European Journal of Computer Science and Information Technology
Vol.3, No.5, pp.30-41, November 2015
Published by European Centre for Research Training and Development UK (www.eajournals.org)

A TEXT-TO-SPEECH SYNTHESIS FOR MARATHI LANGUAGE USING FESTIVAL & FESTVOX

Sangramsing Kayte[1], Monica Mundada[1] and Dr.Charansing Kayte[2]

[1]Research Scholar, Deprtment of Computer Science & IT
[2]Assistant Professor, Department of Digital and Cyber Forensic, Maharashtra
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad.

ABSTRACT: *This research paper describes the Impalement of the first, usable, Marathi Text to Speech system for Maharashtra Marathi using the open source Festival TTS engine. Besides that, this research paper also discusses a few practical applications that use this system. This system is developed using di-phone concatenation approach in its waveform generation phase. Construction of a di-phone database and implementation of the natural language processing modules are described. Natural language processing modules include text processing, tokenizing and grapheme to phoneme (G2P) conversion that were written in Festival's format. Finally, a test was conducted to evaluate the intelligibility of the synthesized speech.*

KEYWORDS: Marathi Speech Synthesis, Text-To-Speech (TTS), Hidden-Markov-Model (HMM), Marathi HTS TTS, speech synthesis, di-phone, Unit Selection.

INTRODUCTION

Marathi is one of the most widely spoken languages of the world (it is ranked between four and seven 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 [1].

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. 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. Pre-processing is the process of identifying the text genre, character encoding issues and multilingual issues. Sentence splitting is the process of segmenting the document text into a list of sentences. Segmenting each sentence into a list of possible tokens can be done by tokenization. In text analysis part, different semiotic classes were identified, and then using a

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

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. There has not been much work done on prosody in this paper. DSP component or waveform generation is the final stage of a TTS. This involves the production of acoustic signals using a particular synthesis approaches such as formant synthesis, articulatory synthesis and concatenation based synthesis. The attempt that has been made here is the second generation di-phone concatenation based synthesis, using widely usable Festival framework [1].

LITERATURE SURVEY

Fundamentals of speech and speech signal processing are discussed in "Text-to-Speech Synthesis" by Paul Taylor. Lungs, larynx and vocal tract are the speech organs that are responsible for speech production. Lungs are the source of an airflow that passes through the larynx and vocal tract, before leaving the mouth as pressure variations constituting the speech signal. The source of most speech occurs in the larynx where vocal folds can obstruct airflow from the lungs. Humans produce a large number of sounds within the constraints of the vocal tract [3] [5].

Each language has a small set of linguistic units called phonemes to describe its sounds. A phoneme is the smallest meaningful unit in the phonology of a language. The physical sound produced when a phoneme is articulated is called a phone. The importance of vocal organs in speech production is clearly discussed. In addition, this book explains about the concept of how the vocal organs respond when vowels and consonants are uttered.

Basics of speech synthesis and speech synthesis methods are discussed in "An Introduction to Text-to-Speech Synthesis" by Thierry Dutoit [6]. Text to speech system organization, functions of each module and conversion of text which is given as input in to speech is clearly explained in this book.

Different speech synthesis methods, which are used for development of synthesis system, are explained in "Review of methods of Speech Synthesis" [4].

The process of text normalization is understood from "Normalization of non-standard words" by Christopher Richards. It research, conversion of non-standard words (NSWs) in to standard words is explained clearly with examples. Non-standard words are the words which are not found in the dictionary. NSWs are tokens that need to be expanded into an appropriate orthographic form before the text-to-phoneme module [7].

The concept of prosody is explained briefly in "Text to speech synthesis with prosody feature" by M. B. Chandak. This book explains how the prosody is predicted from the text which is given as input. In linguistics, prosody includes the intonation, rhythm and lexical stress in speech. The prosodic features of a unit of speech can be grouped into syllable, word, phrase or clause level features. These features are manifested as duration, F0 and intensity [8].

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

Prosodic units do not need to necessarily correspond to grammatical units. The perceived quality of synthetic speech is largely determined by the naturalness of the prosody generated during synthesis. The correct prosody also has an important role in the intelligibility of synthetic speech. Prosody also conveys paralinguistic information to the user such as joy or anger. In speech synthesis system, intonation and other prosodic aspects must be generated from the plain textual input.

Paper by Ramani Boothalingam (2013) presented the comparison of the performance of Unit Selection based Synthesis (USS) and HMM based speech synthesizer. Difference between the two major speech synthesis techniques namely unit selection based synthesis and HMM based speech synthesis is explained clearly in this paper [9].

