IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 20-26
e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
www.iosrjournals.org

A Corpus-Based Concatenative Speech Synthesis System for Marathi

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

Abstract: *Speech synthesis is the process of converting written text into machine-generated synthetic speech. Concatenative speech synthesis systems form utterances by concatenating pre-recorded speech units. Corpus-based methods use a large inventory to select the units to be concatenated. In this paper, we design and develop an intelligible and natural sounding corpus-based concatenative speech synthesis system for the Marathi language. The implemented system contains a front-end comprised of text analysis, phonetic analysis, and optional use of transplanted prosody. The unit selection algorithm is based on commonly used Viterbi decoding algorithm of the best-path in the network of the speech units using spectral discontinuity and prosodic mismatch objective cost measures. The back-end is the speech waveform generation based on the harmonic coding of speech and overlap-and-add mechanism. Harmonic coding enabled us to compress the unit inventory size by a factor of three. In this study, a Marathi phoneme set has been designed and a pronunciation lexicon for root words has been constructed. The importance of prosody in unit selection has been investigated by using transplanted prosody. A Marathi Diagnostic Rhyme Test (DRT) word list that can be used to evaluate the intelligibility of Marathi Text-to-Speech (TTS) systems has been compiled.*

I. Introduction

Speech synthesis is the process of converting written text into machine-generated synthetic speech. In the collected works, there are three main approaches to speech synthesis: articulatory, formant, and concatenative [1]. Articulatory synthesis tries to model the human articulatory system, i.e. the vocal cords, the vocal tract, etc. Formant synthesis employs some set of rules to synthesize speech using the formants that are the resonance frequencies of the vocal tract. Since the formants constitute the main frequencies that make sounds distinct, speech is synthesized using these estimated frequencies. On the other hand, concatenative speech synthesis is based on the idea of concatenating pre-recorded speech units to construct the utterance. Concatenative systems tend to be more natural than the other two since original speech recordings are used instead of models and parameters. In concatenative systems, speech units can be either fixed-size di-phones or variable length units such as syllables and phones. The latter approach is known as unit selection, since a large speech corpus containing more than one instance of a unit is recorded and variable length units are selected based on some estimated objective measure to optimize the synthetic speech quality.

In this research, we propose an intelligible and natural sounding corpus-based speech synthesis system for Marathi. The system consists of an analysis component which converts the text into a linguistic and prosodic description, a unit selection component based on Viterbi decoding, and a waveform generation component based on the harmonic coding of speech and the overlap-and-add mechanism. The research in this paper is directed towards agglutinative languages in general and Marathi in particular. Speech synthesis systems are currently being developed for languages like English and successful results are obtained. However, the studies on Marathi which is an agglutinative language and has a highly complex morphological structure are quite limited. In this study, we take the special characteristics of Marathi into account, propose solutions for them, and develop a speech synthesis system for the language [2]. To the best of our knowledge, this is the first unit selection based system published for Marathi

II. System Architecture

The architecture of the system is shown in Figure 1. The components shown are common in most of the speech synthesis systems that use unit selection. The system can be mainly divided into three parts: analysis (front-end), unit selection, and generation (back-end). The analysis module is responsible for producing an internal linguistic and prosodic description of the input text. This description is fed into the unit selection module as the target specification. The unit selection module uses this specification to choose the units from the speech database such that a cost function between the specification and the chosen units is minimized. The

waveforms for the selected units are then concatenated in the generation module, where the smoothing of concatenation points is also handled [1].

1.1. Text corpus

The fragments that form the text corpus have been collected from online Marathi text materials. These text fragments have been preprocessed and divided into phrases by making use of the punctuation marks. They have been checked manually and only the phrases that were complete and well-formed have been included while the rest have been discarded. Then a Greedy algorithm has been employed which aims to choose the phrases according to their phonetic context. The algorithm assigns a score to each phrase, calculated as the total frequency of the tri-phone contexts found in the phrase normalized by the number of the tri-phones [3]. Then the phrase having the greatest score is selected. The algorithm updates the frequencies of the tri-phones in the selected phrase to zero and runs on the remaining phrases. The algorithm produced 3000 phrases.

