IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 34-39
e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
www.iosrjournals.org

A Marathi Hidden-Markov Model Based Speech Synthesis System

Sangramsing Kayte[1], Monica Mundada [1, 2] Dr. Charansing Kayte
Department of Computer Science & Information Technology
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad
2Department of Digital and Cyber Forensic, Aurangabad Maharashtra, India

***Abstract**: The research presents the capability of a Hidden Markov Model-based TTS system to produce Marathi speech. In this synthesis method, routes of speech parameters are generated from the trained Hidden Markov Models. A final speech waveform is synthesized from those speech parameters. In our experiments, spectral properties were represented by Mel Cepstrum Coefficients. Both the training and synthesis issues are investigated in this research using marked Marathi speech database. Experimental evaluation depicts that the developed text-to-speech system is accomplished of producing effectively regular speech in terms of intelligibility and pitch for Marathi.*

Keywords: *Marathi Speech Synthesis, Text-To-Speech (TTS), Hidden-Markov-Model (HMM), Marathi HTS.*

I. Introduction

A speech synthesis system is a computer-based system that produce speech automatically, through a grapheme-to-phoneme transcription of the sentences and prosodic features to utter. The synthetic speech is generated with the available phones and prosodic features from training speech database [1][2][21]. The speech units is classified into phonemes, diaphones and syllables. The output of speech synthesis system differs in the size of the stored speech units and output is generated with execution of different methods. A text-to-speech system is composed of two parts: a front-end and a back-end. The front-end has two major tasks. First, it converts raw text containing symbols like numbers and abbreviations into the equivalent words. This process is often called text normalization, preprocessing, or tokenization. Second task is to assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units like phrases, clauses, and sentences. Although text-to-speech systems have improved over the past few years, some challenges still exist. The back end phase produces the synthesis of the particular speech with the use of output provided from the front end. The symbolic representations from first step are converted into sound speech and the pitch contour, phoneme durations and prosody are incorporated into the synthesized speech.

The paper is structured in five sections. The techniques of speech synthesis are described in section 2. Database for synthesis system is explained in section 3. Section 4 explains speech quality measurement. Section 5 is dedicated with experimental analysis followed by conclusion.

The first obligation of a text-to-speech (TTS) system is intelligibility and the second one is the naturalness. Actually the concept of naturalness is not to restitute the reality but to suggest it. Thus, listening to a synthetic voice must allow the listener to attribute this voice to some pseudo-speaker and to perceive some kind of expressivities as well as some indices characterizing the speaking style and the particular situation of elocution [3]. Modern speech synthesizers are able to achieve high intelligibility. However, they still suffer from a rather unnatural speech. Recently, to increase the naturalness, there has been a noticeable shift from di-phone based towards corpus-based unit selection speech synthesis observed [4]. Between these corpus-based unit selection technique is by far the best for producing the natural speech. But it requires a large database often in size of gigabyte. There are many corpus based and di-phone based TTS available for different Indian languages but those are still not reaches the acceptable quality of naturalness. More over the same has not yet been implemented for resource-limited or embedded devices such as mobile phones.

In this view, Hidden Markov Models (HMMs) have proven to be an efficient parametric model of the speech acoustics in the framework of speech synthesis because of its small database size and ability to produce intelligent and natural speech. Although having been originally implemented for Japanese language, the HMM based speech synthesis (HSS) [5] approach has also been applied to other languages, e.g., English [6], German [7], Portuguese [8], Chinese [9], etc. Input contextual labels and questions for context clustering are the only language dependent topics in the HSS scheme. This paper describes first experiments on statistical parametric HMM-based speech synthesis for the Marathi language. For building of our experimental TTS system, HTS toolkit is employed.

DOI: 10.9790/4200-05613439

II. Basic System

The HMM-based speech synthesis technique comprises training and synthesis parts, as depicted in Figure 1.