Unit selection systems usually select from a finite set of units in the speech database and try to find the best path through the given set of units. When there are no examples of units that would be relatively close to the target units, the situation can be viewed either as lacking in the database coverage or that desired sentence to be synthesized is not in the domain of the TTS system. To achieve good quality synthesis, the speech unit database should have good unit coverage.

To obtain various voice characteristics in TTS systems based on the selection and concatenation of acoustical units, a large amount of speech data is needed. It is very difficult to collect and segment large amount of speech data for different languages. Storing big database in devices having only a small amount of memory is not possible. So, in order to construct a speech synthesis system that can generate various voice characteristics, without big speech database, HMM based speech synthesis was proposed.

In HMM based speech synthesis system, the parameters of speech are modeled simultaneously by HMMs. During the actual synthesis, speech waveforms are generated from HMMs themselves based on maximum likelihood criteria.

The advantage of using HMM based speech synthesis technique for developing TTS system is discussed briefly in “An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005” by Heiga Zen and Tomoki Toda. The concept of accuracy measurement is outlined. The parameters on which the accuracy of a TTS system depends are discussed [10].

The quality of a TTS system is often determined by using four measures namely intelligibility, naturalness, accuracy and listening ability. These four measures are not independent from one another. For example, serious errors in accuracy will lead to less intelligibility speech, and this will be perceived as less natural and the listening ability of the synthesized speech becomes worse.

DEVELOPMENT

Marathi written system

Marathi is an Indo-Aryan language spoken by about 71 million people mainly in the Indian state of Maharashtra and neighboring states. Marathi is also spoken in Israel and Mauritius. Marathi is thought to be a descendent of Maharashtra, one of the Prakrit languages which

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

developed from Sanskrit. Marathi first appeared in writing during the 11th century in the form of inscriptions on stones and copper plates. From the 13th century until the mid-20th century, it was written with the Modi alphabet. Since 1950 it has been written with the Devanāgarī alphabet [15].

Devanāgarī alphabet for Marathi Vowels and vowel diacritics [15]

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 aṁ aḥ

[ə] [a] [i] [i] [u] [u] [ru] [e] [əi] [o] [əu] [ãⁿ] [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 paṁ paḥ

Consonants[15]

क	ka	[kə]	ख	kha	[k^hə]	ग	ga	[gə]	घ	gha	[$g^ɦ$ə]	ङ	ṅa [ŋə]
च	ca	[tsə/tʃə]	छ	cha	[ts^hə]	ज	ja	[ʤə/zə]	झ	jha	[$ʤ^ɦ$ə/$z^ɦ$ə]	ञ	ña [ɲə]
ट	ṭa	[ʈə]	ठ	ṭha	[$ʈ^h$ə]	ड	ḍa	[ɖə]	ढ	ḍha	[$ɖ^ɦ$ə]	ण	ṇa [ɳə]
त	ta	[t̪ə]	थ	tha	[$t̪^h$ə]	द	da	[d̪ə]	ध	dha	[$d̪^ɦ$ə]	न	na [n̪ə]
प	pa	[pə]	फ	pha	[p^hə/fə]	ब	ba	[bə]	भ	bha	[$b^ɦ$ə]	म	ma [mə]
य	ya	[jə]	र	ra	[rə]	ऱ	ṛa	[ɾə]	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	[ʤ̃ɲə]	श्र	śra [ʃrə]

Sample text in Marathi

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

MARATHI PHONOLOGY

The phoneme inventory of the Marathi language is similar to that of many other Indo-Aryan languages. An IPA chart of all contrastive sounds in Marathi is provided below.

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

Vowels in native words are:

Vowels			
	Front	Central	Back
High	i		u
Mid	e	ə	o
Low		a	

There are no nasal vowels.Like other alpha syllabaries, Devanagari writes out syllables by adding vowel diacritics to consonant bases. The table below includes all the vowel symbols used in Marathi, along with a transliteration of each sound into Latin script and IPA.