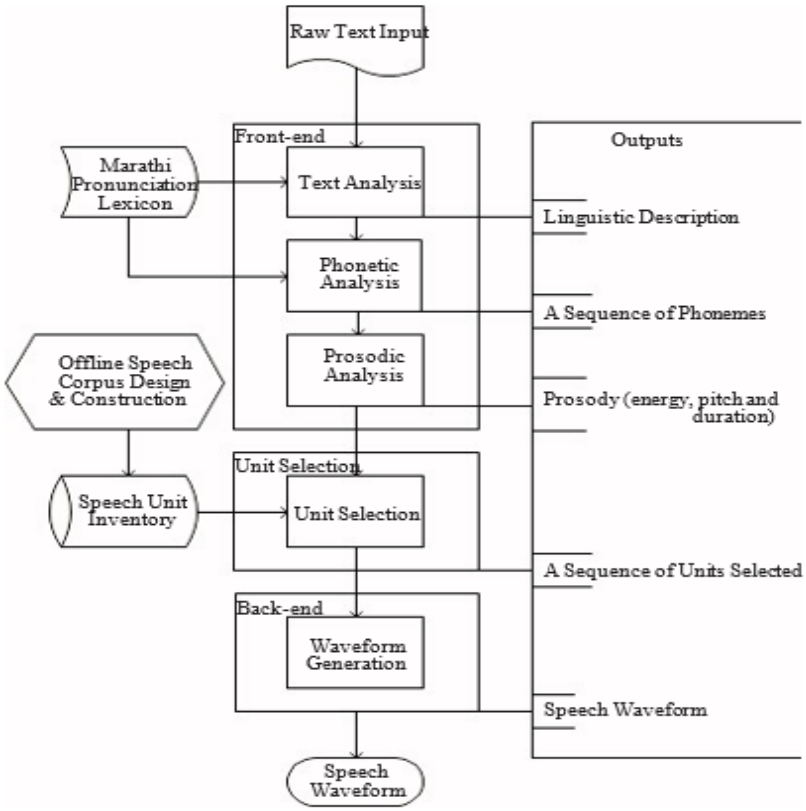

Figure 1. Corpus-based concatenative Marathi speech synthesis system architecture [1].

1.2. Speech corpus

The speech corpus used by the algorithms developed in this research contains about 20 hours of speech recorded by a professional male speaker covering the 3000 Marathi phrases in the text corpus. The speech corpus has been phonetically aligned by a speech recognition engine and then the phone boundaries have been corrected manually. The corpus has been divided into two sets: training set and test set. The test set contains 1000 phrases used for the purpose of evaluating the synthesis quality. From the remaining recordings (training set), two speech unit inventories of different sizes have been constructed. One contains all the recordings in the training set (about 18 hours of speech) and the other contains 5000 phrases (about 2 hours of speech) extracted as explained above [2][3].

III. Forming Linguistic and Prosodic Description

In a language, phonemes are the smallest units of sound that distinguish one word from another [2]. Marathi alphabet contains 56 letters classified as 15 vowels and 41 consonants.

1.3. Marathi pronunciation lexicon

A Marathi lexicon has been built containing about 3500 root words and their pronunciations. The lexicon is used to determine the pronunciations of the words and to expand the abbreviations and acronyms. The small size of the lexicon is because of the relatively simple pronunciation schema of Marathi compared to English. Marathi is a phonetic language in the sense that a simple grapheme-to-phoneme conversion (i.e. one-to-

one mapping of letters to phonemes) is possible for most of the words due to the close relationship between orthography and phonology. Most of the words in the lexicon are those for which such a direct mapping cannot yield the correct pronunciation due to vowel lengthening.