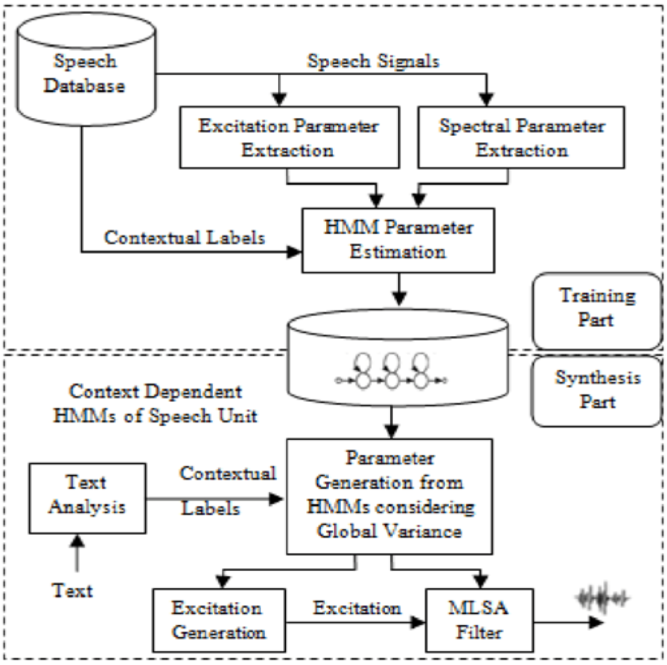


Figure-1: HMM-based speech synthesis system

2.1. Training

In the training part, spectrum and excitation parameters are extracted from the annotated speech database and converted to a sequence of observed feature vectors which is modeled by a corresponding sequence of HMMs. Each HMM corresponds to a left-to-right no-skip model where each output vector is composed of two streams: spectrum part, represented by mel-cepstral coefficients [10] and their related delta and delta-delta coefficients; and the excitation part, represented by Log F0 and their related delta and delta-delta coefficients. Mel-cepstral coefficients are modeled by continuous HMMs and F0s are modeled by multi-space probability distribution HMM (MSD-HMM) [11]. To capture the phonetic and prosody co-articulation phenomena Context-dependent phone models are used. State typing based on decision-tree and minimum description length (MDL) [12] criterion is applied to overcome the problem of data sparseness in training. Stream-dependent models are built to cluster the spectral, prosodic and duration features into separated decision trees.

2.2. Synthesis

In the synthesis phase, input text is converted first into a sequence of contextual labels through the text analysis. Then, according to such label sequence, an HMM sequence is constructed by concatenating context-dependent HMM. After this, state durations for the HMM sequence are determined so that the output probability of the state durations are maximized. Then the mel-cepstral coefficients and F0 routes are generated by using the parameter generation algorithm based on maximum probability criterion with dynamic feature and global variance constraints. Finally, speech waveform is synthesized directly from the generated mel-cepstral coefficients and F0 values by using the MLSA filter [13].

III. Developing Marathi Text-To-Speech

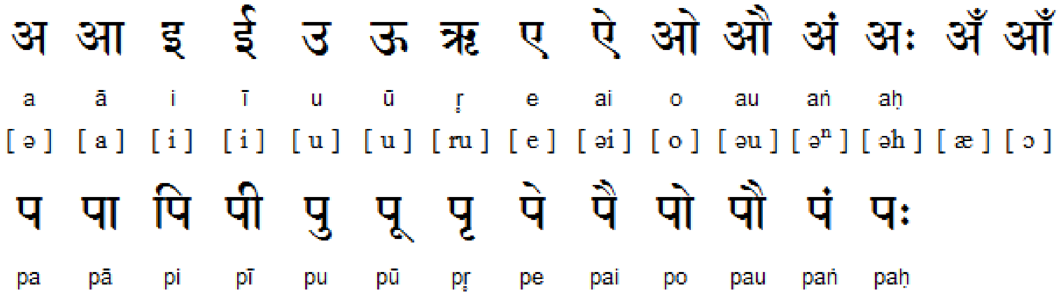
The speech database research lab here for training purpose is originally developed by IIT-Hyderabad [14]. Overall 1000 sentences are used for training which consist of 12 type of sentences i.e. Complex affirmative, Complex negative, Simple affirmative with verb, Simple affirmative without verb, Simple negative, Compound affirmative, Compound Negative, Exclamatory, Imperative, Passive, WH questions, Yes-No questions.