Table 1: Vowel phoneme inventory

Devanagari	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	अं	अः
Transliterated	a	ā	i	ī	u	ū	ṛ	e	ai	o	au	aṃ	aḥ
IPA	/ə/	/a/	/i/	/u/	/ru/	/e/	/əi/	/o/	/əu/	/əm/	/əhə		

Marathi furthermore contrasts /əi, əu/ with /ai, au/.[16]

There are two more vowels in Marathi to denote the pronunciations of English words such as of /æ/ in act and /ɔ/ in all. These are written as ⟨अॅ⟩ and ⟨ऑ⟩. Marathi retains several features of Sanskrit that have been lost in north-Indian Sanskrit-based languages such as Hindi and Bengali, especially in terms of pronunciation of vowels and consonants. For instance, Marathi retains the original Sanskrit pronunciation of ⟨अं⟩ [ə^{n}], ⟨ऐ⟩ [əi], and ⟨औ⟩ [əu]. However, as was done in Gujarati, Marathi speakers tend to pronounce ऋ ṛ somewhat similar to [ru], unlike most other Indic languages which changed it to [ri] (e.g. the original Sanskrit pronunciation of the language's name was saṃskṛtam, while in day-to-day Marathi and Gujarati it is saṃskrut. In other Indic languages, it is closer to

sanskrit). Spoken Marathi allows for conservative stress patterns in words like राम (rama) with an emphasis on the ending vowel sound, a feature that has been lost in Hindi [19].

Natural language processing in Festival

Festvox [2] provides different natural language processing modules for building a new voice. These modules can be generated automatically which appears as a form of scheme files. The scheme files need to be customized for a new language. The language specific scripts (phone, lexicon and tokenization) and speaker specific scripts (duration and intonation) can be externally configured and implemented without recompiling the system [2]. Since the templates are scheme files, which is typically an interpreted language, so recompilation is not required. The following NLP related tasks are involved when building a new voice in Festvox:

- Defining the phone-set
- Tokenization and text normalization
- Pronunciation: Lexicon and grapheme to phoneme conversion.
- Implementation of syllabification and stress

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

- Prosody: Phrase breaking, accent prediction, assignment of duration to phones and generation of f0 contour.

Whitespace and the punctuation marks, which is used in our implementation to tokenize Marathi text. After tokenization, text normalization is performed. In text normalization, the first task is to identify the semiotic classes. The following section discusses the semiotic class [12] (as opposed to say NSW) identification, tokenization and standard word generation and disambiguation rule. Moreover, this work has been done separately before implementing into Festival.

The phoneme inventory of Marathi is similar to that of many other Indo-Aryan languages. An IPA chart of all contrastive sounds in Marathi is provided below

Table 2: Consonant phoneme inventory

Consonants							
	Labial	**Dental**	**Alveolar**	**Retroflex**	**(Alveolo-) palatal**	**Velar**	**Glottal**
Nasal	**plain**	m	n̪		ɳ	(ɲ)	(ŋ)
	murmured	mʱ	n̪ʱ		ɳʱ		
Stop	**voiceless**	p	t̪	t͡s	ʈ	t͡ɕ~t͡ʃ	k
	aspirated	pʰ	t̪ʰ		ʈʰ	t͡ɕʰ~t͡ʃʰ	kʰ
	voiced	b	d̪	d͡z~z	ɖ	d͡ʑ~d͡ʒ	g
	murmured	bʱ	d̪ʱ	d͡zʱ~zʱ	ɖʱ	d͡ʑʱ~d͡ʒʱ	gʱ
Fricative			s	ʂ	ɕ~ʃ		h~ɦ
Approximant	**plain**	ʋ		l	ɭ	j	
	murmured	ʋʱ		lʱ		(jʱ)	
Flap/Trill	**plain**			ɾ	ɻ		
	murmured			ɾʱ			

Older aspirated *tsʰ, dzʱ have lost their onset, with *tsʰ merging with /s/ and *dzʱ being typically realized as an aspirated fricative, [zʱ]. This /ts, dz, zʱ/ series is not distinguished in writing from /tʃ, tʃʰ, dʒ, dʒʱ/.

Pronunciation

This system takes the word based on orthographic linguistic representation and generates a phonemic or phonetic description of what is to be spoken by the subsequent phases of TTS. In generating this representation we used a lexicon of known words and a grapheme-to-phoneme (G2P) algorithm to handle proper names and unknown words.

We developed a system lexicon where the entries contain orthography and pronunciation in IPA. Due to the lack of a digitized offline lexicon for Marathi we had to develop it manually by linguistic experts. To the best of our knowledge this is the first digitized IPA incorporated and syllabified lexicon. The lexicon contains 93K entries where 80K entries entered by hand and the rest of them were automatically generated by G2P system [13]. The performance of this G2P system is 91%. Therefore, the automatically generated entries had to be checked manually to maintain the quality of the lexicon by expert linguists. The system is now available

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

in online for public access [CRBLP 2010]. Another case needs to be handled in order to implement the lexicon into Festival. The Unicode encoded phonetic representation needs to be converted into ASCII to incorporate into festival.

