Screen Readers for Linux and Windows – Concatenation Methods and Unit Selection based Marathi Text to Speech System

Sangramsing Kayte
Research Scholar
Deprtment of Computer Science & IT
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad

Monica Mundada
Research Scholar
Deprtment of Computer Science & IT
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad

Charansing Kayte, PhD
Assistant Professor
Deprtment of Digital and Cyber Forensic, Maharashtra

ABSTRACT

The research paper briefs about the implementation of screen readers for Marathi in Windows and Linux platform using unrestricted domain Marathi Text To Speech with Indian English support. The application is an integration of MTTS with open source Screen readers NVDA and ORCA. MTTS is a syllable based unit selection concatenative system, built around open source festival engine. IE support is provided for the smooth navigation and handling the English words occurring while accessing internet and other applications. The TTS is a concatenative based system in which syllable is the highest unit for concatenation. The TTS output resembles natural human voice since it uses the original speech segments for concatenation. Testing has been done with normal and differently abled users. Tuning of the system for improving the user friendliness has been done based on the feedback from the DA The system gets a Mean Opinion Score of 86.4% when evaluated by a group of DA.

Keywords

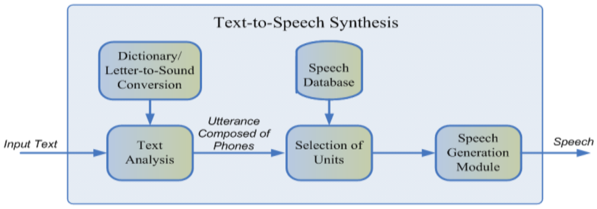
Marathi Text-to Speech, Festival, syllable, assistive technology, unit selection based speech synthesis.

1. INTRODUCTION

In this digital age visually impaired people are left out, owing to their disability [1][2]. An assistive system with Marathi TTS will change the lives of visually impaired persons with an entry to digital world. When all the information's from computer are getting through audio with the help of screen readers, any visually-impaired person can get access to computer and explore the digital world effortlessly. Screen readers play an important role in the life of visually impaired or learning disabled persons by making it possible to access computer without assistance. English is the main communication language for most of the screen readers, and its accent is very difficult to understand, so in this work we have handled IE with native Marathi accent and integrate with the screen readers. This work is done as part of the consortia project having five institutions, working on TTS six Indian Languages Hindi, Telugu, Tamil, Bengali, Marathi and Malayalam [3]. The Screen reader for Marathi with IE support is an integration of the syllable based TTS with open source screen readers is distributed freebie as an AT [4].

2. PROPOSED SYSTEM

2.1 Block diagram of a unit selection based TTS system [2]



2.2 Offline database preparation

The prime requirement for the development of unit selection based TTS is the training of database for the language. Speech corpus covering the units and the acoustic transcription, dictionary for mapping text with what is spoken and the input text are the inputs for creating a speech database [5].

2.3 Corpus Collection

The text for creating the speech database is collected from different sources like like blogs, online sources, and short stories. Few contents are manually for domain coverage

2.4 Corpus Cleaning and Processing

Format conversion is done to convert the text in different format to a unique format (UTF8). Text normalization is done to handle the number, abbreviations, suffix patterns etc in the collected text.

2.5 Text Corpus

Even though Indian languages are said to be phonetic in nature there exist pronunciation variation which needs to be taken care for syllable coverage. Text selection is done maximizing the coverage of high frequent syllables for Marathi. Pronunciation rules are applied to account the unit coverage.

2.6 Speech corpus

The selected sentences are read by professional voice. Selection of voice is done based on the consistency in the quality of voice. Recording is done in an acoustically treated room. The recording is done with specification 16 bit, 16 KHz, mono in raw wave format [1] [6].

2.7 Acoustic Transcription

Acoustic transcription for the recorded sentences is done using HTK tool [7] [8]. The sentences are transcribed at the cluster (syllable) unit level. Phone level transcriptions are also included in the database to implement the fallback mechanism for missing syllable units.

