OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Comparative analysis of Hmm based Marathi TTS using English and Marathi Text font

Monica Mundada*, Dr.Bharti W. Gawali**

**(Department of Computer Science & Information Technology Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India).*

*** (Department of Computer Science & Information Technology Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India.)*

ABSTRACT: *Now-a-days systems are designed for the artificial synthesis of human speech for the given text as input. This helps to develop various assertive devices in fields such as reading for blind people, telecommunication services, language education, and aid to handicapped persons, talking books and toys, call centre automation etc. This article focuses on comparative analysis of Hidden Markov Model (HMM) based speech synthesis for Marathi language using English and Marathi font. The study evaluates the synthetic voice produced using HMM with both Marathi Devanagari text input and English font. Synthetic speech produced by this model sounds more natural and can be easily customized to meet different requirements of different applications and individual users.*

Keywords: *HMM, TTS, NLP, DSP, UNICODE.*

I. INTRODUCTION

Text-to-speech synthesis enables automatic translation of a sequence of type-written words into their spoken form. This paper deals with TTS synthesis of Marathi language. A few attempts have been made in the past to cover different aspects of a possible TTS system for various Indian languages [1]. Marathi is one of the most widely spoken languages of the world it is ranked between 4 and 7 based on the number of speakers, with nearly 100 million native speakers. However, this is one of the most under-resourced languages which lack speech applications. The aim of this project is to develop a freely available Marathi text to speech system. A freely available and open-source TTS system for Marathi language can greatly aid human computer interaction: the possibilities are endless – such a system can help overcome the literacy barrier of the common masses, empower the visually impaired population, increase the possibilities of improved man-machine interaction through on-line newspaper reading from the in telnet and enhancing other information systems [2]. A touch screen based kiosk that integrates a Marathi TTS has the potential to empower the 49% of the population who are illiterate. A screen reader that integrates a Marathi TTS will do the same for the estimated 100 thousand visually impaired citizens of Maharashtra.

A Text to Speech is a computer based system capable of converting computer readable text into speech. There are two main components such as Natural Language Processing and Digital Signal Processing [3]. The NLP component includes pre-processing, sentence splitting, tokenization, text analysis, homograph resolution, parsing, pronunciation, stress, syllabification and prosody prediction. Working with pronunciation, stress, syllabification and prosody prediction sometime is termed as linguistic analysis. Whereas, the DSP component includes segment list generation, speech decoding, prosody matching, segment concatenation and signal synthesis. In text analysis part, different semiotic classes were identified, and then using a parser each token is assigned to a specific semiotic class. After that, verbalization is performed on non-natural language token. Homograph resolution is the process of identifying the correct underlying word for ambiguous token. The process of generating pronunciation from orthographic representation can be done by pronunciation lexicon and grapheme-to-phoneme (G2P) algorithm. Prosody prediction is the process of identifying the phrase break, prominence and intonation tune. In English ASCII characters are used where as In Marathi Unicode characters are used. ASCII takes 8-bits for each character. Unicode takes 16-bits for each character. Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language.

The paper is structured in five sections. The features of Marathi language along with UNICODE values and description are explained in section 2. The training and synthesis of HMM are explained in section 3. Section 4 explains experimental analysis and observations. Section 5 is dedicated with conclusion and references.