All the sentences were tagged with marking of phoneme, syllable, and word boundaries along with the appropriate Parts of Speech (POS) and phrase/clause markers. During the training prosodic word boundary are used as word boundary instead of syntactic word boundary. Those prosodic word labeling was carried out manually.

3.1. Phonemes

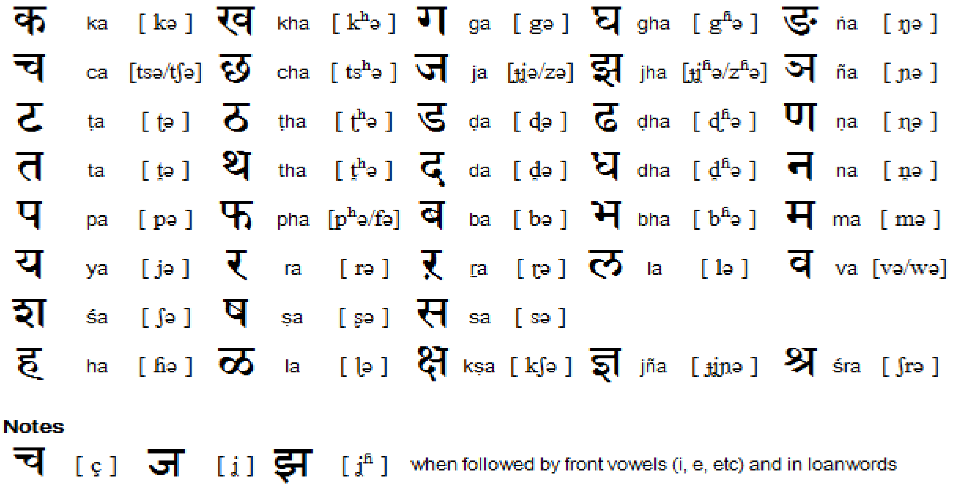
Marathi phoneme inventory consists of 38 consonants including two glides and 15 vowels (including 13 nasal vowels). But the occurrence of Marathi phoneme // is very rear hence this phoneme is not considered during the training and testing. So altogether 66 phonemes, including one silence are used for training as given in Table 1-2 along with their manner and place of articulation. All the diphthongs are marked as vowel-vowel combination [20].

TABLE-1: MARATHI VOWELS CONSONANT INVENTORY



अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	अं	अः	अँ	आँ
a	ā	i	ī	u	ū	ṛ	e	ai	o	au	aṅ	aḥ		
[ə]	[a]	[i]	[i]	[u]	[u]	[ru]	[e]	[əi]	[o]	[əu]	[$ə^n$]	[əh]	[æ]	[ɔ]
प	पा	पि	पी	पु	पू	पृ	पे	पै	पो	पौ	पं	पः		
pa	pā	pi	pī	pu	pū	pṛ	pe	pai	po	pau	paṅ	paḥ		