The quality of a unit selection concatenative TTS depends on the quality of the speech database, coverage of units, and accuracy of labels. Current database contains 10 hour of speech covering 10K syllables in different context. During training the syllables are clustered based on a questions concerning prosodic and phonetic context. These questions are related to the succeeding and preceding units or whether the unit is stressed or not. For each syllable a decision tree is generated whose leaves are the list of units in the database that are best identified by the questions which lead to that leaf [4] [9]. Concatenation units are identified and the segments from the wave files are concatenated for generating speech.

2.8 Text input

At synthesis time the input for generating synthetic speech is UTF8 format. The tokenize module handles the numbers and abbreviations. Numbers up to 10 digits are expanded and above are exploded and handled as single digits. For e.g. cell number 9422237003 system will read like nine four two two two three seven zero zero three. Text that not comes under Marathi Unicode range is not permitted in the system.

Language identification is done to handle English words. CMU dictionary is used as the reference for creating pronunciation dictionary for English. Each entry in the Carnegie Mellon University (CMU) English dictionary is mapped to native language phone [10] [11].

2.9 Pronounciation genaration / Letter to sound rules (LTS Rules)

The pronunciation variations are handled by applying LTS rules and using lexicon look up [4]. One of the factors which affect the quality of output speech is the accuracy of pronunciation [12]. Letter to sound rules and dictionary look up is used to handle the pronunciation and pronunciation variations in Marathi. Pronunciation variation for /k/ gemination is also taken into account to ensure the availability of proper and sufficient units for /k/ gemination in the speech database [4].

2.10 Speech Engine

The system is based on the syllable-based concatenative synthesis approach and resembles natural human voice. The TTS is built using festival speech synthesis system developed at The Centre for Speech Technology Research, University of Edinburgh [13]. Festival serves as a general framework for building speech synthesis systems that runs on multiple platforms. Festival is written in C++ and uses the Edinburgh Speech Tools, with a Scheme based command interpreter for control.

2.11 Screen readers

DA access computers using AT's like screen readers. With the aid of screen readers can access computer without human assistance. The Screen reader will identify and interpret what is displayed on the screen to the user with speech. Selection of screen reader is by considering many factors like platform, cost etc. JAWS, Window-Eyes, Dolphin Supernova etc are most commonly used screen readers and they are very high cost commercial products. We have chosen open source screen readers such as NVDA [9] and ORCA[8] for windows and Linux.

2.12 Integration with Screen readers

Linux: By default Orca is provided by the LINUX operating system distributions like Fedora, and Ubuntu. The festival TTS for Linux can be easily integrated with orca. This is done by editing the lexicon scheme file of the voice folder - proclaimed as UTF-8 encoding. We can select the voice in the list of festival voices under ORCA preferences menu.

Windows: To integrate with NVDA Festival synth driver developed by Olga Yakovleva is used. A parser module is implemented in C or C++ for Marathi syllabification rules. This parser module is added to the existing festival. Recompilation is done to incorporate the Marathi parser to festival. The recompiled festival is used for integration of NVDA

2.13 Navigation and web access issues

Frequently appearing English words in the menus and in the internet are the main issues in screen readers enabled with native languages. When English words are encountered the system breaks if it is not handled in the TTS. IE has been incorporated in MTTS to handle this issue. English words are handled by pronunciation dictionary, which maps English words to native syllable units. A pattern based mapping approach has been taken to improve the pronunciation accuracy of English words [1] [3] [6]. The words which are not in the dictionary are handled by spelling out the words.

In order to improve the accuracy of pronunciation of English words frequently used English words (frequently seen while navigating a Windows/Linux) is recorded and stored in database.

