OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

The Heart of Generating Hidden Markov Model Which Are Essential For Synthesis

Sangramsing N. Kayte, Dr.Charansing N. Kayte*, Dr.Bharti Gawali
Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India)
**Digital and Cyber Forensic, Aurangabad, Maharashtra.*

ABSTRACT:- *Speech synthesis is the process of production of artificial speech. The system used for generation of speech from text is called as text-to-speech system. In TTS system, text and voice models for a particular language or multiple languages are given as input to the system, which generates speech as output corresponding to the provided voice models. Speech synthesis systems can be extremely useful to people who are visually challenged, visually impaired and illiterate to get into the mainstream society. More recent applications include spoken dialogue systems and communicative robots. Hidden Markov Model based Speech synthesis is the emerging technology for TTS. HMM based speech synthesis system consists of training phase and synthesis phase. In the training part, phone and excitation parameters are extracted from speech database and modeled by context dependent HMMs. In synthesis part, the system will extract the suitable phone and excitation parameters from the previously trained models and generates the speech.*

KEYWORDS:- *HMM, parameters, database*

I. INTRODUCTION

A Text-to-Speech system synthesizes a speech signal corresponding to a given text by creating the grapheme to phoneme relation. Two of the major synthesis techniques are concatenative speech synthesis and HMM based speech synthesis [1]. A successful concatenative speech synthesis technique is the unit selection synthesis [2]. It involves combining the appropriate pre-recorded natural speech units corresponding to those present in the text, such that the target and the concatenation costs are minimized [3]. The speech units can be word or sub word units [4]. The quality of the synthesized speech varies with the size of the speech unit. With longer units, the naturalness is preserved to a greater extent in the synthesized speech and also the number of concatenation points is less [5]. However, the amount of data required to build such a system that yields a good quality of synthesized speech is very large. Multiple occurrences of each speech unit in different contexts should be available in the data [6] [8]. Many speech synthesis systems can synthesize high quality speech but cannot synthesize speech with various voice characteristics such as speaker individualities, speaking styles, emotions, etc [9] [10]. To obtain various voice characteristics in speech synthesis systems based on selection and concatenation of acoustical units, a large amount of speech data is necessary[7] . However, it is difficult to collect such speech data. In order to construct speech synthesis systems which can generate various voice characteristics, the HMM based speech synthesis system (HTS) was proposed. Like other data driven speech synthesis approaches, HTS has a compact language dependent module. So, it could easily be extended to other languages [3].

II. HIDDEN MARKOV MODEL

Hidden Markov Models are widely used statistical models for characterizing sequences of speech spectra and have been successfully applied to speech recognition systems. The performance of HMM based speech recognition systems has been improved by techniques which utilize the flexibility of HMMs: context-dependent modeling, dynamic feature parameters and speaker adaptation techniques [3]. The HMM based TTS system described in this thesis also uses HMMs as speech units. HMMs are categorized into discrete HMMs and continuous HMMs, which can model sequences of discrete symbols and continuous vectors respectively [11]. An HMM is a finite state machine which generates a sequence of discrete time observations. At each time unit i.e., frame, the HMM changes states according to state transition probability distribution and then generates an

observation Ot at time t according to the output probability distribution of the current state. Hence the HMM is a doubly stochastic random process model. An N-state HMM is defined by state transition probability distribution A= output probability distribution B= and initial state probability distribution. For convenience, the compact notation (A, B, π) is used to indicate the parameter set of the model [12].

