Text to Speech Synthesis for Hindi Language using Festival Framework.

Mrs. Mangal Joshi[1], Samridhi Agarwal[2], Shabnam Shaikh[3], Priya Pitale[4]

[1]Professor, Dept. of Electronics and Telecommunication (E&TC) Engineering, Cummins College of Engineering, Maharashtra, India
[2,3,4]Student, Dept. of E&TC Engineering, Cummins College of Engineering, Maharashtra, India

---***---

Abstract - *The paper is based on developing a complete system that takes text file as input from the user and gives output in audio form. There exist many text to speech synthesizer but very few have the regional toning and soothing voices that are natural sounding. A visually impaired person using speech synthesis as a platform to hear the data instead of reading and a tongue-tied person's ability to express through speech synthesizer as a surrogate voice is one of the motives. Syllable units are chosen mainly because Indian languages are syllable-centric in nature. Festival is a platform which serves as text to speech synthesizer for many languages. The system uses the festival software which is based on syllable segmentation method. Extraction of syllables and the concatenation constitute to the process of converting Hindi text into speech form.*

***Key Words*: Hindi TTS (Text to Speech), Syllable segmentation, festival, speech synthesis, Hindi Language, etc.**

1. INTRODUCTION

Speech research aims to build the systems that have human-like capabilities in generating, understanding and encoding speech for the range of machine to human interactions. Speech is one the primary medium for communication, so it is natural for human being to expect to have communication in spoken form with computer devices. Hindi is an Indo-Aryan language spoken by 545 million people, 425 million of them are native speakers. Hindi has 13 vowels and 33 consonants and it is spoken using a combination of both.

Text-to-Speech synthesis has the potential to make ICT (Information and Communication Technology) based services accessible to people which is very beneficial. However, good quality Hindi TTS systems that can be used potentially are not yet existing. None of the existing TTS systems are of a quality that can be compared to TTS systems in languages like English, German and French. The main reason for this is to develop a TTS system in a new language like Hindi needs inputs for resolving language-specific issues. We are choosing the Festival framework for developing Hindi TTS. As Festival does not provide the complete language processing support specific to some languages, there is a need for augmentation to facilitate the development of TTS systems in certain new languages.

This syllable-based TTS system aims to work using concatenative speech processing technique. It will be boon for Hindi speakers if the user interfaces with the computer is in Hindi and that too in the form of speech. One of the greatest applications of text to speech converter is Natural language interface to the user. The synthesizer will act as an automatic text reader for blind or specially-abled people. Another important application of it can be reading web pages, emails as well as newspapers. According to the previous researches in 2016, the [1] concatenative method for speech synthesis was used. The pitch frequencies of the sound signals were extracted by cepstral pitch detection algorithm in the noiseless environment. In the same year [2], the system so developed accepted the written text of any language of Devanagari script via MS word utility through MATLAB which then converted into Romanized script under text analysis and tokenization was used in order to map the respective phoneme.

1.1 Text to Speech Conversion

The text to speech synthesis has two stages:

1.) Training phase: Hindi words are segmented into syllable sized units using segmentation techniques. Each segment is given a unique label. An audio file database of the unique labels is then provided to the controller.
2.) Synthesis phase: The text file (to be synthesized) is imported to the program. Here, the logic is provided to the controller for segmentation of the Hindi words into syllables. The syllables will be matched with the database provided initially in the training phase. Using concatenation, the word from the dictionary (database) is created. Through audio amplifiers and speakers, the speech will be generated. The characteristics like fluency, softness, accuracy will be tested. Accurate speech is expected according to the input provided.

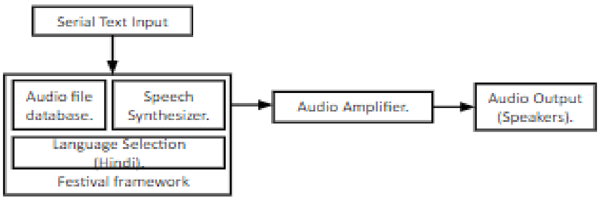

Fig. -1: Basic TTS system

1.2 Festival Framework

The Festival Speech Synthesis System is a general multi-lingual speech synthesis system which was developed by Alan W. Black Centre for Speech Technology Research (CSTR) at the University of Edinburgh. Festival is designed so that it can support multiple languages, it comes with support for the English language (British and American pronunciation), Welsh and Spanish. Voice packages exist for several other languages, such as Spanish, Finnish, Italian, Polish and Russian. Some of the Festival Speech tools are festival-2.4, festvox-2.1, speechtools-2.4, festlexCMU. As Festival does not provide the complete language processing support specific to some languages there is a need for augmentation to facilitate the development of TTS systems in certain new languages. For this application, we used Festvox as our basic framework. Festvox 2.7.0 is an open source text to speech architecture.

