Corpus Driven Malayalam Text-to-Speech Synthesis for Interactive Voice Response System

Arun Soman, Sachin Kumar S., Hemanth V. K.,
M. Sabarimalai Manikandan, K. P. Soman

Centre for Excellence in Computational Engineering and Networking
Amrita Vishwa Vidyapeetham, Amrita School of Engineering,
Coimbatore-641112, India

ABSTRACT
In a text-to-speech system, spoken utterances are automatically produced from text. In this paper, we present a corpus-driven Malayalam text-to-speech (TTS) system based on the concatenative synthesis approach. The most important qualities of a synthesized speech are naturalness and intelligibility. In this system, words and syllables are used as the basic units for synthesis. Our corpus consists of speech waveforms that are collected for most frequently used words in different domains. The speaker is selected through subjective and objective evaluation of natural and synthesized waveform. The proposed Malayalam text-to-speech system is implemented in Java multimedia framework (JMF) and runs on both in Windows and Linux platforms. The proposed system provides utility to save the synthesized output. The output generated by the proposed Malayalam text-to-speech synthesis system resembles natural human voice. Our text to speech reader software converts a Malayalam text to speech wav file that has high rates of intelligibility and comprehensibility.

Keywords
Text-to-Speech, Concatenation, Speech Synthesis, Text Normalization, Romanization

1. INTRODUCTION

Over the past years, there has been a great development in speech understanding and synthesis technology. The voice user interface (VUI) is plays an important role in human-machine communication applications such as computer systems, mobile multimedia, online ticket information, market information, customer services, personal banking information, voice-enabled equipment maintenance devices, and paperless tasks. Most of these voice-enabled applications have imparted huge financial benefits for the multimedia industries. Among the applications of speech technology, the automatic speech production, which is referred to as text-to-speech (TTS) system is the most natural-sounding technology. The text-to-speech (TTS) system will convert ordinary orthographic text into acoustic signal which is indistinguishable from human speech [1]-[10]. Today's interest is high quality speech application combined with computer resources. Text-to-speech synthesis system can be useful for several multimedia applications. For developing a natural human machine interface, the TTS system can be used as a way to communicate back through human voice. The TTS can be a voice for those people who cannot speak. The TTS system can be used to read text from emails, SMSs, web pages, news, articles, blogs, and Microsoft office tools and so on. In such reading applications, the TTS technology can reduce the eye-strain. The TTS system can be useful for disabled person to make effective communications. The existing TTS systems can be broadly classified into three groups: i) articulatory synthesis; ii) formant synthesis; iii) concatenative synthesis [11], [12]. In the past decades, the TTS has been the main research focus automatic speech production in Indian languages. Some of TTS systems for Indian languages like Hindi, Telugu, Tamil and Bengali have been developed using the unit selection and festival framework [2], [6]. In literature, each approach has its own purposes, strengths, and limitations. In practice, listener should be able to understand language information of the user textual information in generated synthesized speech waveform. In most of multimedia applications, listeners demand high quality of synthesized speech compared with natural speech. Generally speaking, the intelligibility and comprehensibility of synthesized speech should be relatively good in the naturalistic environments. Furthermore, listeners are able to clearly perceive the message with little attention, and act on synthesized speech of a command correctly and without perceptible delay in noisy environments. Although many TTS approaches, the intelligibility, naturalness, comprehensibility, and recallability of synthesized speech is not good enough to be widely accepted by users. There is still considerable room for further improvement of performance of the text-to speech production system.