TABLE-2: MARATHI CONSONANTS INVENTORY



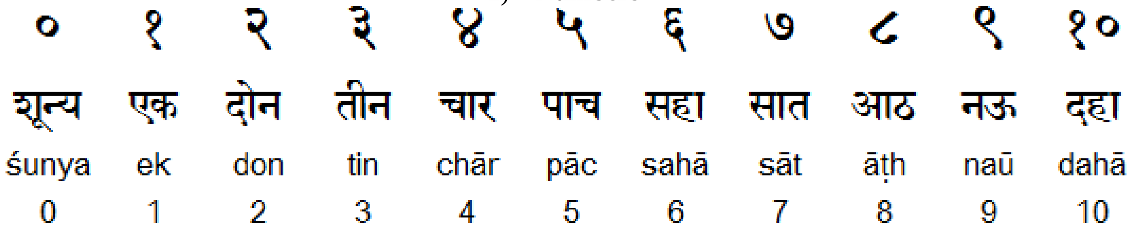
क	ka	[kə]	ख	kha	[$k^hə$]	ग	ga	[gə]	घ	gha	[$g^ɦə$]	ङ	ṅa	[ŋə]
च	ca	[tsə/tʃə]	छ	cha	[$ts^hə$]	ज	ja	[ʤə/zə]	झ	jha	[$ʤ^ɦə$/$z^ɦə$]	ञ	ña	[ɲə]
ट	ṭa	[ʈə]	ठ	ṭha	[$ʈ^hə$]	ड	ḍa	[ɖə]	ढ	ḍha	[$ɖ^ɦə$]	ण	ṇa	[ɳə]
त	ta	[t̪ə]	थ	tha	[$t̪^hə$]	द	da	[d̪ə]	ध	dha	[$d̪^ɦə$]	न	na	[n̪ə]
प	pa	[pə]	फ	pha	[$p^hə$/fə]	ब	ba	[bə]	भ	bha	[$b^ɦə$]	म	ma	[mə]
य	ya	[jə]	र	ra	[rə]	ऱ	ṟa	[ɽə]	ल	la	[lə]	व	va	[və/wə]
श	śa	[ʃə]	ष	ṣa	[ʂə]	स	sa	[sə]						
ह	ha	[ɦə]	ळ	ḷa	[ɭə]	क्ष	kṣa	[kʃə]	ज्ञ	jña	[ʤɲə]	श्र	śra	[ʃrə]

Notes

च [ç] ज [ʝ] झ [$ʝ^ɦ$] when followed by front vowels (i, e, etc) and in loanwords

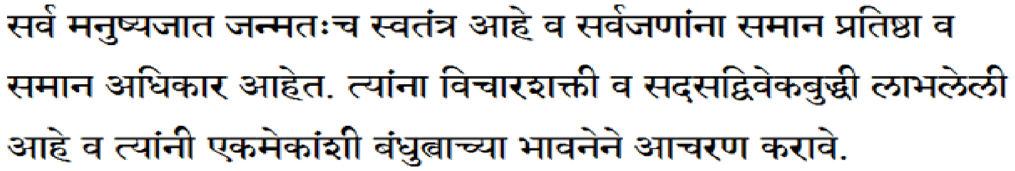
TABLE-3:MARATHI PHONEME NUMBER INVENTORY CONSISTS OF 11 NUMBERS INCLUDING TWO GLIDES AND 3 SAMPLE TEXT IN MARATHI

1) **Numbers**



०	१	२	३	४	५	६	७	८	९	१०
शून्य	एक	दोन	तीन	चार	पाच	सहा	सात	आठ	नऊ	दहा
śunya	ek	don	tin	chār	pāc	sahā	sāt	āṭh	naū	dahā
0	1	2	3	4	5	6	7	8	9	10

TABLE-4: SAMPLE TEXT IN MARATHI



सर्व मनुष्यजात जन्मतःच स्वतंत्र आहे व सर्वजणांना समान प्रतिष्ठा व
समान अधिकार आहेत. त्यांना विचारशक्ती व सदसद्विवेकवुद्धी लाभलेली
आहे व त्यांनी एकमेकांशी वंधुत्वाच्या भावनेने आचरण करावे.

