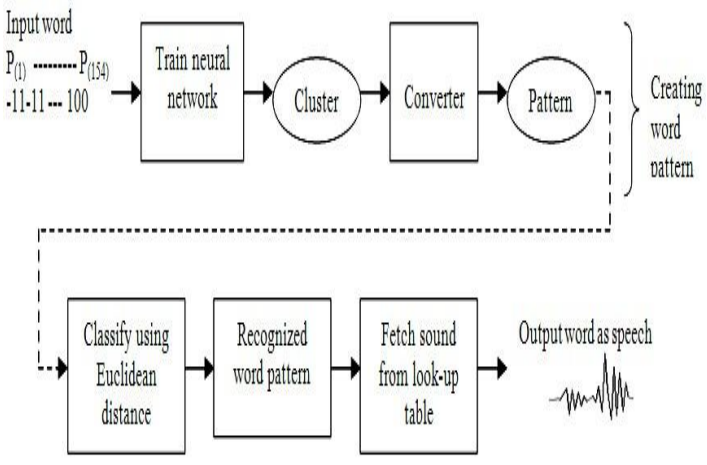


Fig. 3: Word synthesis model

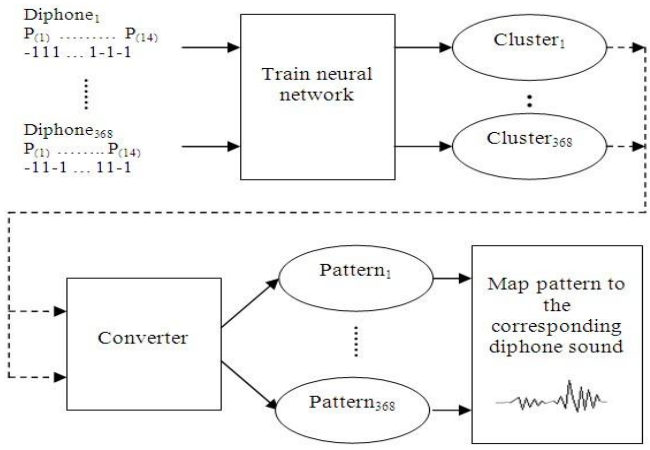


Fig. 4: Di-phone training model

If the Euclidean distance exceeds a certain threshold, this means that the word hasn't been recognized as one of the trained words. In this case, it will be added along with its corresponding speech to the look-up table [15][16][19].

The di-phone model: We need a more flexible model which adapts to any new input data. Thus, for unrestricted speech synthesis we have to use shorter pieces of speech signal, such as di-phones and tri-phones. The concept of this model is to use the di-phones as the speech synthesis units.

Di-phone database: This step aims to generate a mapping between the textual di-phones and their equivalent speech units. Each di-phone is represented by two characters, consequently producing a vector of 14 elements. The training process is similar to that of the word model, except that the produced pattern is mapped to the equivalent speech unit of that di-phone. This process is repeated for all the di-phones in the database. The training process is shown in Fig. 4.

Synthesis using di-phone units: To convert input words into speech, the words are automatically broken down to their di-phone sequence. Each di-phone will be converted into a feature vector then trained by the network to finally produce the pattern. This pattern is classified by the Euclidean distance and the corresponding di-phone speech is fetched. This process is repeated for all the di-phones. The output di-phone units are saved in a speech buffer until text reading is finished. After that, the speech segments are concatenated together to produce a spoken utterance as shown in Fig. 5.

The tri-phone model: This model uses longer segmental units (tri-phones) in attempt to decrease the density of concatenation points, therefore provide better speech quality. The di-phones in the speech database were used to build a database of 300 tri-phones, each tri-phone is built up by concatenating two di-phones.

Training tri-phones: This procedure is the same as the one used to train di-phones, with a difference of the size of the input and output units. A tri-phone is presented by three characters producing a feature vector of 21 elements. When the pattern is generated, it's mapped to the equivalent tri-phone speech unit. This process is repeated until the whole 300 tri-phones are trained.