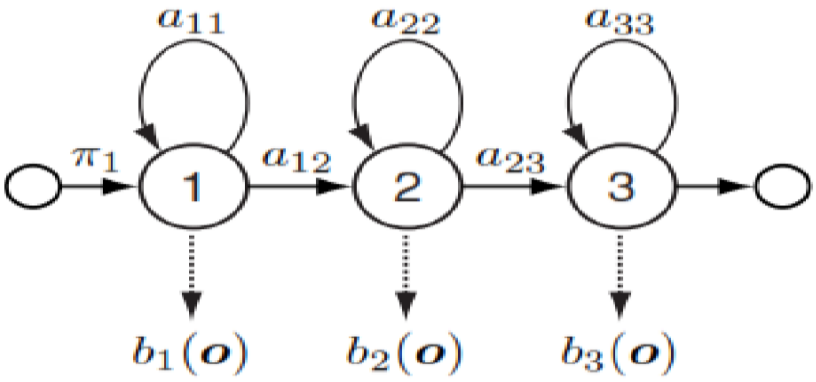

Figure 1: A 3-state left to right model

Figure 1 shows examples of HMM structure in which the state index increases or stays the same as the time increases. One common HMM topology is to use three states. Each state is linked to the next state and back to itself again. This last transition is known as the self-transition probability and is basically the probability that the next observation is generated from the present state. A phone's state transition probabilities govern the durational characteristics of the phone; if the self-transition probabilities are high, it is more likely that more observations will be generated by that phone which means the overall phone length will be longer [13].

III. BASIC HTS SYSTEM

HMM based speech synthesis system mainly consists of two parts. One is the training part and the other is synthesis part. Figure 2 illustrates an overview of the basic HMM based speech synthesis system. In the training part, spectrum and excitation parameters are extracted from speech database and modeled by phoneme HMMs [3].

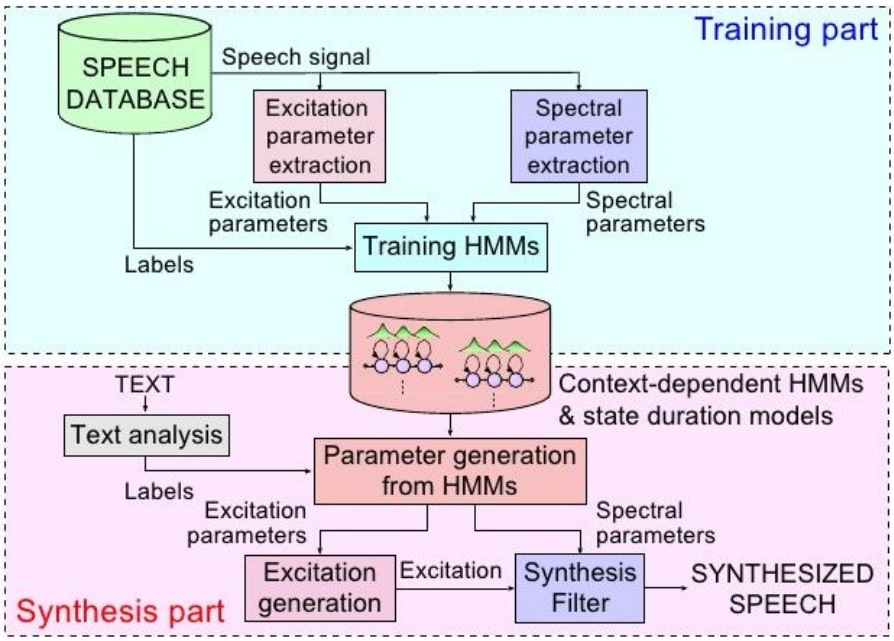

Figure 2: HMM Based Speech Synthesis System

In this system, feature vector consists of spectrum and F0 parts. The spectrum part consists of Mel-cepstral coefficients, their dynamic features i.e., delta and delta-delta. F0 part consists of log F0, its delta and

delta-delta. In the synthesis part, phonemes HMMs are concatenated according to the text to be synthesized [14]. Then spectrum and excitation parameters are generated from the HMM. The excitation generation module and synthesis filter module synthesize speech waveform using the generated excitation and spectrum parameters. The attraction of this approach is that voice characteristics of synthesized speech can easily be changed by transforming HMM parameters [15].