1.4. Text-to-phoneme conversion

The input text is first parsed into sentences and words by making use of space characters and punctuation marks. It is then stored in an internal data structure which is a linked list of sentence nodes, each of which is a linked list of word nodes. The sentence node structure was designed to hold sentence level information such as sentence type and the word node structure was designed to hold word level information such as POS tagging and word accent. At this stage, text normalization was also performed. The no orthographic symbols are converted into orthographic ones in the sense that abbreviations and acronyms are expanded into full forms and digit sequences are converted into written forms. The characters that cannot be represented in speech are discarded. The punctuation marks are preserved [2].

Table 1. Marathi vowels phoneme set [2].

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 phoneme set [2].

क	ka	[kə]	ख	kha	[$k^{h}ə$]	ग	ga	[gə]	घ	gha	[$g^{ɦ}ə$]	ङ	ṅa	[ŋə]
च	ca	[tsə/tʃə]	छ	cha	[$ts^{h}ə$]	ज	ja	[ʄjə/zə]	झ	jha	[ʄ$j^{ɦ}$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	[nə]
प	pa	[pə]	फ	pha	[p^{h}ə/fə]	ब	ba	[bə]	भ	bha	[$b^{ɦ}ə$]	म	ma	[mə]
य	ya	[jə]	र	ra	[rə]	ऱ	ṟa	[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	[ʄɲə]	श्र	śra	[ʃrə]

Marathi phoneme inventory consists of 41 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 56 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 [5].

1.5. Prosodic analysis

Although the system was designed to use a prosodic analysis component, currently it does not include such a component. Prosody module can provide pitch, duration, and energy information which can be used in the unit selection process to synthesize the text. We plan to add pitch contour synthesis and duration modeling in future research. However, to evaluate the effect of using prosodic analysis, we tailored the system in such a way that it can optionally use transplanted prosody from the original speech utterances. Transplanted prosody means that the duration and intonation values from recorded speech are used in the unit selection process [6]. This approach was used in the experiments to see the effect of real prosody on the output speech quality.

IV. Unit Selection Using Viterbi Algorithm

The output of the analysis module is a sequence of phonemes corresponding to the input text, each having energy, pitch, and duration values. We refer to this sequence as the target sequence. The phones are used as the basic units in this research. The speech corpus had already been processed to build a unit inventory storing the phonemes with the same prosodic features (energy, pitch, duration) and the context information. Since we use a large speech database, there is more than one instance for each phoneme, each possibly having

different phonetic context, and prosodic and acoustic realizations. Therefore, for each phoneme in the target sequence, there exist a large number of choices from the unit inventory. In concatenative speech synthesis, choosing the right units is very important for the quality of the synthesized voice. An appropriate selection of units may also allow to get rid of prosodic modifications of the selected units, which generally degrade the output speech quality. The unit selection module tries to choose the optimal set of units from the unit inventory that best match the target sequence.

1.6. Determining the optimal unit sequence

We implemented a Viterbi decoding algorithm to find the optimal unit sequence in the network of the nodes. A state transition network formed of the units in the speech inventory is shown in Figure 2, where the thick arrows indicate the connections between the selected units. The Viterbi algorithm tries to find the optimal path through the network [1, 9]. Since the number of units in unit inventory is very large, we employed some pruning methods to limit the number of units considered. By making use of a window size of three, for a target unit, we select only those units whose left and right three units are identical to those of the target unit. If there exist no such units, the search is repeated with a window size of two and finally with a window size of one.