We have implemented the G2P algorithm[13] to handle unknown words and proper name. In Festival, the UTF-8 textual input was converted into ASCII based phonetic representation in a Festival's context sensitive rule [2]. The rules were re-written in UTF-8 multi-byte format following the work done for Telugu [14] and Sinhala. The method was proven to work well with promising speed. The rules proposed in [13] were expanded up to 3880 rules when re-written in Festival context sensitive format.

Another attempt has been made to reduce the size of the lexicon for TTS. The lossless compression [12] technique was applied to reduce the size of the lexicon. Lossless compression technique is a technique where the output is exactly the same as when the full lexicon is used. It is just the generalities of the lexicon that has been exactly captured in a set of rules. This technique reduces the size of our lexicon to ~50K entries from 93K.

Syllabification and stress

Festival's default syllabification algorithm based on sonority sequencing principle [2] is used to syllabify the Marathi words. Besides the default syllabification algorithm, our lexicon has also been syllabified along with pronunciation.

Little work has been done on Marathi stress. Identifying the stress pattern for Marathi is beyond the scope of this paper. Considering Marathi as a stress less language we have used Festival's default stress algorithm. In our implementation of lexicon we have not incorporated the stress marker.

Prosody Implementation

Prosody is one of the important factors contributing to natural sounding speech. This includes phrasing, accent/boundary prediction, duration assignment to phones and f0 generation. The presence of phrase breaks in the proper positions of an utterance affects the meaning, naturalness and intelligibility of the speech. Festival supports two methods for predicting phrase breaks. The first one is to define a Classification and Regression Tree (CART). The second and more elaborate method of phrase break prediction is to implement a probabilistic model using probabilities of a break after a word, based on the part of speech of the neighboring words and the previous word [2]. However, due to the lack of a POS tagger for Marathi, we have not able to construct a probabilistic model yet. Therefore, we decided to use a simple CART based phrase breaking algorithm described in [2]. The algorithm is based on the assumption that phrase boundaries are more likely between content words and function words. A rule is defined to predict a break if the current word is a content word and the next is seemingly a function word and the current word is more than 5 words from a punctuation symbol. Since function words are limited in a language so we specified them as function words and considered rest of them as content words. The function words that we used here to implement the phrase break model is shown in Table 4. These function words need to be converted into ASCII form to incorporate into festival phrase breaking algorithm.

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

Table 4: Function words [16][17]

अशा प्रकारे अमेरिकेचा या युद्धात प्रवेश झाला
रत्नागिरी जिल्ह्याच्या माहितीसाठी येथे टिचकी द्या
पुणे शहरातील एक मध्यवर्ती ठिकाण
जलंधर शहरातील क्रिकेटचे मैदान आहे
आपण चर्चा करताना कृपया चार वापरून सही करावी
केवळ प्रबंधक ही पाने बदलू शकतात काय "

To predict accent and boundary tone, Festival uses simple rules to produce sophisticated system. To make a more sophisticated system such as a statistical model one needs to have an appropriate set of data. Due to the lack of the availability of this data we used a simple accent prediction approach [2] which proved surprisingly well for English. This approach assigns an accent on lexically stressed syllable in all content words.

Festival uses different approach for F0 generation such as F0 by rule, CART tree and tilt modeling. In our implementation we used rule based approach. An attempt has been made to make a CART tree based model from the data; however, surprisingly that has not been work well.

Several duration models support by Festival such as fixed models, simple rules models, complex rules models and trained models.

Development of di-phone database

Developing a speech database is always time consuming and laborious. The basic idea of building a di-phone database is to explicitly list all phone-phone combination of a language. It is mentioned in section 3.2 that Marathi language has 43 consonants and 28 vowels phonemes. In general, the number of di-phone in a language is the square of the number of phones. Since Marathi language consists of 65 phones, so the number of di-phones are (65X65) 4225. In addition, silence to phones are (1X65) 65, phones to silence are (65X1) 65 and a silence. So the total number of di-phones is 4336. In the first step, a list has been made to maintain all the possible vowel consonant combination with the following pattern: VC, CV, VV, CC, SIL_V, SIL_C, V_SIL, C_SIL and SIL. Here SIL is silence, V is vowel and C is consonant. Silence is considered as a phoneme, usually taken at the beginning and ending of the phonemes to match the silences occurring before, between and after the words. They are therefore an important unit within the di-phone inventory. These di-phones were embedded with carrier sentences using an external program. The di-phone is inserted in the middle of the word of a sentence, minimizing the articulatory effects at the start and end of the word. Also, the use of nonsense words helped the speaker to maintain a neutral prosodic context. Though there have been various techniques to embed di-phone with carrier sentences, here nonsense words were used to form carrier sentences [2]. In this list, there could be redundant di-phones those need to be marked and omitted. The study of phonotactics says that all phone-phone pair cannot be exist in a language. Due to the lack of existing work and linguistic experts we were not able to work on this phenomenon. Therefore, the whole di-phone list was selected for recording.