II. FEATURES OF MARATHI LANGUAGE

Marathi language uses Devanagari, a character based script. A character represents one vowel and zero or more consonants. Consonant clusters are represented by combination of ligatures; so, there are hundreds of characters [4]. Many characters share glyphs. To take advantage of such shared glyphs, some true type font designers have used codes corresponding to glyphs rather than phonemes. This makes the task of rendering Marathi text in Devanagari script easier. However, this causes problems for us, since we are interested in statistical analysis of phonemes. The grapheme-to-phoneme mapping of text in true type fonts is not straightforward, especially because of non-linear nature of Devanagari script. For example, the word "ki" is a sequence two phonemes: /k/ and /i/. However, the corresponding character in Devanagari script has the glyph corresponding to phoneme. There are 50 Marathi phonemes. About 2/3rd of the sentence sets contain at least one token of each and every Marathi phoneme. Even the least phonetically rich sentence set contains at least 45 phonemes out of 50. It would be interesting to discuss about the phonemes that did not occur in some sentence set. Out of 10 rare phonemes, 6 are asphirated stop consonants; 2 are retroflex sounds; the rest are velar and palatal nasals. Such a tilt of the distribution of rare phonemes towards asphirated stops is to be expected since asphirated phonemes are known to occur rarely in natural text. There are 10 asphirated stop consonants in Marathi. So, it is difficult to have all of the 10 rare phonemes in addition to other phonemes in just 10 sentences. The rarest phoneme is the velar nasal (/ng/); it does not occur in 64 out of 1000 sentence sets. Table 2.1 shows the Marathi alphabet with their UNICODE values and description.

Table 2.1 Marathi alphabets and their respective UNICODE values and meanings

Unicode Value	Devanagari Character	Description
0901	ँ	CHANDRABINDU_SIGN
0902	ं	Anuswara
0903	ः	Wisarga
0904	ऄ	SHORT_A
0905	अ	A
0906	आ	AA
0907	इ	I
0908	ई	II
0909	उ	U
090A	ऊ	UU
090B	ऋ	RRU
090C	ऌ	LRU
090D	ऍ	ARDHACHANDRA_E
0972	ॲ	ARDHACHANDRA_A
090E	ऎ	SHORT_E
090F	ए	E
0910	ऐ	EI
0911	ऑ	ARDANCHANDRA_O
0912	ऒ	SHORT_O
0913	ओ	O
0914	औ	OU
0915	क	K
0916	ख	KH
0917	ग	G
0918	घ	GH
0919	ङ	NGA
091A	च	CH
091B	छ	CHHA
091C	ज	J
091D	झ	Z
091E	ञ	NYA

091F	ट	T
0920	ठ	TH
0921	ड	D
0922	ढ	DH
0923	ण	N
0924	त	T
0925	थ	Th
0926	द	D
0927	ध	Dh
0928	न	N
0929	ऩ	DRAVID_n
092A	प	P
092B	फ	F
092C	ब	B
092D	भ	BH
092E	म	M
092F	य	Y
0930	र	R
0931	ऱ	DRAVID_R
0932	ल	L
0933	ळ	L
0934	ऴ	DRAVID_LLLA
0935	व	W
0936	श	SH
0937	ष	SHH
0938	स	S
0939	ह	H
093C	़	NUKTA
093D	ऽ	AWAGRAHA
093E	ा	KANA
093F	ि	RHASWA_IKAR
0940	ी	DEERGHA_IKAR
0941	ु	RHASWA_UKAR
0942	ू	DEERGHA_UKAR
0943	ृ	RRUKAR
0944	ॄ	LRUKAR
0945	ॅ	ARDHACHANDRA_E_SIGN
0946	ॆ	SHORT_MATRA
0947	े	MATRA
0948	ै	DOUBLE_MATRA
0949	ॉ	ARDHACHANDRA_O_SIGN
094A	ॊ	SHORT_KANA-MATRA
094B	ो	KANA-MATRA
094C	ौ	KANA-DOUBLE_MATRA
094D	्	WIRAMA
0950	ॐ	OM
0951	॑	UDATTA
0952	॒	ANUDATTA
0953	॓	GRAVE_ACCENT
0954	॔	ACCUTE_ACCENT
0958	क़	K_WITH_NUKTA
0959	ख़	KH_WITH_NUKTA
095A	ग़	G_WITH_NUKTA

095B	ज़	J_WITH_NUKTA
095C	ड़	D_WITH_NUKTA
095D	ढ़	DH_WITH_NUKTA
095E	फ़	F_WITH_NUKTA
095F	य़	Y_WITH_NUKTA
0960	ॠ	RRRU
0961	ॡ	LRRU
0962	◌ॢ	RRRU_Vowel_Sign
0963	◌ॣ	LRRU_Vowel_Sign
0964	।	DANDA
0965	॥	DOUBLE_DANDA
0966	०	0
0967	१	1
0968	२	2
0969	३	3
096A	४	4
096B	५	5
096C	६	6
096D	७	7
096E	८	8
096F	९	9
0970	॰	ABBREV_SIGN