A. Training Part

In the training part, spectrum and excitation parameters are extracted from raw files in the speech database and modeled by context dependent phoneme HMMs. In HTS, output vector of HMM consists of spectrum part and excitation part. The spectrum part consists of Mel-cepstral coefficient vector including the 0th order coefficients, their dynamic features i.e., delta and delta-delta coefficients [16]. On the other hand, the excitation part consists of log fundamental frequency (log F0), its delta and delta-delta coefficients [17]. The extracted parameters model speaker characteristics and speaking styles and they are used to train the context-dependent phoneme HMMs. Spectrum parameters are modeled by multivariate Gaussian distributions, whereas excitation parameters are modeled by multi space probability distribution hidden Markov models (MSD-HMMs) [18]. Mel-cepstral, log fundamental frequency (log F0), and state durations are simultaneously modeled in a unified framework but clustered in isolation using a decision tree based clustering technique called minimum distance length (MDL). The MDL technique ties contextual factors i.e., phoneme identity, stress related and location contexts that are almost similar. This is done because it is both impractical and impossible to prepare a speech database that can model all combinations of contextual factors. A re-estimation of the clustered context dependent phoneme sequence will then be performed using the expectation maximization (EM) algorithm. Clustering is also used to generate excitation and spectrum parameters for newly observed vectors, i.e., observation vectors not included in the training corpus. State durations are modeled by context dependent n-dimensional Gaussian distributions which are then clustered by a decision tree. State densities capture or model the temporal structure of speech. Mel-cepstral coefficients, log F0 and state durations are modeled simultaneously in a unified framework of HMM. Context dependent multi space probability distribution (MSD) hidden semi Markov models (HSMMs) are used to model feature vectors of both continuous and discrete HMMs of the F0 part and continuous HMMs of the spectrum part. HMMs have state duration densities to model the temporal structure of speech. As a result, HTS models not only spectrum parameter but also F0 and duration in a unified framework of HMM. It is noted that it does not require label boundaries for training when an appropriate initial HMM set is available because all parameters of HMMs are determined automatically through the embedded training of HMMs [19]. Training procedure of the context dependent HMMs is almost the same as that in speech recognition systems. The main differences are that not only phonetic contexts but also linguistic and prosodic ones are taken into account and state duration probabilities are explicitly modeled by single Gaussian distributions.[20] i) Spectrum Modeling. To control the synthesis filter by HMM, its system function should be defined by the output vector of HMM, i.e., Mel-cepstral coefficients. Thus a Mel-cepstral analysis technique is used which enables speech to be re-synthesized directly from the Melcepstral coefficients using the MLSA (Mel Log Spectrum Approximation) filter [21].

B. F0 Modeling

HMMs are categorized into discrete and continuous HMMs, which can model sequences of discrete symbols and continuous vectors respectively. Both the conventional discrete and continuous HMMs cannot be applied to observations which consist of continuous values and discrete symbols. Both these conventional discrete and continuous HMMs cannot be applied to pattern modeling since the observation sequence of fundamental frequency (F0) is composed of one-dimensional continuous values and discrete symbol which represents „unvoiced‟ . Therefore the conventional discrete or continuous HMMs cannot be applied to F0 pattern modeling.

Several methods have been investigated for handling the unvoiced region which is given below:

I. Replacing each „unvoiced‟ symbol by a random vector generated from a probability density function (PDF) with a large variance and then modeling the random vectors explicitly in the continuous HMMs.

II. Modeling the „unvoiced‟ symbols explicitly in the continuous HMMs by replacing each “unvoiced” symbol with 0 and adding an extra PDF for the „unvoiced‟ symbol to each mixture. iii. Assuming that F0 values always exist but they cannot be observed in the unvoiced region and applying the EM algorithm.