In this paper, we propose corpus driven Malayalam text-to-speech system. The proposed methodology offers text-to-speech technology so that those with literacy difficulties, learning pronunciation of words and characters. We use concatenative-based approach to synthesis desired speech through pre-recorded speech waveforms. Over past decades, this approach was very difficult to implement because of limitation of computer memory. With the advancements in computer hardware and memory, a large amount of speech corpus can be stored and used to produce high quality speech waveforms for a given text. Thus, the synthesized speech preserves the naturalness and intelligibility. In this work, we collected speech wav files for 1,40,000 Malayalam words and syllables. We selected one person for recoding these words, who has uniform characteristics of speaking, pitch-rate and energy profile, and developed speech corpus [2]. Each sound files are unique. Speech corpus collected includes text from dictionary

words, commonly used words, words from Malayalam newspapers and story books, and covers different domain such as sports, news, literature, education etc. In this work, we applied concatenation approach at word level and syllable level. The quality of synthesized speech via concatenative approach is very close to natural speech. The rest of this paper is organized as follow. In Section 2, we discuss Malayalam phonology briefly. In Section 3, we describe the proposed Malayalam text-to speech system. Finally, we provide synthesized speech waveforms and conclude in Section 4.

2. MALAYALAM PHONOLOGY

Malayalam is one among the Dravidian languages in India. Malayalam is the official language of Kerala state and the Union Territories of Lakshadweep and Pondicherry. Malayalam language contains 2500 unique phonemes. Difficulties in developing Malayalam TTS include understanding Malayalam phonetics, database creation of Malayalam language, syllable level concatenation, complexity of the language etc. There are 37 consonants and 15 vowels in Malayalam language [13].

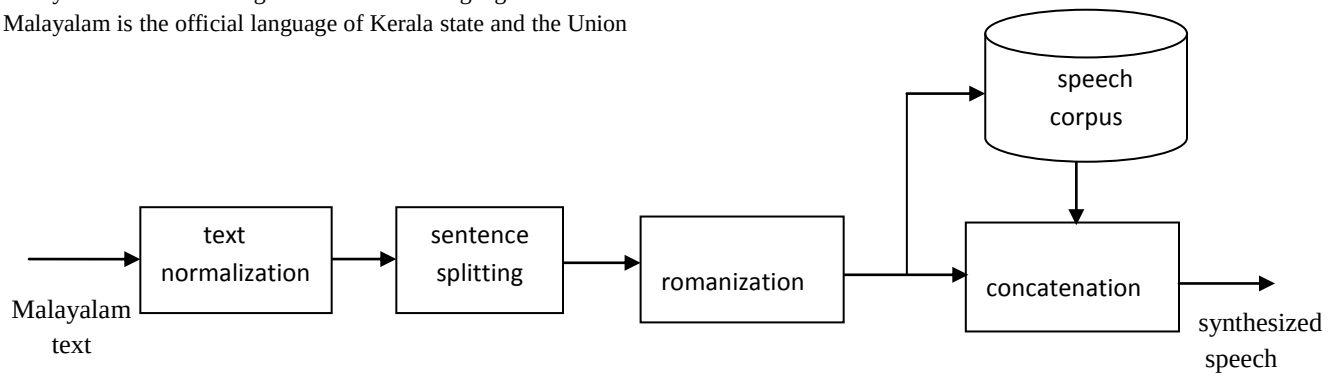

Figure 1: Block Diagram of Malayalam Text-To-Speech Synthesis System

3. MALAYALAM TEXT-TO-SPEECH SYNTHESIS SYSTEM

3.1 Text Normalization

In this stage, we perform removing of punctuations such as double quotes, full stop, comma and all. Then we will get pure sentence. We need to know that the sentence ends after a full stop (.) and not between abbreviations. It is somewhat easy to tokenize a word with help of full stop as most of the sentences will be ending with full stop. But there are some other cases where it ends with semicolon or some other punctuation like previous case. This problem can be solved by expanding the abbreviation and removing the unwanted punctuation. All the Malayalam abbreviations cannot be expanded, because some mostly used abbreviations are stored in a separate database. When certain abbreviation comes in the text, then it will search in the database for that abbreviation. If that abbreviation is present the system will replace the text, if not it will be leaving the original text as it is. It is difficult to add all the abbreviations in the database, so most commonly used abbreviations are used. The unwanted punctuation like (: , ; ” ‘ ` $) etc. are to be removed from the given paragraph to avoid confusion and not to give any disturbance in the naturalness of the speech. Each and every text in the input should be assigned some sound file for the concatenation. The second step in text normalization is normalizing non-standard words. Non standard words are tokens like numbers or abbreviations, which need to be expanded into sequences of Malayalam words before they can be pronounced. Ambiguity is the main problem with the non-standard words. For example, the number 1900 can be spoken in at least three different ways, depending on the context.