2. METHODOLOGY

The festival toolbox operates on a command line interface in a LINUX operating system. To get Festvox on LINUX OS,

1. Obtain the Festvox version 2.7.0 from the festival software website. Create a directory dedicated to the software. Keep all the setup related files in the same folder such as (festival-2.4-release.tar, speech_tools-2.4 release.tar, festvox-2.7.0release.tar, festlex_OALD.tar.gz, and festlex_POSLEX.tar)
2. Unpack all the tar.gz files using the tar commands.
3. Compile the speech tools by following the steps in the installation guide using gmake test and gmake install commands.
4. Compile the festival by following the installation guide in the festival folder. (commands -./configure , gmake test, gmake install)
5. After the successful installation of a festival, Compilation of festvox is needed.
6. To run the festival, Export the three variables named PATH, FESTVOXDIR, ESTDIR.
7. Next, execute the festival command to go to the festival shell.

For Introducing a new language voice in festival or to make your own festival enhanced voice simulator, template is to be designed.

1. Firstly create a directory to hold the voice inside the festival folder.
2. Build the basic structure of the new voice to be added.

Define the phone set for the language which is the set of symbols defining whether it is a consonant or vowel. The places of vowels as per each nasal sound is located and given. While we made the basic template a schema file for phone set is created, where we have to define this. Considering each parameter for the Hindi language corresponding schema files (.scm) are created which will generate more natural sounding audio output.

The database is created by recording Hindi words and labeling them according to their contents. Store all the recorded .wav files under one directory [3]. Phone synthesis can be tested after this step and if any labeling errors present can be corrected. Pronunciation can be defined using a large set of databases (lexicons) or using a letter to sound rules. Festival uses lexicon structure for pronunciation. After adding various intonation models (.scm files) basic synthesizer can be tested in the festival.

3. CHALLENGES IN TEXT TO SPEECH SYNTHESIS

Text to speech synthesis is the approach of converting text input into an audio form. There are many TTS synthesizers available for many languages. There is also so many software available for Indian languages. Text to speech conversion is totally dependent on language. Thus, it is very simple for some languages and complicated to others. A large set of different rules and their exceptions is needed to produce correct pronunciation and prosody for synthesized speech.

Text processing is the initial stage of converting text into speech. Challenges arising in text processing is while reading the numbers, units, abbreviations, dates, special characters, and symbols, etc. For example, the number 1990 can be read as nineteen ninety if it's a year and one thousand nine hundred ninety if it's number. Also, Kg. should be read as kilograms. To process such texts accurately, a very large set of rules need to be applied. Once the text is processed another difficult task is to produce correct pronunciation for the word which requires a large database. Amount of stress applied for a particular word in a sentence, finding proper duration for pronunciation and correct intonation is also needs to be taken into consideration. A timing at sentence level or grouping of words into phrases correctly is difficult because prosodic phrasing is not always marked in the text by punctuation, and phrasal accentuation is almost never marked.

Most of the languages have some special features which make the development of speech either easier or difficult. Letter to sound rules are also differ from language to language. There is a lot of work has been done already on these parameters to improve the synthesized speech output. For some of the synthesizers, the output is quite good but the naturalness of the speech still needs to be improved. The synthesized speech sounds more like machine-generated than human-like. This may be irritating after a point for a person who is using TTS for reading a book or information from the internet. The proposed work is focused on the improvement of the naturalness of Hindi TTS output.

4. CONCLUSION

Speech Synthesis systems are used in various applications. These systems can be useful as an assistant to visually impaired person and are currently used in many educational institutes as a learning machine for kids. Speech Synthesis and recognition techniques are also used by many research companies and web browsers. Using Festival Framework speech synthesizer for many languages are developed. Here we are mainly working on the Hindi language. We are considering a few parameters to improve the naturalness of the utterance created. The parameter like schwa deletion [4] can be considered to make the synthesized voice more human-like. To include these parameters for the natural sound schema files are created.

REFERENCES

[1] G. D.Ramteke, R. J.Ramteke, “Hindi Spoken Signals for Speech Synthesizer” , 2nd International Conference on Next Generation Computing Technologies (NGCT-2016), 2016.

[2] Shilpi Kannojia, Ghanpriya singh, Dr.Sanjay Mathur, “A Text to Speech synthesizer using acoustic unit based concatenation for any Indian Language of Devanagari Script”, 11th international conference on Indudtrial and Information System, 2016.

[3] Somnath Roy, “A Technical Guide to Concatenative Speech Synthesis for Hindi using Festival”, International Journal of Computer Applications (0975 – 8887) Volume 86 – No 8, January 2014.

[4] Kalika Bali, Partha Pratim Talukdar N. Sridhar Krishna, A.G. Ramakrishnan, “Tools for the development of a hindi speech synthesis system ”.