Although, the last approach is appropriate from the viewpoint of statistical modeling, it intends to estimate F0 values which are not existent contradictorily. From the first two approaches, statistical techniques like context-dependent modeling, speaker or environment adaptation techniques cannot be derived as they are based on assumptions. To model such observation sequences, a new kind of HMM based on multispace probability distribution (MSD-HMM). The MSD-HMM includes discrete HMM and continues mixture HMM as special

cases and further can model the sequence of observation vectors with variable dimensionality including zero-dimensional observations i.e., discrete symbols[22].

C. Duration modeling

State durations of each HMM are modeled by a multivariate Gaussian distribution. The dimensionality of state duration density of an HMM is equal to the number of states in the HMM, and the n-th dimension of state duration densities is corresponding to the n-th state of HMMs.

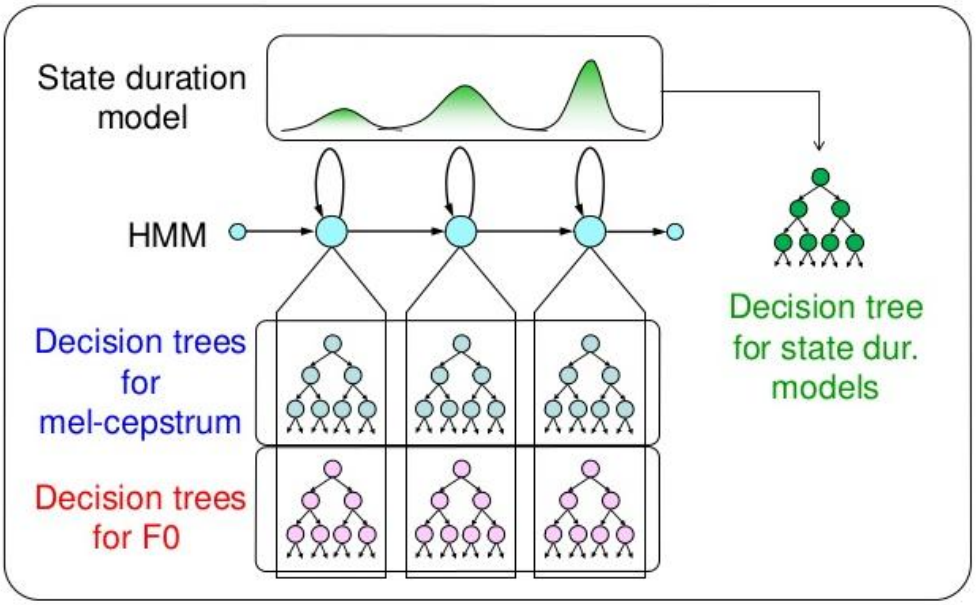

Figure 3: Decision trees for context clustering

D. Decision tree based context clustering

There are many contextual factors like phone identity factors, stress related factors, location factors that affect spectrum, F0 pattern and duration. To capture these effects, context dependent HMMs are used. However, as contextual factors increase, their combinations also increase exponentially. Therefore, model parameters cannot be estimated accurately with limited training data. Furthermore, it is impossible to prepare speech database which includes all combinations of contextual factors. To overcome this problem, a decision tree based context clustering technique is applied to distributions for spectrum, F0 and state duration in the same manner as HMM based speech recognition [23-31]. The decision tree based context clustering algorithms have been extended for MSD-HMMs. Since each of spectrum, F0 and duration has its own influential contextual factors, they are clustered independently as shown in figure 3 State durations of each HMM are modeled by a n-dimensional Gaussian and context dependent n-dimensional Gaussians are clustered by a decision tree. The spectrum part and F0 part of state output vector are modeled by multivariate Gaussian distributions and multi space probability distributions, respectively [23-31].