3. VOICE BUILDING CHALLENGES AND SYNTHESIS

The open source Festival TTS that runs on multiple platforms has language independent modules for building synthetic voices. For customizing the festival frame work for a Marathi we require the phone set, language specific rules, syllabification rules and corpus covering the syllables. Cluster size is reduced in festival, which controls the number of nodes in each branch of tree .The synthesis time for searching a particular unit in a tree can be reduced by this. Fundamental pitch penalty weight is adjusted to avoid fluctuations in the F0 contour of the synthesized speech [14] [15].

A Parser module is added to parse the input text in accordance with the appropriate syllable sequence. The initial, final and medial syllables are clustered separately and stress/accent is assigned for the syllables with long vowel.

When synthesizing the festival's unit selection algorithm will selects appropriate decision tree and searches for a suitable unit which is closer to its cluster center. The algorithm selects the units by optimizing the cost of joining two adjacent units while synthesizing the text. Missing syllables are handled by a fall backing mechanism [3]. If the unit is not covered in the database, it searches for the next lower syllable by trimming consonants from the syllable and replaces the unit with a suitable combination of Consonant and lower syllable. For example a pattern like CCVC is replaced with either C CVC or CCV C.

4. SYSTEM PERFORMANCE AND TESTING

MOS is calculated for subjective quality measurement. It is calculated for the synthesized speech using the Unit selection synthesis and 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 [1] [3] [6]. The total scores are summed up. Each score will range from 1-5 Positive questions will be calculated as scale position minus one and with negative questions the scale position minus five is the score and multiplying the total scores by 2.5 to get the overall value of the system usability out of 100.The selection criteria for listeners in which they of 60 minutes [14]. The tests are conducted in the laboratory environment, the MOS score obtained after evaluation is 86.4 %. Do not participated in any quality test for synthetic speech, and familiar with Marathi. Test conducted for a maximum

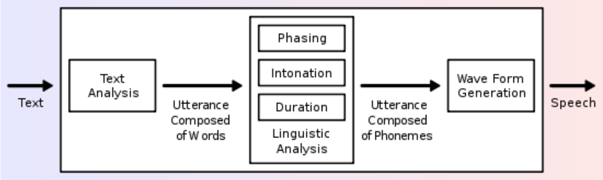


Fig 2. Block diagram of Text to speech synthesis cycle

5. CONCLUSIONS

Current unit selection system showed a 92-45% coverage of syllables for the various inputs. The system produces a good quality speech. In future we are planning to implement the following

- A small footprint TTS running on any android based platforms
- Emotional speech synthesis – especially story telling
- Integration of TTS with OCR/ASR for text reading and enabling hands free navigation

6. REFERENCES

[1] 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

[2] 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

[3] 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

[4] 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

[5] 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

[6] 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

[7] Sangramsing N.kayte "Marathi Isolated-Word Automatic Speech Recognition System based on Vector Quantization (VQ) approach" 101th Indian Science Congress Jammu University 03th Feb to 07 Feb 2014.

[8] Hidden Markov Model Tool Kit and License http://htk.eng.cam.ac.uk/download.shtml

[9] Automatically clustering similar units for unit selection in speech synthesis by Alan W Black and Paul Taylor, 10.1.1.69.472.pdf

[10] Implementation of Bilingual TTS using Festival Frame work" by Aswathy P V, Sajini T and V K Bhadran for the National conference at Kerala university, Kariavattom

[11] 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

[12] High quality Arabic concatenative TTS, by Abdelkader Chabchoub and Adnan Cherif, Signal & Image Processing : An International Journal (SIPIJ) Vol.2, No.4, December 2011 Signal Processing Laboratory, Science Faculty of Tunis, 1060 Tunisia, 2411sipij03.pdf

[13] A. Black and K. Lenzo, "Building voices in the Festival speech synthesis system," http://festvox.org/bsv/,

[14] Monica Mundada, Sangramsing Kayte "Classification of speech and its related fluency disorders Using KNN" ISSN2231-0096 Volume-4 Number-3 Sept 2014.

[15] 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)