ആയിരത്തി തൊള്ളായിരം

പത്തൊൻപ്ത് നൂറ്

ഒന്ന് ഒൻപത് പൂജ്യം പൂജ്യം

To solve this problem we have to add number system into TTS. It will be coming under future work. The algorithm will cancel all the numbers coming in text, and then it will become normal text without numbers or any extra punctuation [14].

3.2 Sentence Splitting

In this stage, the given paragraph will be splitted as sentences. Separating out sentences can also be done in parallel through Graphical Processing Unit computing. From these sentences, words are separated out. Example is given below

ഞാൻ ഇവിടെ നിൽക്കുന്നു → (ഞാൻ, ഇവിടെ, നിൽക്കുന്നു)

The written sentence can be segmented easily by using white space as delimiter,

1997→(ആയിരത്തി, തൊള്ളായിരത്തി, തൊണ്ണൂറ്റി, ഏഴ്)or (ഒന്ന്, ഒൻപത്, ഒൻപത്, ഏഴ്)

There are numerous cases where the conventional delimiter segmentation approach fails. The classification and segmentation cannot be done one after the other. The first step to perform a provisional segmentation into potential written form is called token and then examining each token in turn resolves the ambiguity. This process is called tokenization; the step which generates the words from the token is called text analysis [1]. The main reason for dividing the problem this way is that it makes much easier to perform the text analysis step. By segmenting the text, generally text analysis algorithms take only one token at a time. Analysing algorithm will take input from either side of the sentence. And this makes the writing of rules and the training of algorithms quite a bit easier.

We have to consider the punctuation property which is one of the main issues in tokenization. In sentence such as "So, he went all that way (at consider cost) and found no-one here" the classic uses of punctuation are seen. Firstly it creates a boundary between two tokens which might otherwise be assumed to be linked, or it creates a link between two tokens. This might otherwise be assumed to be separate. Generally a 'full stop' ends a sentence, a 'comma' ends a clause and both create a boundary between the tokens that precede and follow them. The status markers, '?' and '!', all indicates the end of the sentence and each assign a different status to that sentence. Hyphens and apostrophes link tokens that might otherwise be assumed to be more separate. Some problems are also involved in technical language, such as rules of using white space in conjunction with punctuation are different; for example time 10:34 am, in which there is no whitespace after the colon. Some of the verbalizations of technical language speak the punctuation as words, so that the word for yahoo.com are 〈 yahoo, dot, com 〉 , where the '.' is spoken as dot. It helps in case like these to make some distinctions. This shows the need of distinguishing underlying punctuation from text punctuation.

There are two type of underline punctuation in conventional writing, one is dash and other one is hyphen. This use of one character for two underlying punctuation marks causes ambiguity and this must be resolved. This can resolved by using the model of verbalization, so that in yahoo.com the sequence is analysed firstly as an email address, then verbalized into natural language and then spoken. In such cases, representation for the word that the punctuation is spoken as word comma, word colon and so on, to distinguish from the silent form of comma and colon.

3.3 Romanization

Romanization [15] is the representation of written word with a roman alphabet. In this system Romanised form of Malayalam words/syllables are generated. For representing the written text the method used for Romanization is transliteration and for spoken word, the method is transcription. For example,

ജ്ഞാനം: jnjA1naM1

Steps mentioned above are commonly called as Text Processing

3.4 Speech corpus

Text-to-speech system based on concatenative synthesis needs well arranged speech corpus. The quality of synthesized speech waveform depends up on the number of realization of various units present in the speech corpus. A good quality microphone should be used to avoid noise in speech wav file. In text-to-speech, the accuracy of the system is calculated in the ways of naturalness and intelligibility of the synthesized speech. In this work, a female voice is used for recoding Malayalam words and syllables. We have analyzed the speech with respect to (1) the quality of the synthesized speech (2) variations in natural prosody and (3) the perceptual distortion with respect to prosodic and spectral modifications. Speaker is selected through subjective and objective evaluation of natural and synthesized wave form [1]-[16]. The best speaker means the speech produced by that speaker should have the following capabilities with respect to energy profile, speaking rate, pronunciation and intonation [2]. In concatenative TTS, the quality of speech corpus is very important because the characteristic of synthesized speech are directly related to nature of speech corpus [2].