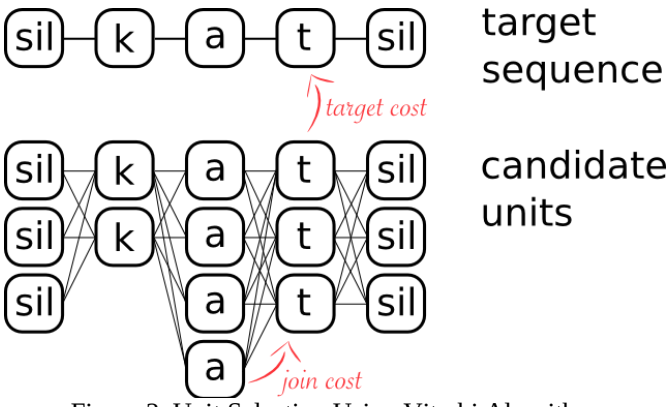

Figure 2. Unit Selection Using Viterbi Algorithm

V. Unit Concatenation and Waveform Generation

The unit selection module outputs a sequence of units from the speech inventory to be used for the generation of waveform for the input text. The waveform generation module concatenates the speech waveforms of the selected units. We used a speech representation and waveform generation method based on harmonic sinusoidal coding of speech [7, 8]. Analysis-by-synthesis technique was used for sinusoidal modeling.

The sinusoidal coding encodes the signal with a sum of sinusoids whose frequency, amplitude, and phase are adequate to describe each sinusoid. The harmonic coding is a special case of the sinusoidal coding where the frequencies of the sinusoids are constrained to be multiples of the fundamental frequency. The harmonic coding takes the advantage of the periodic structure of the speech and is very effective in coding voiced and unvoiced signals.

The harmonic coding is a parametric coding method. Unlike waveform coders which try to construct the original waveform, parametric coders (vocoders) try to encode the speech into a parametric representation that captures its perceptually important characteristics. Harmonic coders represent the speech signal using the magnitudes and phases of its spectrum at multiples of the fundamental frequency. Low bit rate harmonic coders even use the synthetic phase rather than original phase to lower the bit rate. However, a high quality speech synthesis requires that the speech representation should be transparent to the listener. Therefore, we used the original phase in the harmonic coding of speech. The coded speech quality heavily depends on the correct parameter estimation. For robust parameter estimation, we used an analysis-by-synthesis methodology.

A perfectly periodic signal can be represented as a sum of sinusoids:

$$x[n] = \sum_{k=0}^{T_0-1} A_k \cos(nk\omega_0 + \varphi_k),$$

where T_0 is the fundamental frequency of the signal, $\omega_0 = 2\pi/T_0$, φ_k is the phase of the kth harmonics, and A_k is the amplitude of the kth harmonics. For the quasiperiodic speech signals, the same equation can be used to approximate the signal. This approximation can even be used to model the unvoiced sounds. In this case, the fundamental frequency is set to 100 Hz. The error in representing the speech by a harmonic model is estimated as:

$$\varepsilon = \sum_{k=-T_0}^{T_0} \omega^2[k]\,(x[k] - \tilde{x}[k])^2,$$

Where ω is a Hamming window, x is the real speech signal and x˜ is the harmonic model for the speech signal. For parameter estimation of the harmonic coding, we use this function for error minimization criterion. Finding model parameters is a least squares problem. The values for A_k and φ_k that minimize the error are calculated by solving the linear set of equations obtained by differentiating the error function. The derivation of the linear equations is given in [8]. We used QR factorization method for solving the set of linear equations to obtain the model parameters.

The correct pitch period estimation is an important part of harmonic coding. The algorithm that we used for pitch estimation is based on the normalized autocorrelation method. The normalized autocorrelation is calculated as:

$$R_n(k) = \frac{\sum_{n=0}^{N-1} x[n]x[n+k]}{\sqrt{\sum_{n=0}^{N-1} x^2[n] \sum_{n=0}^{N-1} x^2[n+k]}}.$$

The search for the pitch was constrained to a region between 50Hz and 500Hz. We also performed some post-processing to smooth the pitch track, since the normalized autocorrelation method is error-prone. The smoothing process takes into consideration the factor that the pitch does not change drastically from frame to frame. We applied median smoothing that keeps a history of the pitch values, sorts it, and takes the one in the middle.