E. Synthesis Part

In the synthesis part of HTS, first, an arbitrarily given text to be synthesized is converted to a context based label sequence. The text which is given as input to the synthesizer is normalized with the help of text normalization module. This input text is converted in to phonetic representation which is the output of front-end part of the system. HTS is the back-end part of the system. So, HTS tool kit does not include a text analyzer. For text analysis, there is a variety of text analyzers that could be used such as Festival, Flite, etc. So, the HMM models are given to the front-end in order to complete the whole process. Second, according to the label sequence, a sentence HMM is constructed by concatenating context dependent phoneme HMMs. During speech synthesis, an HMM corresponding to the input text is constructed by concatenating the context dependent HMMs. The State durations of constructed HMM are determined by maximizing the output probability of the state durations. According to the obtained state durations, a sequence of Mel-cepstral coefficients and excitation parameter is generated from the sentence HMM by using the speech parameter generation algorithm. The main feature of the system is the use of dynamic feature; by inclusion of dynamic coefficients in the feature vector[23-31]. The speech parameter sequence generated in synthesis is constrained to be realistic, as defined by the statistical parameters of the HMMs. Finally, speech waveform is synthesized directly from the generated Mel-

cepstral coefficients and F0 (excitation parameter) values by using the MLSA (Mel Log Spectrum Approximation filter. In HMM based speech synthesis technique, the speech characteristics can be altered by modifying HMM parameters [18].

IV. CONCLUSION

In this research work, a brief introduction on Hidden Markov Models (HMM) is given. HMM based speech synthesis system is explained in detail and the two parts of the HTS system are also discussed. The training phase is the heart of this method, generating HMMs which are essential for synthesis. Finally, the synthesis phase which outputs a speech waveform is outlined. Next gives detailed information about the speech database chosen for the purpose of training, tools used for implementation of HTS system. Further it explains how the tools are used in the design of the system.

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 "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

[3]. 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

[4]. 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

[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 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

[7]. 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

[8]. 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

[9]. 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.

[10]. 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

[11]. Paul Taylor, "Text to Speech Synthesis", University of Cambridge, pp.442-446

[12]. 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

[13]. 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

[14]. 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

[15]. Sangramsing N. Kayte,Monica Mundada,Dr. Charansing N. Kayte, Dr.Bharti Gawali "Rule-based Prosody Calculation for Marathi Text-to-Speech Synthesis" Sangramsing N. Kayte et al. Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part - 5) November 2015, pp.33-36

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

[17]. 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)

[18]. Keiichi Tokuda, Heiga Zen and Alan W Black, "An HMM Based Speech Synthesis System Applied To English" in Proc. of IEEE Workshop on Speech Synthesis, 2002, pp.227-300.

[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]. H. Zen, K. Oura, T. Nose, J. Yamagishi, S. Sako, T. Toda, T. Masuko, A.W. Black, and K. Tokuda, "Recent development of the HMM Based Speech Synthesis System (HTS)". In Proc. 2009 Asia-Pacific Signal and Information Processing Association (APSIPA), Sapporo, Japan, Oct 2009.

[21]. 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

[22]. Sangramsing N. Kayte,Monica Mundada,Dr. Charansing N. Kayte, Dr.Bharti Gawali "Implementation of Text To Speech for Marathi Language Using Transcriptions Concept" Sangramsing N. Kayte et al. Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part - 5) November 2015, pp.33-36

[23]. 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

[24]. 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

[25]. 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

[26]. 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

[27]. 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

[28]. 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

[29]. Sangramsing N. Kayte, Monica Mundada, Dr. Charansing N. Kayte, Dr.Bharti Gawali " Artificially Generated of Concatenative Syllable based Text to Speech Synthesis System for Marathi" OSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. II (Nov -Dec. 2015), PP 44-49e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[30]. Sangramsing N. Kayte, Monica Mundada, Dr. Charansing N. Kayte, Dr.Bharti Gawali "Automatic Generation of Compound Word Lexicon for Marathi Speech Synthesis" OSR 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

[31]. Sangramsing N. Kayte, Monica Mundada, Dr. Charansing N. Kayte, Dr.Bharti Gawali "Approach of Syllable Based Unit Selection Text-To-Speech Synthesis System for Marathi Using Three Level Fall Back Technique OSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. II (Nov -Dec. 2015), PP 31-35e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197