3.2. Tone

In Marathi language tone is not phonemically significant. In a simple declarative sentence with neutral focus, most words and/or phrases in Marathi is said to carry a rising tone with the exception of the last word in the sentence, which carries only a low tone. In that sense Marathi is called bound stress language. In a declarative sentence with neutral focus, the intonation pattern is falling. In sentences involving focused words or phrases, the rising tones last until the right edge of the focused word; all following words carry a low tone. WH sentences follow the same intonation pattern as the sentences involving focused words, but in yes-No sentences

the overall intonation pattern is raising. It observed that most of the bangle prosodic words whether it is spoken in sentences or isolated the F0 contours are rising [16][19].
To match these criteria the TOBI tone markings is done for each of the training sentences.

3.3. Context Based Clustering

In HMM based speech Synthesis the input contextual labels which are used to determine the corresponding HMM in the models set, depends on the language. Thus, contextual information which are fully represented in such contextual labels were necessary to be considered in order to obtain a good reproduction of the prosody.

Table 3 enumerates the main features taken into account and the main language dependent contextual factors are derived from Table 1-2. These questions represent a yes/no decision in a node of the tree. Correct questions will determine clusters to reproduce a fine F0 contour in relation to the original intonation.

IV. Synthesis Of Input Text

In order to generate the context-dependent label format from the given text first POS is marked. As existing automatic POS tagging for Marathi language is not up to the mark so it is done manually. Since the training is done based on the prosodic word, the input text prosodic word labeling is performed based on the rule as describe below [17].

TABLE 3: LIST OF THE CONTEXT FEATURES

Units	Features
Phoneme	-{preceding, current, succeeding} phonemes - position of current phoneme in current syllable
Syllable	-whether or not {preceding, current, succeeding} syllables are stressed -number of phonemes in {preceding, current, succeeding} syllables - position of current syllable in current word -number of stressed syllables in current phrase { before, after} current syllable -number of syllables, counting from previous stressed to current syllable in the utterance -number of syllables, counting from current to next stressed syllable in the utterance
Prosodic Word	-part-of-speech of {preceding, current, succeeding} words -number of syllables in {preceding, current, succeeding} words - position of current word in current phrase -number of content words in current phrase {before, after} current word -number of words counting from previous content word to current word in the utterance -number of words counting from current to next content word in the utterance
Phrase	-number of {syllables, words} in {preceding, current, succeeding} phrases - position of current phrase in current utterance - TOBI endtone of current phrase
Utterance	-number of {syllables, words, phrases} in the utterance

4.1. Prosodic Word Labeling

Rule 1: Hyphenated words and repeated words always form a prosodic word.
Rule 2: Two consecutive proper nouns, within the same prosodic phrase form a prosodic word.
Rule 3: If a common noun (length _3 syllables) is preceded by an adjective (length _ 3 syllables) then they are combined together to form a prosodic word.
Rule 4: A common noun and a verbal noun join together to form a prosodic word.
Rule 5: A postposition and the preceding word together form a prosodic word.
Rule 6: A verb (main or auxiliary) and the following particle together form a prosodic word.
Rule 7: A main verb and the following auxiliary verb (viz., a compound verb) combine together to form a prosodic word. Rule Rule 8: A common noun (or an adjective or a verbal noun) and a verb form a prosodic word. After that the annotated text is converted to phoneme string using Grapheme to Phoneme (G2P) rules described in [18]. A complete block diagram of the above process is shown in Figure 2.

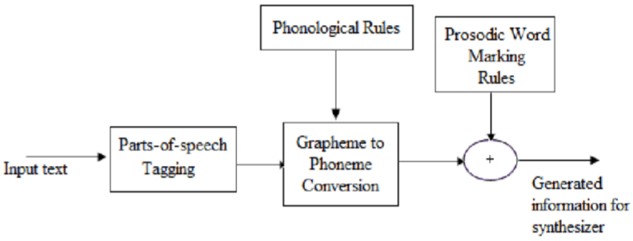
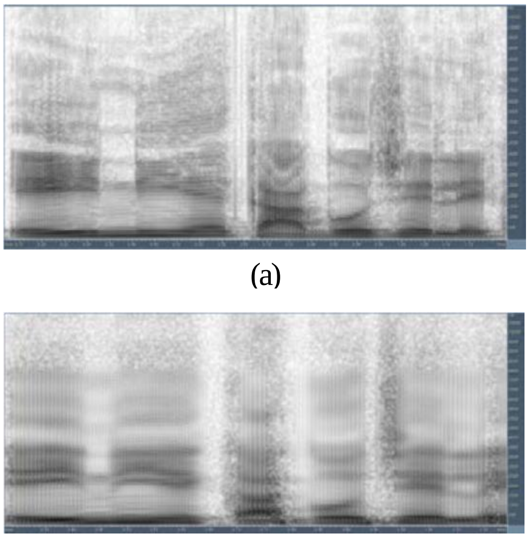