The model parameters are calculated in a pitch-synchronous manner using overlapping windows of two pitch periods. The scalar quantization of model parameters is performed. The unit speech inventory was compressed about three times using quantized model parameters.

The waveform generation using the model parameters for speech waveforms of units is done by taking the inverse FFT of the parameters and then overlap-and-add mechanism is used for smooth concatenation of the units.

VI. Experiments and Results

To evaluate the quality of the synthetic voice produced by the developed system, we carried out formal listening tests. The tests were of two type. The first one requires the listeners to rank the voice quality using a Mean Opinion Score (MOS) like scoring. The other test is a diagnostic rhyme test.

MOS tests are commonly used for both evaluating the effectiveness of speech coding algorithms and assessing the quality of synthesized speech. The MOS scores for speech synthesis are generally given in three categories: intelligibility, naturalness, and pleasantness.

The MOS test was carried out by synthesizing a set of 50 sentences that were selected from the speech corpus randomly and did not participate in the training set. The reason of choosing the sentences for which we have also available the original speech waveforms is that the original recordings are also used in the tests to ensure the reliability of the test results. 10 subjects (2 females) were used and they listened the sentences using headphones. The sentences were at 16 kHz and 16 bits. The subjects were instructed to rate the sentences on a scale of 1-5 where 1 is very poor and 5 is excellent. Some speech samples of speech coders having different MOS scores were presented to the subjects to ensure consistency in evaluating the speech quality. The subjects were also familiarized with the speech synthesis by listening some example utterances of varying quality.

We built five different systems and evaluated their quality. The first system uses the original recordings from the test speech corpus that were coded by our harmonic coder and reconstructed. The second system uses the unit selection synthesizer with a speech unit inventory containing about 19 hours of speech recording. The third system uses a speech inventory containing about 3 hours of recording. The latter two systems do not use prosody information and no prosody targets are specified for the target units in unit selection. The last two systems are the same as the previous two, except that the original prosody from the original recordings is used in the unit selection process [6].

Each of the 50 test sentences were synthesized by each of the five systems. Then five test sets were constructed in the following way: 10 sentences from each system were gathered to form a test set. Each set contained all of the 50 test sentences, i.e. repeating of the same sentence from different systems was not allowed. The subjects were also divided into five groups with two subjects in each. Then each test set was listened by a different group. The subjects gave ratings in terms of intelligibility, naturalness, and pleasantness to each sentence. The average MOS scores are shown in descending success rates in Table 2. Figures 3 and 4 show the scores for each system and category. The differences in system ratings were found to be significant using ANOVA analysis. The analysis yielded an F-value of about 21 whereas the critical F-values are about 3.3 and 5.0 for P=0.01 and P=0.001, respectively.

It is quite interesting that while system C is better than system E both of which use 3 hours of speech, this is not the case for systems D and B which use 19 hours of speech. In other words, for 3 hours of speech

corpus, using original prosody improves the naturalness of generated speech, whereas for 19 hours of speech corpus, it degrades the generated speech quality. It can be argued that for systems that use relatively less amount of speech corpus, using prosody information in unit selection helps to select better units in terms of prosody, hence increasing the overall naturalness of synthetic speech. On the other hand, for larger corpus, we have more units in the corpus and the unit selection is more probable to find a better acoustic and prosodic match. In these systems, using prosody information may cause the unit selection to favor prosody over acoustic appropriateness which is probably more important than prosody for naturalness.

Table 2. Systems and average scores for the MOS test.