Although with the above mentioned characteristics, the speech of selected speaker should introduce less distortion when speech segments are manipulated as per requirements. We digitized speech signal with sampling rate of 16 KHz and 16-bit resolution (PCM uncompressed data format) [2]. The speech wave files are saved according to the requirement. In future, we will optimize the text. The speech wave files corresponding to the Malayalam words are named according to their corresponding Romanized names. The words collected comprises dictionary words, commonly used words, Malayalam newspapers and story books, also different domain such as sports, news, literature and education for building unrestricted TTS [2].

3.5 Concatenation

The final stage is the concatenation process. All the arranged speech units are concatenated using a concatenation algorithm. The concatenation of speech files is done in java media framework. The main problem in concatenation process is that there will be glitches in the joint. The concatenation process combine all the speech file which is given as a output of the unit selection process and then making in to a single speech file. This can be played and stopped any where needed. The main aim of this project is to achieve good naturalness in output speech [11], [12].

3.6 Speech synthesis

In speech synthesis we are utilizing two approaches, the first one is word level synthesis that means all the words that are present in the input text is already in the speech corpus so synthesised output naturalness is very high. Second, when input word is not present in the database we synthesis the word using syllable

level concatenation. In this case naturalness will be comparatively less than word level synthesis. For Example

വണ്ടി ഇടുക്കി എത്തി.

The system first removes ". " from the input text then it check the spaces between words and break the given input into different lines. So the input sentence become 3 words,

വണ്ടി
ഇടുക്കി
എത്തി

Then the java program searches the normalized database whether the word is present or not using the help of mapping file. This mapping file includes Malayalam word with the corresponding Romanized form.

For Example,

വണ്ടി : vaN1T1i
ഇടുക്കി: iT1ukki
എത്തി: etti

"vaN1T1i" is the Romanized form of the Malayalam word വണ്ടി . The audio file is named as in the Romanized format, here we can notice that "1" is placed after "N" because the system will recognize small letter and capital letter as same. There may be a contradiction for "vanti" and "vaNTi" so in order to avoid that we put numbers next to the capital letters. We record data on different days so there may be variation in the recorded audio file. To avoid this we are normalizing the database. While synthesis we will get a normal speech without any fluctuations. Figure 2 shows the synthesized output of the given input word.

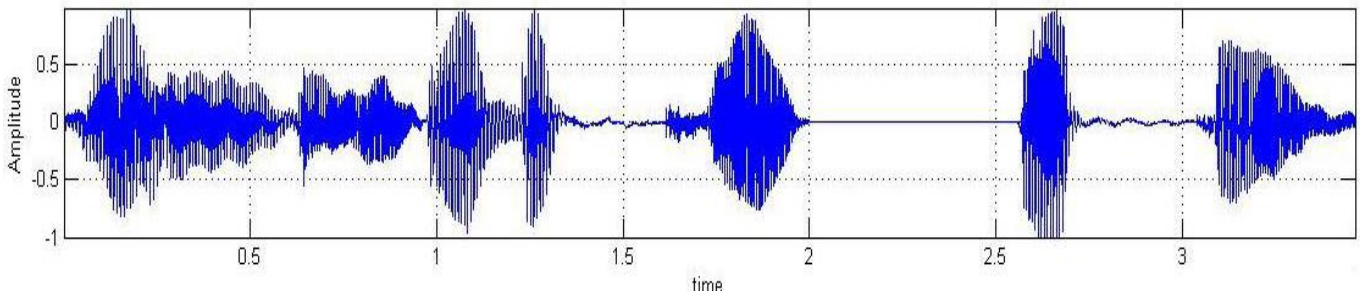

Fig 2:വണ്ടി ഇടുക്കി എത്തി (vaN1T1i iT1ukki etti)