Synthesis using tri-phones: To generate spoken utterances in this model, the words are automatically segmented into tri-phones. These tri-phones are converted into feature vectors of 21 elements and they go through the same procedure as the di-phones,

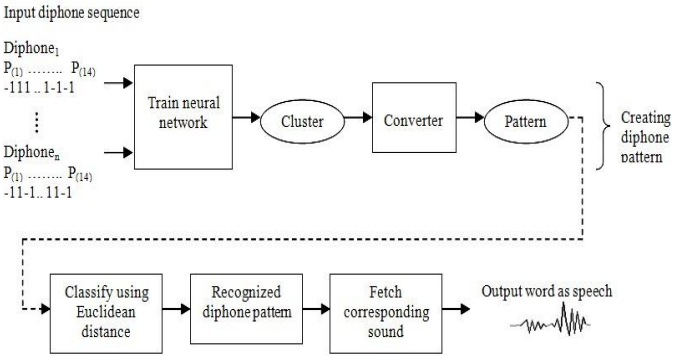


Fig. 5: Di-phone synthesis model

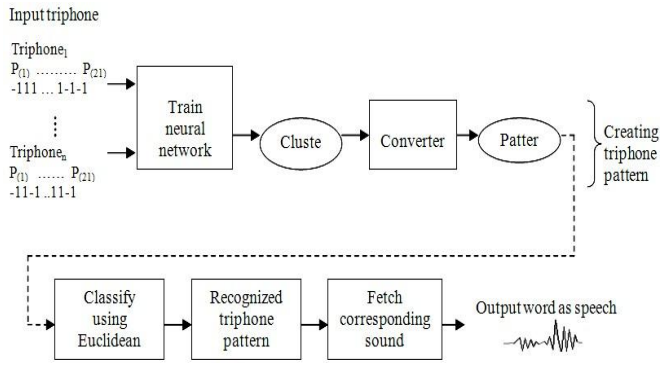


Fig. 6: Tri-phone synthesis model

but the output pattern is mapped to the equivalent tri-phone speech unit. At last, tri-phone segments are concatenated to produce the written sentence as speech. Figure 6 shows the components of the tri-phone synthesis system [20-23].

3. RESULTS

The proposed system was built and evaluated using the Matlab 14 programming language. To evaluate the accuracy of the synthesizer, different sets of sentences and words are input to the three models word, di-phone and tri-phone. In order to evaluate the quality of the system, a subjective listening test was conducted. The test sets were played to eight volunteer listeners (2 females and 2 males), which their ages range from 18-30 years. All the listeners are native Marathi speakers and have no experience in listening to synthesized speech. The speech was played by loudspeakers in a quiet room and each listener was tested individually. As a first step, a set of eight sentences was used to evaluate the word model, in which all the words were recognized by the neural network and output the right speech waveform. Then the output speech was played to the listeners in order to determine how much of the spoken output one could understand, the average of the recognized words by the listeners 98% was. Further, a larger set of sentences was built and tested by the model, but it wasn't evaluated by the listeners. The set contains 30 sentences including the eight sentences tested before. The sentences vary in length from short sentences to a paragraph. The average accuracy of the recognized sentences by the neural network is 99%, Fig. 7 shows the accuracy of each sentence. Finally, to test the whole set of words in the dictionary, the 1000 words were input to the neural network in sequence in four different runs. The average accuracy of the four runs is 99.05%.

To evaluate the di-phone model, a set of six sentences and nine discrete words where tested both by the neural network and by the listeners. The test conducted by the listeners is the Mean Opinion Score (MOS) test which provides a numerical indication of the perceived quality of received media after compression and/or transmission [7][20-23]. The rating scheme is described in Table 1

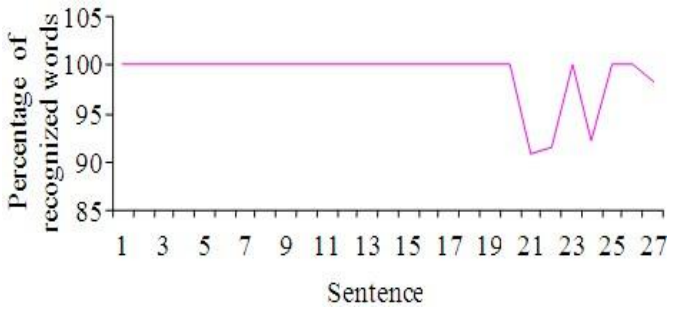
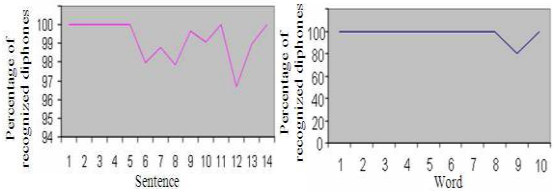


