North-East Research Conclave
Sustainable Science and Technology Extended Abstract May 20-22, 2022
Track (AM/SM/AF/AI/BD/DM/HR/ID/MT/PR/SD/SE/EG/TD/TL/RD): AI

Spelling Generation with Text Decomposition Technique in Unicode and its Application in Text to Speech Conversion

Chandamita Nath[1] and Bhairab Sarma[2]

[1] University of Science & Technology, Meghalaya
e-mail: dipudoili99@gmail.com,
[2]University of Science & Technology, Meghalaya
e-mail: sarmabhairab@gmail.com

Abstract:

Speech is the most popular natural way of communication among people. For interaction with the computer also, speech is considered as the highly used and desired medium of all forms of communication from the beginning of the man-machine interaction [1]. This is the main reason why Text-to-Speech conversion has been studied in various ways to make communication with machines more likely as human. Text to Speech (TTS) conversion is an activity under Natural Language processing (NLP). Many approaches are used in TTS. Among them a dictionary based approach is a traditional approach and merely used in many cases. However there are few drawbacks in this approach. The performance of this system is based on the size of the dictionary. Moreover, searching time is another factor in this approach. For English language conversion it can be tuned to some extent. For Indian languages, this technique does not work perfectly. The main reason behind it is the formation structure of characters. English is considered as the common language for communication. Characters are coded here with ASCII code which is 7 bit structure [2]. In case of Indian languages, where characters are coded with Unicode which is 16 bit format, tuning with sound becomes complicated. There are lots of challenges in NLP when Indian languages are selected as target language. The objective of this paper is to introduce a new concept of Sound Database in the field of Text to Speech (TTS) conversion system, which is one of the major applications of Natural Language Processing (NLP). In this research we develop a Spelling Generator for those texts which are unavailable in the dictionary decomposing text into independent character. Hindi and Assamese language will be considered as target language. Assamese is the official language as well as scheduled language of India spoken by the people of Assam and the entire North-east. Character formation techniques are similar for all Indo-Aryan languages with Unicode format. All Indian languages are highly inflectional in nature. Upon inflection, structure of the word gets modified and it pronounces too. Therefore pronounce generation is a difficult task for these languages. To adopt this approach, a prototype system has been developed using PHP as front end and MySQL as backend. In this research

work, parallel task also developed to implement in this spelling generator. Tokenization is one of the important task where a written sunk of text is broken down into independent token preserving grammatical rule of the experimented language. During tokenization, features extraction techniques were also evolved. Features extraction is applied for root word steaming [3, 4].

The prototype model consists of two main modules. Module 1 (as given in figure 1) is a of spelling generator that reads each dictionary entry word with a common voice. The Module 2 (in Figure 2) depicts a sound database generation.

Module1: Spelling Generator

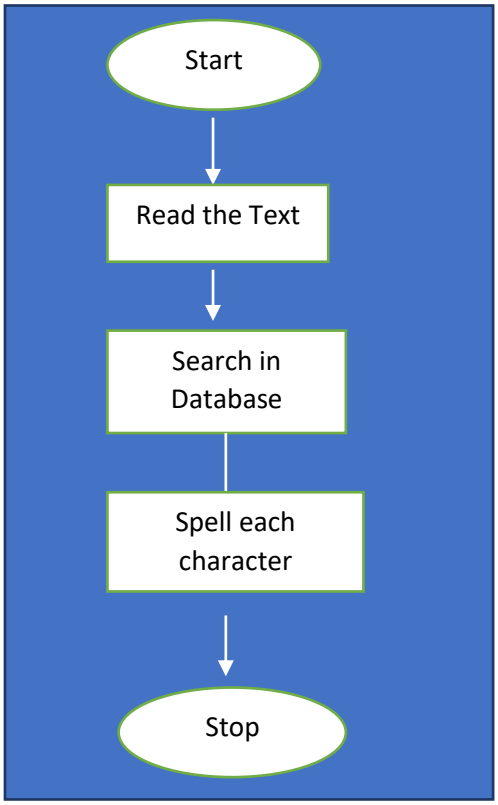

Figure 1: Spelling Generator

Module2: Sound Database

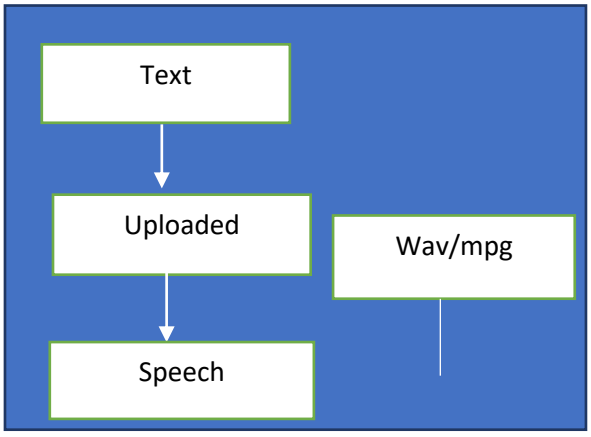

Figure 2:Sound Database

The system was tested against some text collected from some prominent Assamese and Hindi news papers. Output results of these samples testing for Assamese text are as given two tables. Provisions for reading a single word and uploading a full text file were available in the system. Results of both the cases are visualized in Table 1 and Table 2.

Case 1:In the case of uploading input as words-

TABLE 1: SPELLING ACCURACY (IN PC)

No. of Words Uploaded	No. of Words Spell Correctly	No. of Words Spell Wrongly	No. of Words not Spelled	Percentage of Accuracy
100	90	5	5	90
200	160	25	15	80

Case 2: In the case of uploading input as a text file -

TABLE 2: SPELLING ACCURACY FOR FILE (IN PC)

No. of Files Uploaded	No. of Words Spell Correctly	No. of Words Spell Wrongly	No. of Words not Spelled	Percentage of Accuracy
1st File (100 words aprox)	30	50	20	30
2nd File (200 words aprox)	60	100	40	30

References:

1. Sasirekha.D, Chandra.E,(2012) ,"Text to Speech: A Simple Tutorial", *International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307*

2. A.Akshay et.al.,(2018),"A Survey on Text to Speech Conversion", *International Journal of Trend in Research and Development, Vol5(2).*

3. Sarma B, Purkayastha BS, (2012), "A Practical Tokenizer for Part of Speech Tagging of English Text", published in International Journal of Research in Computing & Management(IJRCM): http://www.irjcm.com, Vol. 2, Issue No. 10, ISSN: 2231-5756

4. Ifeanyi. Nwakanma et.al.,(2014), "Text-To-Speech Synthesis (TTS*)", International Journal of Research in Information Technology. Volume 2, Issue 5.*

5. ShetakePoonam,.S.,et.al.,(2014),"Review Of Text To Speech Conversion Methods" *International Journal of Industrial Electronics and Electrical Engineering, Volume-2, Issue-8.*