System	Description	MOS
A	The original recordings with harmonic coding	4.91
B	Speech synthesis using 19 hours of speech	4.20
C	Speech synthesis using 3 hours of speech with original prosody	4.11
D	Speech synthesis using 19 hours of speech with o riginal prosody	4.01
E	Speech synthesis using 3 hours of speech	4.00

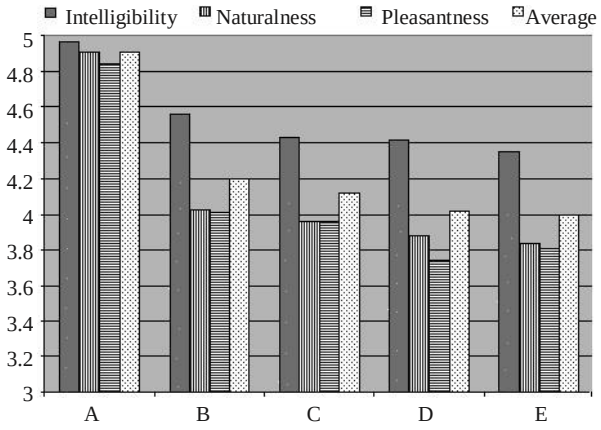

Figure 3. MOS scores with respect to system type.

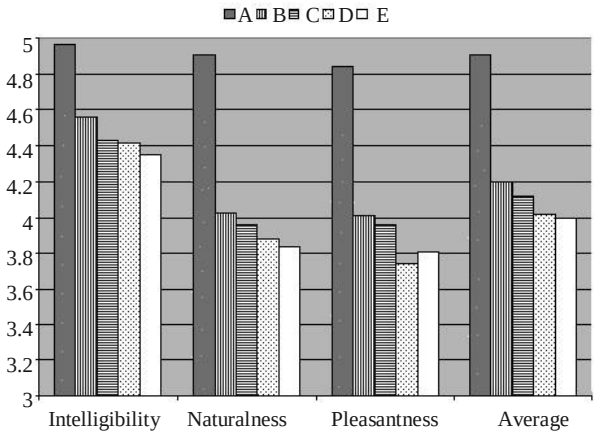

Figure 4. MOS scores with respect to test category.

We also conducted an intelligibility test. Diagnostic Rhyme Test (DRT) uses monosyllabic words that have consonant-vowel-consonant pattern. This test measures the capability of discrimination of the initial consonants for the system evaluated. The DRT word list of ANSI standard for English contains 192 words arranged in 96 rhyming pairs which differ only in their initial consonant sounds. The list has been divided into six categories depending on the distinctive features of speech. The categories have been constructed in terms of voicing, nasality, sustenation, sibilation, graveness, and compactness characteristics of the sounds. For assessing the intelligibility of the synthesized speech in Marathi, we constructed a DRT word list for based on the categories of the DRT word list of English as shown in Table 3. The DRT list was designed to exploit the distinctive features of Marathi speech at maximum.

Using the DRT word list for Marathi, we carried out an intelligibility test for our system. The randomly selected words from each pair of the DRT word list were synthesized using the system. The output speech waveforms were played to 10 native Marathi listeners who were then asked to choose which one of the words given in pairs from the DRT list they heard. The listeners were assured to have a good hearing and discrimination of sounds. The test results are shown in Table 4 as the percentage of the number of correct selections for the two systems evaluated.

Table 3. DRT word list for Marathi

Number	Marathi	in English	Number	Marathi	in English	Number	Marathi	in English
1	म्हणून	as	31	ते	to	61	ठेवले	put
2	मी	I	32	आणि	and	62	मुख्य पान	home
3	त्याच्या	his	33	एक	a	63	वाचा	read
4	त्या	that	34	मध्ये	in	64	हात	hand
5	तो	he	35	आम्ही	we	65	पोर्ट	port
6	होते	was	36	हे करू शकता	can	66	मोठ्या	large
7	साठी	for	37	बाहेर	out	67	शब्दलेखन	spell