3.7 Quality test

Voice quality testing is performed using subjective test. In subjective tests, human listeners hear and rank the quality of processed voice files according to a certain scale. The most common scale is called MOS (Mean Opinion Score) [16] and is composed of 5 scores of subjective quality, 1-Bad, 2-Poor, 3-Fair, 4-Good, 5-Excellent. The MOS score of a certain vocoder is the average of all the ranks voted by different listeners of the different voice file used in the experiment. Here tests are conducted with 5 students with the age group of 23—26 years. The tests were conducted in the laboratory environment by playing the speech signals through headphones. In the test, these 5 students were asked to audio perception test for the three synthesized sentences. Then they were asked to judge the distortion and quality of the speech. The evaluation of Malayalam TTS is shown in Table 1. Sentence-1's score is obtained by making 5 people to speak and evaluate the speech output of the test sentence with words present in the speech corpus. To obtain the score for Sentence-2, the test sentence is made with words present in the speech corpus and those which are not in the speech corpus. The score for Scentence-3 is obtained by making the test sentence with words not in the speech corpus. Figs. 3 and 4 illustrate the synthesized speech waveforms generated for two Malayalam sentences.

Table 1: Subjective test results.

Test- Set	MOS
Sentence-1	4.00
Sentence-2	3.2
Sentence-3	2.72

4. RESULT AND CONCLUSION

The speech files for Malayalam syllable and words are recorded and stored in PCM format. The speech files created are of good quality so that it will retain the naturalness of the synthesized output. A mapping file for the Malayalam words and syllables are also created. Text processing algorithm helps to produce speech synthesis successfully. Graphical user interface (GUI) is created to type the Malayalam sentences and show the output of the system. Figure 5 shows the GUI of Malayalam text-to-speech system. When we type the Malayalam sentence in the textbox shown in GUI and then clicks button 'Play', we will get the speech wave file that can be played. The text area can be cleared at any point of time by clicking the 'Clear' button. 'Save' button is used to save the synthesized output to the desired location in '.wav' format. This saved wave file can be opened in any audio editing tools like Audacity, Wavepad Sound Editor etc, and signal processing tools like MATLAB for further analysis.

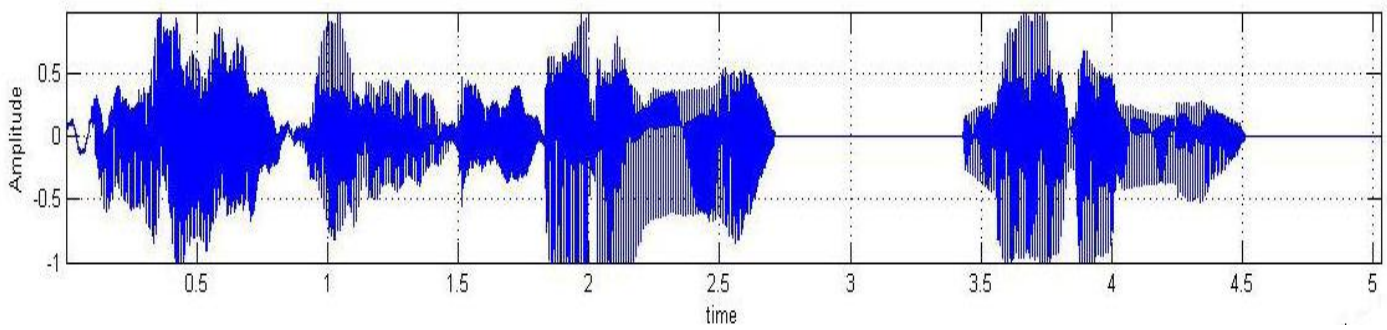

Fig. :3 ഞാന് ഇന്ന് ഒരു വണ്ടി കയറി(njA1nZ1 inn oru vaN1T1i kayarri)