Fig. 7: Accuracy of the word model

Table 1: Mean opinion score rating scheme

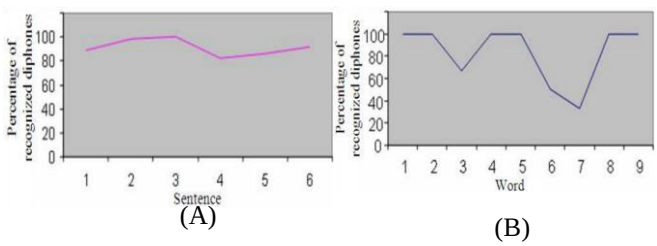
Bad	Description
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad



(A) **(B)**

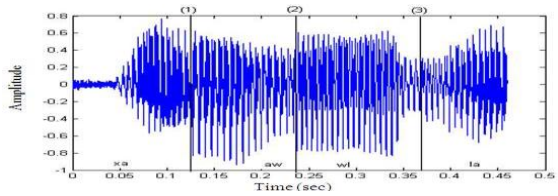
Fig. 8: Di-phone recognition accuracy (a): Sentences (b): Discrete words

The accuracy obtained by the neural network was 89% in recognizing the di-phones and the average rated score given by the listeners is 4.37. For further testing, a larger set was created and tested again by this model. The new set consists of fourteen sentences and ten discrete words including the set tested before. The new sentences were also created from words outside the dictionary. The accuracy of the recognized di-phones by the neural network is 91%. Figure 8 shows di-phone recognition accuracy for the new set of sentences and the discrete words. The same set used to evaluate the di-phone model the first time is used to evaluate the tri-phone model. The recognition accuracy of the six sentences and nine words obtained by the neural network is 82%. This result is not as good as the ones obtained by the previous two models. This is due to the small number of tri-phones in our database, which doesn't cover a wide range of tri-phone combinations. The tri-phone recognition accuracy is shown in Fig. 9. When applying interpolation on the output speech, the results showed that the linear interpolation made no changes on the signal. Meanwhile the spline interpolation did have an effect but it's not the desired one since this kind of interpolation caused the signal to oscillate. The cubic interpolation could successfully smooth the transitions between di-phones, but it had a slight effect in improving the quality of the speech when it was played.

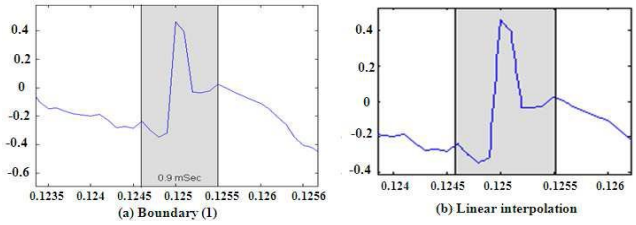


(A) (B)

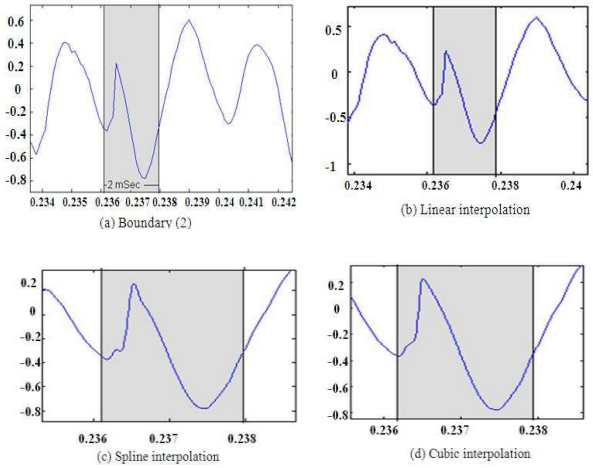
Fig. 9: Tri-phone recognition accuracy (a): Sentences (b): Discrete words