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

Since speaker choice is perhaps one of the most vital areas for recording so a careful measure had taken. Two potential speakers was chosen and their recording were played to a listening group and asked them which they prefer. According to the measurement of the listening group a male speaker was chosen who is a professional speaker and aged 28.

As far as recording conditions is concerned, we tried to maintain as high quality as possible. The speech data was digitized at a sample rate 44.1 kHz, sample width 24-bit resolution and stored as wave format. After each recording, the moderator checked for any misleading pronunciation during the recording, and if so, the affected utterances were re-recorded.

There were a few challenges in the recording. First, speaker was asked to keep the speaking style consistent. Second, speaker was supervised to keep the same tone in the recording.

The most laborious and painstaking task is to clean the recording and then hand-labeled the di-phone using the speech analysis software tool

'Praat' During labeling, at first we labeled phone boundary, then automatically marked the diphone boundary using Praat script. Another important factor is that, every boundary should be placed in zero crossing. Failing to do so produces audible distortions, this in turns generates clicks. Afterwards, a script was written to transform Praat textgrid files into diphone index file (.est) [2] as required by Festival.

Festival, in its publicly distributed form only supports residual excited Linear Predictive Coding (LPC). This method requires pitch marks, LPC parameters and LPC residual values for each diphone in the diphone database. The script make_pm_wave provided by speech tools [A.W. Black et al. 2003] was used to extract pitch marks from the wave files. Then, the make_lpc command was invoked in order to compute LPC coefficients and residuals from the wave files [A.W. Black et al. 2003]. To maintain an equal power we used proprietary software tool to normalize it in terms of power so that all diphones had an approximately equivalent power. After that the diphone database was grouped in order to make it accessible by Festival's UniSyn synthesizer module, and to make it ready for distribution.

INTEGRATION WITH APPLICATIONS

The Marathi Text to Speech runs on Linux, Windows and Mac OSX. There is also a webenabled front-end for the TTS, making this tool available at any time and from anywhere.

Since Festival is incapable of reading UTF-8 text files with byte-order marker (BOM) so manual BOM removal patch was used which was written. This patch was incorporated with Festival text processing module.

To develop windows version we had motivated by the work carried out in the Welsh and Irish Speech Processing Resources (WISPR) project. Following the work of WISPR, we implemented TTS using Microsoft Speech Application Programming Interface (MS-SAPI) which provides the standard speech synthesis and speech recognition interface within Windows applications]. Consequently, the MS-SAPI compliant Marathi voice is accessible via any speech enabled Windows application. The system has been tested with NVDA and Dolphin screen reader. Moreover, it is also tested with WordTalk , a free text-to-speech plug-in for

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

Microsoft Word which runs as a macro. Currently Bengali speaking print disabled community accessing local language content using Marathi Text to speech system via screen reader.

Besides, there are few other applications that currently testing this system such as talking dictionary, DAISY book, agro-information system and news reader. Using this system one of the newspapers in Marathidesh developed their audio version of newspaper to make mp3 of their daily content.

EVALUATION

Any real system needs to undergo rigorous testing before deployment. Though TTS testing is not a simple or widely agreed area, it is widely agreed that a TTS system has two main goals on system test; that is a synthesized speech should be i) intelligible and ii) natural. Intelligibility test can be performed by word recognition tests or comprehension tests where listeners are played a few words either in isolation or in a sentence and asked which word(s) they heard. In naturalness test, listeners are played some speech (phrase or sentence) and simply asked to rate what they hear. This can be done by mean opinion score. Since these testing may not always be the best approach so people also use unit testing approach. As our goal was to make a general-purpose synthesizer, a decision was made to evaluate it under the intelligibility criterion and unit testing on a few components. The most commonly used word recognition test - modified rhyme test (MRT) [12] was designed to test Marathi TTS system. Based on the MRT we designed a set of 77 groups - 5 words each. Therefore a set of 385 words came into testing. The words in each group are similar and differ in only one consonant and the users were asked to account which word they have heard on a multiple choice sheet. Based on the test the overall intelligibility of the system from 6 listeners is 96.96%. Besides the intelligibility test, we have performed a unit test on text normalizer and G2P converter. The performance of text normalizer is 87% only for ambiguous tokens and that of G2P converter is 89%.