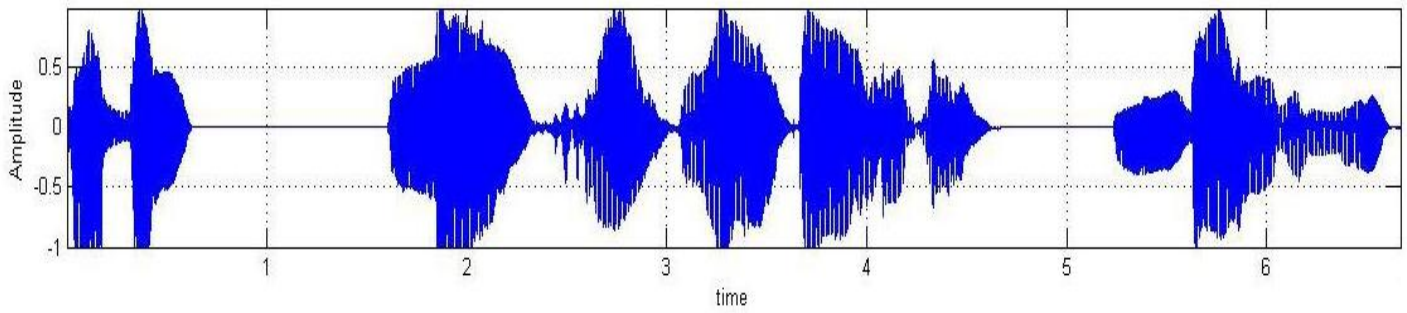

Fig. :4 അതില് നിറയെ ആളുകള് ഉണ്ടായിരുന്നു (atilZ1 nirraye A1L1ukaL1Z1 uN1T1A1yirunnu)