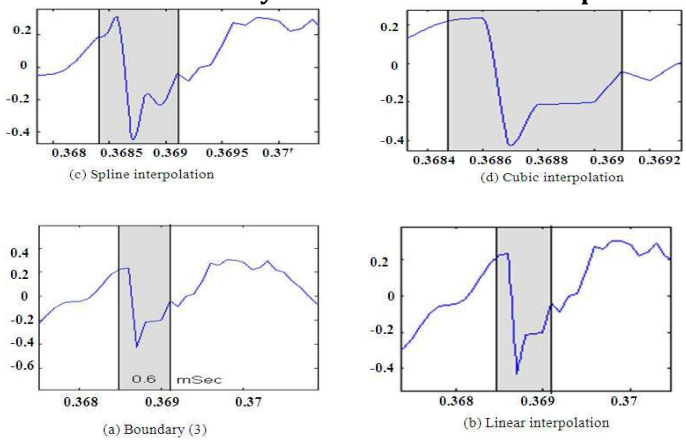
'Ram-Ram' waveform



Boundary between 1st and 2nd di-phones



"Ram-Ram" Boundary between 2nd and 3rd di-phones



Boundary between 3rd and 4th di-phones

Fig. 10: Interpolating the word "Ram-Ram" The reason of this shortcoming is the very small duration of the segments we processed where the longest interpolated time span is 2 m sec. which is not adequate to cause a perceptible change in the signal. Figure 10 shows the effect of interpolating the Marathi equivalent of the word "Ram-Ram".

4. CONCLUSION

In this research, we presented a Marathi text-to-speech synthesis system. Artificial neural networks with unsupervised learning paradigm where used to build the system and different types of speech units were used to synthesize the desired utterances, which are: words, di-phones and tri-phones. The experimental results over the system showed its ability to produce unlimited number of words with high quality voice and high accuracy in converting the written text into speech. Where the obtained accuracy by the word and di-phone models was 89% and by the tri-phone model was 82%.

REFERENCES

[1] 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

[2] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "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

[3] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "A Review of Unit Selection Speech Synthesis International Journal of Advanced Research in Computer Science and Software Engineering -Volume 5, Issue 10, October-2015

[4] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "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

[5] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "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

[6] 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.

[7] Monica Mundada, Sangramsing Kayte "Classification of speech and its related fluency disorders Using KNN" ISSN2231-0096 Volume-4 Number-3 Sept 2014

[8] 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

[9] 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)

[10] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "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

[11] 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

[12] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "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

[13] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte " 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

[14] Sangramsing Kayte, Monica Mundada,Dr. Charansing Kayte " Performance Evaluation of Speech Synthesis Techniques for Marathi Language " International Journal of Computer Applications (0975 – 8887) Volume 130 – No.3, November 2015

[15] Sangramsing Kayte, Monica Mundada,Dr. Charansing Kayte" Speech Synthesis System for Marathi Accent using FESTVOX" International Journal of Computer Applications (0975 – 8887) Volume 130 – No.6, November2015

[16] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "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, 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

[18] Sangramsing Kayte, Monica Mundada, Jayesh Gujrathi, " Hidden Markov Model based Speech Synthesis: A Review" International Journal of Computer Applications (0975 – 8887) Volume 130 – No.3, November 2015

[19] Sangramsing Kayte, Monica Mundada,Dr. Charansing Kayte "Screen Readers for Linux and Windows – Concatenation Methods and Unit Selection based Marathi Text to Speech System" International Journal of Computer Applications (0975 – 8887) Volume 130 – No.14, November 2015

[20] Sangramsing N. Kayte ,Monica Mundada,Dr. Charansing N. Kayte, Dr.Bharti Gawali "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

[21] Sangramsing N. Kayte, Dr. Charansing N. Kayte, Dr.Bharti Gawali* "Grapheme-To-Phoneme Tools for the Marathi Speech Synthesis" Sangramsing Kayte et al.Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part -4) November 2015, pp.86-92

[22] 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)November2015

[23] 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