CONCLUSIONS

Here the development of the first-ever complete Text to Speech (TTS) system has described, that can convert a Unicode encoded Marathi text into human speech. It is distributed under an open source license to empower both the user and developer communities. This TTS system can also be used with any available Screen Reader. In addition to the standalone TTS client, it can be integrated into virtually any application, and can also be accessed as a Web Service. Incorporating this technology in various applications such as screen reader for the visually impaired, touch screen based agro-information system, talking books, telecenter applications, e-content, etc., can potentially bridge the literacy divide in Marathidesh, which in turn goes towards bridging the digital divide. An evaluation of the system has been done based on MRT and unit testing on a few components to check intelligibility. Since the voice developed here is di-phone concatenation based and it lacks proper intonation modeling so it produces robotic speech. Therefore, a natural sounding voice needs to be made in future, which could be performed by developing a unit selection voice. Besides that, a few works need to be done in future to improve the intelligibility of the system such as POS tagger, improvement of G2P algorithm, improvement of text normalizer and working on intonation modeling.

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

REFERENCES

[1] 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 www.iosrjournals.org
[2] A.W. Black, and K.A. Lenzo, 2003, Building Synthetic Voices, Language Technologies Institute, Carnegie Mellon University and Cepstral LLC. Retrieved from: http://festvox.org/bsv/
[3] 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
[4] Newton, "Review of methods of Speech Synthesis", M.Tech Credit Seminar Report, Electronic Systems Group, November, 2011, pp. 1-15
[5] Mark Hasegawa Johnson,"Lecture Notes in Speech Production, Speech Coding and Recognition", University of Illinois, February 2000.
[6] Thierry Dutoit,"An Introduction to Text-to-Speech Synthesis", Springer, Volume 3.
[7] Christopher Richards, "Normalization of non-standard words". Computer Speech and Language (2001), pp.287–333.
[8] M.B.Chandak, Dr.R.V.Dharaskar and Dr.V.M.Thakre,"Text to Speech with Prosody Feature: Implementation of Emotion in Speech Output using Forward Parsing", International Journal of Computer science and Security, Volume (4), Issue (3).
[9] Ramani Boothalingam,V Sherlin Solomi, Anushiya Rachel Gladston,S Lilly Christina, "Development and Evaluation of Unit Selection and HMM-Based Speech Synthesis Systems for Tamil", 978-1-4673-5952-8/13, IEEE 2013 National Conference.
[10] Heiga Zen, Tomoki Toda and Keiichi Tokuda. "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006", INTERSPEECH 2005.
[11] K. Partha Sarathy, A.G.Ramakrishnan "TEXT TO SPEECH SYNTHESIS SYSTEM FOR MOBILE APPLICATIONS" http://mile.ee.iisc.ernet.in /mile/publications/softCopy/SpeechProcessing/WISP07_124_MobileTTS.pdf
[12] Paul Taylor, 2009, Text-to-Speech Synthesis, University of Cambridge, February.
[13] Ayesha Binte Mosaddeque, Naushad UzZaman and Mumit Khan, 2006, Rule based Automated Pronunciation Generator, Proc. of 9th International Conference on Computer and Information Technology (ICCIT 2006), Dhaka, Bangladesh, December 2006.
[14] C. Kamisetty and S.M. Adapa, 2006, Telugu Festival Text-to-Speech System, Retrieved from:http://festivalte.sourceforge.net/wiki/Main_PageCRBLP, 2010, CRBLP pronunciation lexicon, [Online], Available: http://crblp.bracu.ac.bd/demo/PL/
[15] 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
[16] 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
[17] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "Di-phone-Based Concatenative Speech Synthesis System for Hindi" International Journal of Advanced

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)

Research in Computer Science and Software Engineering -Volume 5, Issue 10, October-2015

[18] 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

[19] 19) 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 www.iosrjournals.org

[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

[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 "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

[25] 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

[26] 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 www.iosrjournals.org

ISSN 2054-0957 (Print), ISSN 2054-0965 (Online)