8	वर	on	38	इतर	other	68	जोडा	add
9	आहेत	are	39	होते	were	69	अगदी	even
10	सह	with	40	जे	which	70	जमीन	land
11	ते	they	1	करू	do	71	येथे	here
12	असू	be	42	त्यांच्या	their	72	पाहिजे	must
`	येथे	at	43	वेळ	time	73	मोठा	big
14	एक	one	44	तर	if	74	उच्च	high
15	आहे	have	45	खाईन	will	75	अशा	such
16	या	this	46	कसे	how	76	अनुसरण	follow
17	पासून	from	47	म्हणाला	said	77	कायदा	act
18	द्वारे	by	48	एक	an	78	का	why
19	गरम	hot	49	प्रत्येक	each	79	विचारू	ask
20	शब्द	word	50	सांगा	tell	80	पुरुष	men
21	परंतु	but	51	नाही	does	81	बदल	change
22	काय	what	52	संच	set	82	गेला	went
23	काही	some	53	तीन	three	83	प्रकाश	light
24	आहे	is	54	इच्छित	want	84	प्रकारची	kind
25	तो	it	55	हवा	air	85	बंद	off
26	आपण	you	56	तसेच	well	86	गरज	need
27	किंवा	or	57	देखील	also	87	घर	house
28	होते	had	58	प्ले	play	88	चित्र	picture
29	अगोदर निर्देश	the	59	लहान	small	89	प्रयत्न	try
30	च्या	of	60	शेवट	end	90	आम्हाला	us

Table 4. Systems and average scores for the DRT test.

System	Description	DRT
B	Speech synthesis using 19 hours of speech	0.95
E	Speech synthesis using 3 hours of speech	0.94

By analyzing the MOS and DRT tests conducted, we have also identified the main problems and limitations of the developed system. The major sources of errors degrading synthesized speech quality are as follows: Misalignment of phones in the speech database, prosody related problems such as pitch contour discontinuities, timing errors for phones, energy differences between phones, and errors caused by acoustic variations of phones in different contexts. The latter one shows itself in the concatenation of phones from different contexts due to the lack of phones with similar contexts.

VII. Conclusions

In this paper, a corpus-based concatenative speech synthesis system architecture for Marathi has been proposed and implemented. A new Marathi phoneme set that is suitable and adequate for representing all the sounds in was given. A pronunciation lexicon for the root words in has been prepared. A text normalization module and a grapheme-to-phoneme conversion module based on morphological analysis of have been implemented. Speech corpus has been compressed by a factor of three with slight degradation on the voice quality using the harmonic coding based speech model. As the final system, a unit selection based concatenative speech synthesis system capable of generating highly intelligible and natural synthetic speech for has been developed. Subjective tests have been carried out to assess the speech quality generated by the system. A DRT word list for has been constructed to carry out the intelligibility tests. The final system got 4.2 MOS like score and 0.95 DRT correct word discrimination percentage.

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 "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

[3] 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)

[4] X. Huang, A. Acero and H.W. Hon, Spoken Language Processing, Prentice Hall, New Jersey, 2001.

[5] F. Spyns, F. Deprez, L.V. Tichelen and B.V. Coile, "Message-to-Speech: High Quality Speech Generation for Messaging and Dialogue Systems", Proceedings of the ACL/EACL Workshop on Concept to Speech Generation, pp. 11-16, 1997.

[6] http://www.phon.ucl.ac.uk/home/sampa/home.htm.

[7] R.E. Donovan, "Current Status of the IBM Trainable Speech Synthesis System", Proceedings of the 4th ISCA Tutorial and Research on Speech Synthesis, Edinburgh, 2001.

[8] R.E. Donovan, "Current Status of the IBM Trainable Speech Synthesis System", Proceedings of the 4th ISCA Tutorial and Research on Speech Synthesis, Edinburgh, 2001.

[9] M. Bernd, "Corpus-based Speech Synthesis: Methods and Challenges", Arbeitspapiere des Instituts f˙ur Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 6 (4), pp. 87-116, 2000.