Phonological Rules
Prosodic Word Marking Rules
Input text
Parts-of-speech Tagging
Grapheme to Phoneme Conversion
+
Generated information for synthesizer

Figure 2: Block Diagram for the Linguistic Analysis of Inputted Text

V. Evaluation

Figure- 3 and Figure- 4 show a comparison of spectrogram and F0 patterns between synthesized and original speech signals for a given sentence which is not included in the training database. It can be noticed that the generated spectrogram and F0 contour are quite close to the natural.



(a)

(b)

Figure- 3: Examples of spectrogram extracted from utterance
(a) Natural speech, (b) Synthesized speech

For subjective evaluation of the output speech quality 5 subjects, 3 male (L1, L2, L3) and 2 female (L4, L5), are selected and their age ranging from 24 to 50. All subjects are native speakers of Standard Colloquial Marathi and non-speech expert. 10 original and synthesized sentences are randomly presented for listening and their judgment in 5 point score (1=less natural – 5= most natural).

Table 4 represents the tabulated mean opinion scores for all sentences of each subject for original as well as modified sentences.

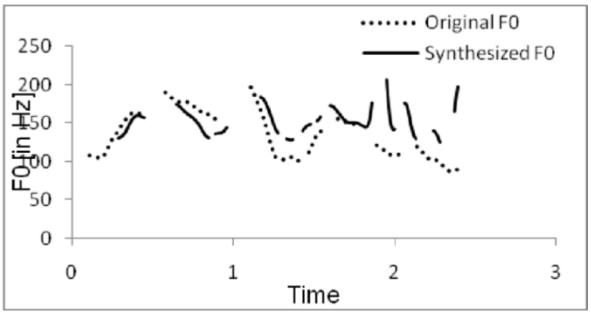


Figure- 4: Examples of F0 contour extracted from utterance
(a) Natural speech, (b) Synthesized speech

TABLE 4 RESULT OF LISTING TEST

Score						
Subject		P1	P2	P3	P4	P5
Original Sentences	Avg	5.9	5.9	5.9	5.9	5.9
	Stdev	1.31	1.05	1.48	1.31	1.70
ESNOLA	Avg	2.5	2.1	2.3	2.6	2.2
	Stdev	1.2	1.7	1.6	1.5	1.7
HTS	Avg	3.7	3.2	3.9	3.8	3.4
	Stdev	1.4	1.7	1.5	1.4	1.6

The output of the Marathi -HTS is also compare with previously developed Epoch Synchronous Non Overlap Add (ESNOLA) [15] based concatenative speech synthesis technique. The total average score for the original sentences is 4.66 and the ESNOLA based synthesis sentence is 2.34 HTS is 3.6.

VI. Conclusion And Future Works

The evaluation results show the efficacy of HMM based Marathi TTS system for generation of highly intelligible speech with naturalness although the training corpus is not made for development of TTS system. In future the above TTS system will be trained by the appropriate training corpus for better quality of output. It was observed during the testing that the intonation of WH and Yes/no sentences was not good in spite of the presence of WH and Yes/no sentences in the training corpus. In future derived F0 contour from the training model can be corrected as per the language input with the help of Fujisaki generation process F0 model.

