RESEARCH ARTICLE OPEN ACCESS

Approach To Build A Marathi Text-To-Speech System Using Concatenative Synthesis Method With The Syllable

Sangramsing N. Kayte[1], Monica Mundada[1],Dr. Charansing N. Kayte[2], Dr. Bharti Gawali*
[1,3]Department of Computer Science and Information Technology Dr. Babasaheb Ambedkar Marathwada University, Aurangabad
[2]Department of Digital and Cyber Forensic, Aurangabad, Maharashtra

ABSTRACT
Marathi is one of the oldest languages in India. This research paper describes the development of Marathi Text-to-Speech System (TTS). In Marathi TTS the input is Marathi text in Unicode. The voices are sampled from real recorded speech. The objective of a text to speech system is to convert an arbitrary text into its corresponding spoken waveform. Speech synthesis is a process of building machinery that can generate human-like speech from any text input to imitate human speakers. Text processing and speech generation are two main components of a text to speech system. To build a natural sounding speech synthesis system, it is essential that text processing component produce an appropriate sequence of phonemic units. Generation of sequence of phonetic units for a given standard word is referred to as letter to phoneme rule or text to phoneme rule. The complexity of these rules and their derivation depends upon the nature of the language. The quality of a speech synthesizer is judged by its closeness to the natural human voice and understandability. In this research paper we described an approach to build a Marathi TTS system using concatenative synthesis method with syllable as a basic unit of concatenation**.**

Keywords - Text processing, speech generation, Marathi syllable phoneme, grapheme.

I. INTRODUCTION

A. Speech Synthesis

A speech synthesis system is a computer-based system that produce speech automatically, through a grapheme-to-phoneme transcription of the sentences and prosodic features to utter. The synthetic speech is generated with the available phones and prosodic features from training speech database [1] [2] [5]. The speech units is classified into phonemes, diaphones and syllables [3] [4]. The output of speech synthesis system differs in the size of the stored speech units and output is generated with execution of different methods. A text-to-speech system is composed of two parts: a front-end and a back-end. The front-end has two major tasks. First, it converts raw text containing symbols like numbers and abbreviations into the equivalent words. This process is often called text normalization, preprocessing, or tokenization. Second task is to assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units like phrases, clauses, and sentences[12][13]. Although text-to-speech systems have improved over the past few years, some challenges still exist. The back end phase produces the synthesis of the particular speech with the use of output provided from the front end. The symbolic representations from first step are converted into sound speech and the pitch contour, phoneme durations and prosody are incorporated into the synthesized speech[12][13][14][15].

The conversion of words in written form into speech is non-trivial. Even if we can store a huge dictionary for most common words; the TTS system still needs to deal with millions of names and acronyms. Moreover, in order to sound natural, the intonation of the sentences must be appropriately generated. Synthesis of speech cannot be accomplished by cutting and pasting smaller units together. Attention has to be paid to smoothing out the discontinuities in such a process so that the resulting signal approximates natural speech. According to the speech generation model used, speech synthesis can be classified into three categories as Articulatory synthesis, Formant synthesis and Concatenative synthesis. Based on the degree of manual intervention in the design, speech synthesis can be classified into two categories Synthesis by rule and Data-driven synthesis[14][15].

Prosody and intonation are quite important for natural sounding speech. There are in existence speech synthesis systems which replicate the prosodic features of human speech [6]. This involves fairly complex parsing of the input sentences and using rather complex rules to determine the intonation patterns.

Generation of sequence of phonetic units for a given standard word is referred to as letter to phoneme rule or text to phoneme rule. Voiced sounds were simulated with a computer model of the vocal fold composed of a single mass vibrating both parallel and perpendicular to the air flow.

The work is divided into 3 main modules.

- ¬ Converting Marathi script to Unicode
- ¬ Differentiating Grapheme
  - o Combination of Consonant-vowel ii. Combination of Consonant-consonant
- ¬ Generating voice
  - o Identify the Grapheme recognizer ii. Identify the Marathi Audio source

B. Converting Marathi Script to Unicode

In English ASCII characters are used where as In Marathi Unicode characters are used. ASCII takes 8-bits for each character. Unicode takes 16-bits for each character. Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language [7][12][15].

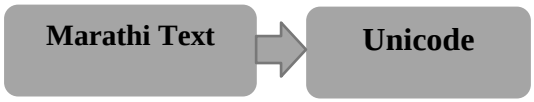

Fig.1. Converting Marathi script to Unicode

II. DIFFERENTIATING GRAPHEME

In natural speech, durations of phonetic segments are strongly dependent on contextual factors. For synthetic speech to sound natural, the module for computing segmental duration must mimic these contextual effects as closely as possible.