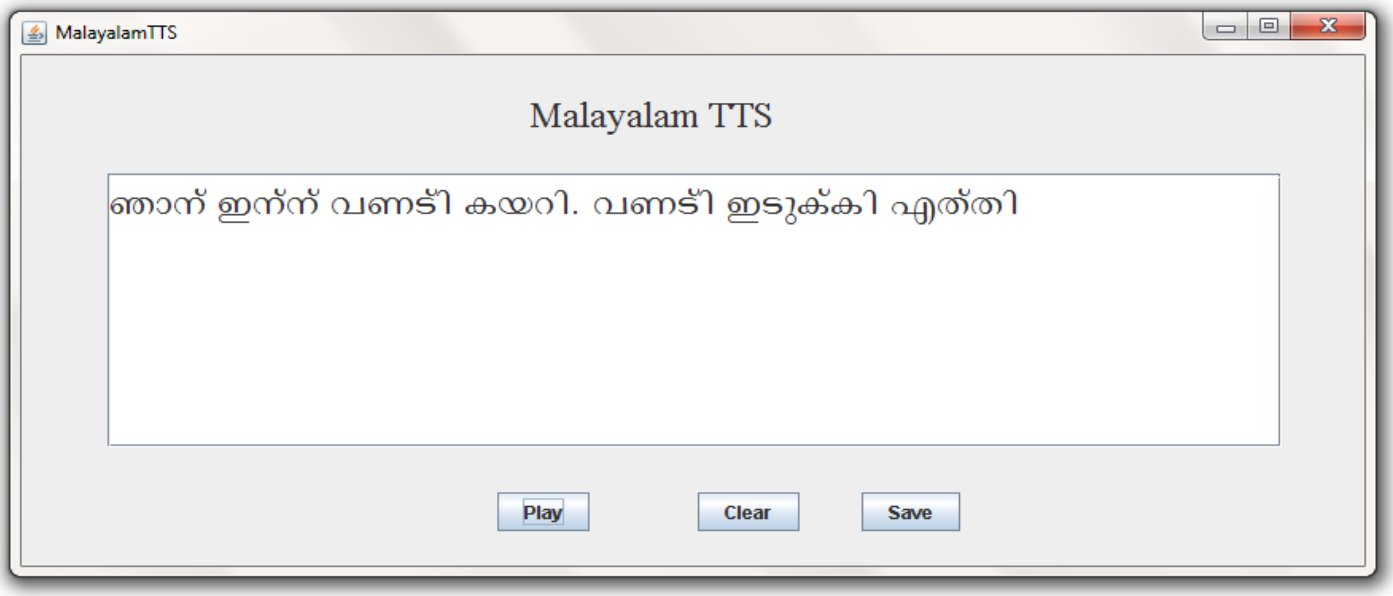

Figure: 5 Graphical User Interface of Malayalam Text-To-Speech Synthesis System

5. REFERENCES

[1] Marian Macchi, Bellcore,"Issues in text-to-speech synthesis" In *Proc. EEE International Joint Symposia on Intelligence and Systems*, pp.318-325, 1998

[2] N.P. Narendra, K. Sreenivasa Rao ,Krishnendu Ghosh, Ramu Reddy Vempada, Sudhamay Maity "Development of syllable-based text to speech synthesis system in Bengali" DOI: 10.1007/s10772-011-9094-4

[3] C. Pornpanomchai, N. Soontharanont, C. Langla, N. Wongsawat, "A dictionary-based approach for Thai text to speech (TTTS)," icmtma, In *Proc. Third Int. Conference on Measuring Technology and Mechatronics Automation*, vol. 1, pp.40-43, 2011

[4] M. N. Rao, S. Thomas, T. Nagarajan, and H. A. Murthy, "Text-to-speech synthesis using syllable like units". In *National conference on communication*, IIT Kharagpur, pp. 227–280, 2005.

[5] D. H. Klatt, "Review of text-to-speech conversion for English". *The Journal of the Acoustical Society of America*, *82*, 737–793, 1987

[6] M. Sreekanth, & A.G. Ramakrishnan, "Festival based maiden tts system for Tamil language". In *Proc. 3rd language and technology conf.*, Poznan, Poland, October pp. 187–191, 2007.

[7] S. P. Kishore and A. W. Black, "Unit size in unit selection speech synthesis," in *Proc. Eurospeech 2003*, Sept. 2003.

[8] N.S. Krishna, P.P. Talukdar, K. Bali, A.G. Ramakrishnan, "Duration modeling for Hindi text-to-speech synthesis system", in *Proc. Of Int. Conference on Spoken Language Processing (ICSLP'04)*, Korea, 2004

[9] A. Hunt, & A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", In *Proc. of EEE int. Conference acoust, speech, and signal processing*, vol. 1, pp. 373–376, 1996

[10] N. Sridhar Krishna, Hema A.Murthy and Timothy A. Gonsalves, "Text-to-Speech (TTS) in Indian Languages*", Int. Conference on Natural Language Processing, ICON-2002*, Mumbai, pp. 317.326, 2002.

[11] Robert J. Utama, Ann K. Syrdal, Alistair Conkie," Six approaches to limited domain concatenative speech synthesis". *INTERSPEECH, ICSLP*, 2006

[12] S.D. Shirbahadurkar, D.S. Bormane, R.L. Kazi, "Subjective and spectrogram analysis of speech synthesizer for Marathi tts using concatenative synthesis". *Recent Trends in Information, Telecommunication and Computing (ITC)*, 2010

[13] K. P. Mohanan, T. Mohanan "Lexical phonology of the consonant system in Malayalam" *Linguistic Inquiry The MIT Press*, volume 15, 1984.

[14] K. Panchapagesan, P.P Talukdar, N.S. Krishna, K.Bali and A.G. Ramakrishnan, "Hindi text normalization*", Fifth International Conference on Knowledge Based Computer Systems (KBCS)*, Hyderabad, India, 2004.

[15] T.Charoenporn, A.Chotimongkol, V.Sornlertlamvanich, "Automatic romanization for Thai", In *Proc. of the 2nd Int. Workshop on East-Asian Language Resources and Evaluation*, 1999

[16] M. Cernak, M. Rusko, "An evaluation of synthetic speech using the PESQ measure", In *Proc. Forum Acusticum, Budapest, 2725-2728*, 2005