References

[1]. Mohammed Waseem, C.N Sujatha, “Speech Synthesis System for Indian Accent using Festvox”, International journal of Scientific Engineering and Technology Research, ISSN 2319-8885 Vol.03,Issue.34 November-2014, Pages:6903-6911

[2]. Sangramsing Kayte, Kavita waghmare ,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.

[3]. Campbell, N., Black, A.: Prosody and the Selection of Source Units for Concatenative Synthesis, J. van Santen, R. Sproat, J. Olive and J. Hirschberg (Eds.), in Progress in Speech Synthesis, pp. 279–282, Springer Verlag, 1996.

[4]. Deketelaere S., Deroo O., Dutoit T., “Speech Processing for Communications: What’s New?”(*) MULTITEL ASBL, 1 Copernic Ave, Initialis Scientific Park, B-7000 MONS(**) Faculté Polytechnique de Mons, TCTS Lab, 1 Copernic Ave, Initialis Scientific Park, B-7000 MONS.

[5]. T. Yoshimura, K. Tokuda, T. Masuko, T.Kobayashi and T.Kitamura, "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM Based Speech Synthesis," Proc. of EUROSPEECH, vol.5, pp.2347–2350, 1999.

[6]. K. Tokuda, H. Zen, and A. W. Black, “An HMM-based speech synthesis applied to English,” in IEEE Workshop in Speech Synthesis, 2002.

[7]. S. Krstulovic, A. Hunecke, M. Schroeder, “An HMM-Based Speech Synthesis System applied to German and its Adaptation to a Limited Set of Expressive Football Announcements,” Proc. of Interspeech, 2007.

[8]. Maia, R., Zen, H., Tokuda, K., Kitamura, T., Resende Jr., F.G.,”Towards the development of a Brazilian Portuguese text-tospeech system based on HMM”, Eurospeech, 2003.

[9]. Qian, Y., Soong, F., Chen, Y., Chu, M.: An HMM-based Mandarin Chinese text-to-speech system. In: Q. Huo et al. (eds.)

[10]. ISCSLP 2006, LNAI, vol. 4274, pp. 223-232. Springer, Heidelberg (2006).

[11]. T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in ICASSP, 1992

[12]. Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T., Multi-space Probability Distribution HMM, IEICE Trans. Inf. & Syst., E85-D(3):pp. 455-464, 2002.

[13]. Shinoda, K. and Watanabe, T., Acoustic Modeling Based on the MDL Principle for Speech Recognition, Proc. EuroSpeech 1997 , pp. 99-102.

[14]. K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech synthesis generation algorithms for HMMbased speech synthesis," in ICASSP, 2000.

[15]. Shyamal Das Mandal, Arup Saha, A.K.Datta “Annotated Speech Corpora Development in Indian languages” Vishwa Bharat vol. 16, pp49-64, January 2005.

[16]. Basu, J., Basu, T., Mitra, M., Mandal, S. 2009. “Grapheme to Phoneme (G2P) conversion for Bangla”. Speech Database and Assessments, Oriental COCOSDA International Conference, pp. 66-71.

[17]. Hayes, B. and Lahiri, A., “Bengali Intonational Phonology”, Natural Language and Linguistic Theory, Springer Science, pp 5658 , 1991.

[18]. Shyamal Kumar Das MandaI and Asoke Kumar Datta,. 2007. Epoch synchronous non-overlap-add (ESNOLA) method-based concatenative speech synthesis system for Bangla. 6th ISCA Workshop on Speech Synthesis, Germany, pp. 351-355.

[19]. S. K. Das Mandal, A. H.Warsi, T. Basu, K. Hirose, and H.Fujisaki, Analysis and Synthesis of F0 Contours for Bangla Readout Speech, Proc. of Oriental COCOSDA 2010, Kathmandu, Nepal, 2010.

[20]. 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 (IMPACT FACTOR: 2.080)

[21]. 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) (IMPACT FACTOR: 3.32)

[22]. 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 Impact Factor: 1.492