III. HIDDEN MARKOV MODEL

The HMM is doubly stochastic process with an underlying stochastic process that is not observable, but can only be observed through another set of stochastic processes that produce sequence of observed symbols Baker at Carnegie Mellon University and Jelinek at IBM provided the first HMM implementations to speech processing applications in the 1970s [5]. HMM has been broadly accepted in today's modern state: its capability to model the non-linear dependencies of each speech unit on the adjacent units and a powerful set of analytical approaches provided for estimating model parameters [6][7].The Hidden Markov Model (HMMs) is widely-used statistical models to characterize the sequence of speech spectra and have successfully been applied to speech recognition and speech synthesis systems. This system simultaneously models, spectrum, excitation, and continuance of speech using content dependent HMMs and generates speech waveforms. HMM creates stochastic models from known utterances and compares the probability that the unknown utterance can be engendered by each model. It also includes various methods for text to voice communication such as Dynamic Features (Delta and Delta-Delta parameters of Speech). This technique uses very less memory, but consumes large CPU resources. This approach gives good prosody features with natural sounding language.

3.1training

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 [8] 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).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) [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.

3.2. 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 [9].

IV. EXPERIMENTAL ANALYSIS

The speech database consist of 500 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. The recording is performed in recording studio with noise free environment. 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. The Hidden Markov Model algorithm is performed on Marathi Text font with their Unicode values. The naturalness and intelligibility is observed by calculating MOS (Mean Opinion Score) for subjective quality measurement. It is calculated for the synthesized speech using the HMM approach. It was counseled to the listeners that they have to score between 01 to 05 (Excellent – 05 Very good – 04 Good – 03 Satisfactory – 02 Not understandable-01) for understandable. The mean of the scores given by each individual subject for ten sentences of the Marathi input Text and for English is given in table 4.1.

Table 4.1 Comparative mean analysis report of Marathi text and English Text as Input using HMM

Subject	Sentence	Marathi Text as Input	English Text as Input
Sub1	1	5	5
Sub2	2	5	4
Sub3	3	4	5
Sub4	4	5	3
Sub5	5	5	4
Sub6	6	4	5
Sub7	7	5	5
Sub8	8	5	5
Sub9	9	3	4
Sub10	10	5	5

The observation shows that during the Marathi text as input the English loan words like cricket, country names like New Zealand are not synthesized up to the mark of naturalness. The mean scores of the sentences are rated by the individual on the criteria of naturalness of synthesized speech and intelligibility involved in the output speech.

V. CONCLUSION

In this paper we have studied the comparative result for Marathi language, both Marathi Unicode and English font as text input. The study show that during the Marathi Text as input the English loan words are not synthesized up to the naturalness. The work will help to build various assertive devices, screen readers with both Marathi and English Text as input. This study even focuses to build synthetic Marathi voice application for users who don't know how to write Marathi and English vice-versa. The evaluation results show the efficiency of HMM based Marathi TTS system for generation of highly intelligible speech with naturalness for development of TTS system.

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

[3]. 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.

[4]. M L Dhore, S K Dixit, R M Dhore, Hindi and Marathi to English NE Transliteration Tool using Phonology and Stress Analysis. pages 111–118,COLING 2012, Mumbai, December 2012.

[5]. Rabiner, L. and Juang, B.H. (1986), "An Introduction to Hidden Markov Models", IEEE ASSP Magazine, Vol. 3, No.1, Part 1 pp. 4-16.
[6]. Picone, J. (1990), "Continues Speech Recognition using Hidden Markov Models", IEEE ASSP Magazine, Vol. 7, Issue 3, pp. 26-41.
[7]. Flahert, M.J. and Sidney, T. (1994), "Real Time implementation of HMM speech recognition for telecommunication applications", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), Vol. 6, pp. 145-148.
[8]. K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis applied to English," in IEEE Workshop in Speech Synthesis, 2002.
[9]. 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.