Grapheme:

Graphemes are "functional spelling units" encompassing one or more letters of the text input, a grapheme in the text input corresponds to a single phoneme.

Phoneme

Phones characterize any sound that can be produced by a human vocal tract, if a phone is part of a specific language; it becomes a phoneme of the language [6]. Phonemes are the elementary sounds of a language[12] [13].

In this research we are going to differentiate A character in Indian language scripts is close to syllable and can be typically of the following form:

Different Combinations:

V-Vowel
C- Consonant
C+V-Consonant+Vowel
C+C-Consonant+Consonant
C+C+V-Consonant+Consonant+Vowel
C+C+C--Consonant+Consonant+Consonant

III. GENERATING VOICE

The function of Text-To-Speech system is to convert the given text to a spoken waveform. This conversion involves text processing and speech generation processes. These processes have connections to linguistic theory, models of speech production, and acoustic-phonetic characterization of language [3]. Text processing including end-of-sentence detection, text normalization. Word pronunciation, including the pronunciation of names and the disambiguation of homographs. In this approach, the pre-recorded speech segments which are to be used in the synthesizer are stored exactly as how it is recorded. Additional information of the speech waveform is attached to the sound to provide proper annotation of the speech waveform. To synthesize a particular language, required units (di-phones) from the database which doesn't contain any language specific information and these selected units were then typically altered by signal processing functions to meet the language specific target specification generated by different modules in the synthesizer [4]. To build a voice/speech for a language text, the steps involved are as follows.

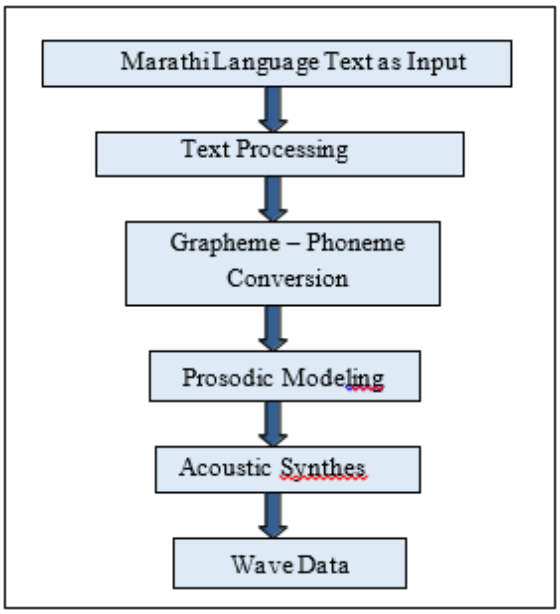

Fig 3. Generating Voice [13] [14]

A. Existing System

Many of the improvements in speech synthesis over the past years have come from creative use of the technologies developed for speech recognition. We can build systems that interact through speech, which the system can listen to what was said, compute or do something, and then speak back, using spoken language generation. Festivox is speech synthesis system developed by speech group at CMU; it provides a general framework for building speech synthesis systems. It offers full text to speech through a number API's from shell level, though a scheme command interpreter, as a C++ library, from Java, and an Emacs interface. The Festvox project: automating the processes involved in building synthetic voices for new languages. Creating festival TTS for other languages, especially Indian languages is very similar and will mostly require language specific changes to be made to the code in these modules. Festival uses diphone as basic unit [3][4][11].

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware products. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech. The overview of the problems that occur during text-to-speech (TTS) conversion and describe the particular solutions to these problems taken within the AT&T Bell Laboratories TTS system [11].

IV. MARATHI LANGUAGE

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 developed from Sanskrit. The basic units of writing system are characters which are orthographic representation of speech sounds. A character in Indian language scripts is close to syllable and can be typically of the following form: C, V, CV, CCV and CVC, where C is a consonant and V is a vowel. There are about 43 consonants and about 28 vowels in Marathi languages. An important feature of Marathi language scripts is their phonetic nature. There is more or less one to one correspondence between what is written and what is spoken. The rules required to map the letters to sounds of Marathi languages are almost straight forward. All Maharashtran language scripts have common phonetic base [8][9][10] [6].

Syllabification Rules

There is almost one to one correspondence between what is written and what is spoken in Maharashtra Languages. Each character in Marathi language script has a correspondence to a sound of that language. In Marathi languages, a consonant character is inherently bound with the vowel sound /a/, and is almost always pronounced with this vowel. This occurs at both word final and word middle positions. A few heuristic rules to detect IVS of a consonant character are noted below. While letter to phone rules are almost straight forward in

Marathi languages, the syllabification rules are not trivial. There is need to come up with some rules to break the word into syllables [3]. We have derived certain simplistic rules for syllabification i.e. rules for grouping clusters of C*VC* based on heuristic analysis of several words in Marathi language [3][4].

V. SYSTEM ARCHITECTURE

The main purpose of the system is to convert an arbitrary text into its corresponding spoken waveform. Text processing and speech generation are two main components of a text to speech system. To build a natural sounding speech synthesis system, it is essential that text processing component produce an appropriate sequence of phonemic units. Generation of sequence of phonetic units for a given standard word is referred to as letter to phoneme rule or text to phoneme rule. The complexity of these rules and their derivation depends upon the nature of the language [8][9][10]. In Marathi TTS the input is Marathi text in Unicode.

Text syllabification will segment the normalized text to syllable unit according to Marathi language rules. Phone set Definition module defines the complete set of phones used in Marathi speech. It also includes feature definitions of these phones. Lexical Analysis module is used to arrive at the phones that make up the pronunciation of a particular word. Since Marathi is phonetic in nature, we do not require a dictionary for lexical analysis. Instead, this module defines letter-to-sound rules (LTS) which are used to arrive at the speech phones based on the spelling of the word

The input text is converted into readable text and the syllabification module will generate the sequence of syllables that should be extracted from the speech corpus to be concatenated and play the sound files.

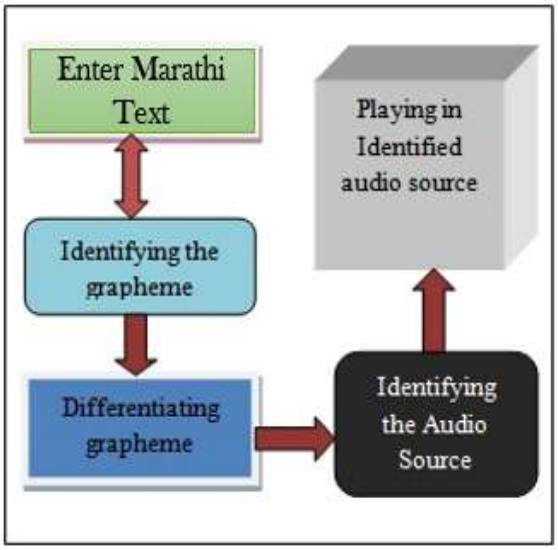

Fig 4. Architecture of the Speech Synthesis System

A Marathi TTS system using syllables as basic unit of concatenation is presented. The quality of the synthesized speech is reasonably natural. The proposed approach minimizes the co articulation effects and prosodic mismatch between two adjacent units. Selection of a unit is based on the presiding and succeeding context of the syllables and the position of the syllable. The system implements a matching function which assigns weight to the syllable based on its nature and the syllable with maximum weight is selected as output speech units. We have observed the efficiency of this approach for Marathi language and found that the performance of this approach is better.

The results showed that there is a strong correlation between the values of the source parameter in the vowel midpoint and the vowel duration. The same parameters tend to decrease on vowel onsets and to increase on vowels offsets. This seems to indicate a prosodic nature of these parameters requiring special treatment in concatenative-based TTS systems that use source modification techniques, such as pitch synchronous overlap add and multiples.

VI. CONCLUSION

Speech synthesis techniques , it is much easier to build a voice in a language with fewer sentences and a smaller Speech. The proposed approach minimizes the co articulation effects and prosodic mismatch between two adjacent units. Selection of a unit is based on the presiding and succeeding context of the syllables and the position of the syllable. The system implements a matching function which assigns weight to the syllable based on its nature and the syllable with maximum weight is selected as output speech units.

We have observed the efficiency of this approach for Marathi language and found that the performance of this approach is better. Marathi TTS system using syllables as basic unit of concatenation is presented. The quality of the synthesized speech is reasonably natural. The proposed approach minimizes the co articulation effects and prosodic mismatch between two adjacent units. Selection of a unit is based on the presiding and succeeding context of the syllables and the position of the syllable. The system implements a matching function which assigns weight to the syllable based on its nature and the syllable with maximum weight is selected as output speech units. We have observed the efficiency of this approach for Marathi language and found that the performance of this approach is better.

REFERENCES

[1.] 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

[2.] 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.

[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, 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

[5.] 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

[6.] 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

[7.] 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

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

[9.] 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

[10.] 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)

[11.] ZenH.,NoseT.,YamagishiJ.,SakoS.,MasukT.,Black A.W., andTokudaK.,"The hmmbased speech synthesis system version2.0," in Proc.ofISCASSW6, Bonn, Germany,2007.

[12.] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte " Performance Calculation of Speech Synthesis Methods for Hindi language IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 13-19e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[13.] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "A Corpus-Based Concatenative Speech Synthesis System for Marathi" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 20-26e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[14.] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "A Marathi Hidden-Markov Model Based Speech Synthesis System" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 34-39e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[15.] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "Implementation of Marathi Language Speech Databases for Large Dictionary" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 40-45e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197