International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Review of various Text to Speech Synthesis Methodologies

Sheilly Padda [1], Sonali Gupta [2], Amandeep Kaur [3]

[1-3]*Assistant Professor, Department of Computer Science & Engg, Chandigarh Engineering College, Landran, Mohali, India*

Abstract:- *There are about millions of blind and visually impaired people worldwide. These people are not able to see and hear anything. Disability of visual text reading has a huge impact on the quality of life for visually disabled people. Though many several devices have been designed for helping visually disabled to see objects using an alternating sense such as sound and touch, the development of text reading device is still at an early stage. Some existing systems for text recognition are typically limited either or require user assistance or may be of high cost. Therefore some low cost system must be developed that will enable to automatically locate and read the text aloud to visually impaired persons. In this paper we will study various ideas to recognize the text character and convert it into speech signal along with some applications of TTS systems. This paper also contains the various terms and concepts of text to speech conversion systems.*

Keywords: **Linguistics, Natural language Processing, Phonetic, Tokenization, Transcription.**

Introduction

In general in any text to speech convertor system the text is first pre-processed. The pre-processing module prepares the text for recognition. Then the text is segmented to separate the character from each other. Segmentation is followed by extraction of letters and resizing them and stores them in the text file. These processes are done with the help of MATLAB. This text is then converted into speech. [1]

Processing of Text in TTS.

The text to speech synthesizer is a computer based system that reads the text. The following Fig1 briefs the steps involved in processing a text to be converted to speech as the output[2].

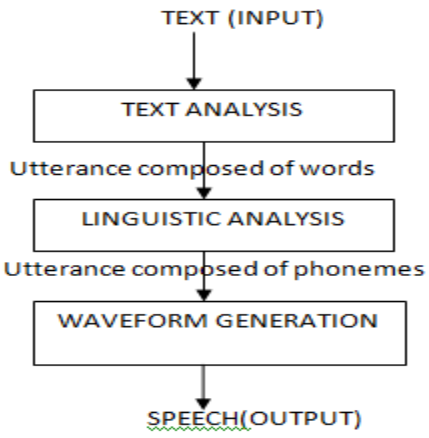

Fig 1: Overall block diagram of TTS

A TTS system if made up of two parts: First part called as frontend which converts raw text like symbols, numbers, and abbreviations into written out words. This process is also called as normalization of text or preprocessing or tokenization [3].The front also marks the text into prosodic units like phrases, clauses and sentences. The process of assigning phonetic transcription to words is known as text to phoneme conversion. The Phonetic transcription is the use of phonetic symbols to represent the sound symbols. The second part called the backend converts the symbolic linguistics specifications into sound. Fig 2 explains the flow of the text to speech module which is explained in detail further below.

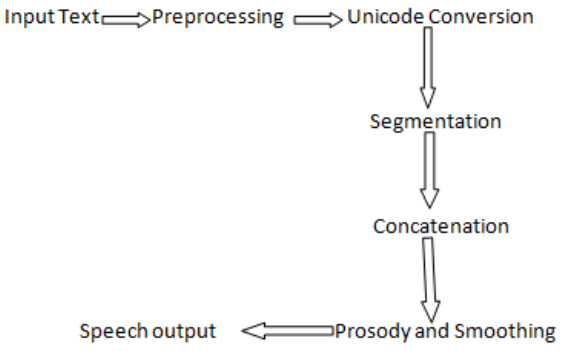

Fig 2: Text to speech problem formulation

Text to speech conversion can be accomplished by starting with the method of pre-processing of the input text. Here the text abbreviations, acronyms and numbers are expanded. The pre-processed text will then be converted to Unicode. Unicode has the explicit aim of transcending the limitations of encodings. Here the pre-processed text is used to identify the fonts of input text and is converted to

Unicode. Now, the encoded text is segmented into syllables and the duplicates are removed.

Text to Speech conversion using OCR

In this type of TTS, Text-to-speech device consists of two main modules, the image processing module and voice processing modules. Image processing module captures image using camera, converting the image into text. Voice processing module changes the text into sound and processes it with specific physical characteristics so that the sound can be understood [5]. Fig 3 below shows the block diagram of Text-To-Speech device, 1st block is image processing module, where OCR converts .jpg to .txt form. 2nd is voice processing module which converts .txt to speech

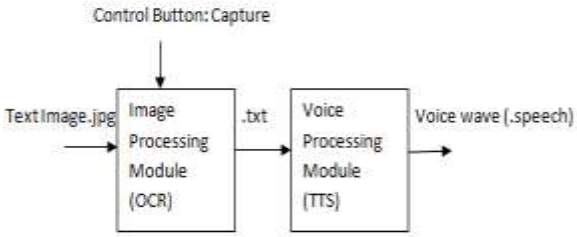

Fig 3: Block diagram of text-to-speech device using OCR

NLP and Speech Synthesis Methodology

Text to speech system also contains two parts namely natural language processing and speech synthesis (digital signal processing)which are explained as below[6].

Natural Language Processing (NLP)

NLP produces phonetic transcript together with prosodic (study of sound) feature of the input text. NLP comprises of three main mechanisms: text analysis, phonetic conversion and prosodic phrasing.

1. **Text study:** In this TTS system, the input sentence is segmented into token and then after the tokenization, each word is specified as part of speech (POS). Part-of-speech is a process assigning correct POS label to each word in a sentence from a given set of tags [7].This provides the transition between the tags and helps capture the context of the sentence.

2. **Phonetic translation**: In this system, Dictionary based approach is used for phonetic transcription of input word. So, the words or any input text that is not included in the dictionary cannot run.

3. **Prosodic Phrasing** : Prosodic Phrasing is to assign the idiom of the input text .The words are grouped together into phrases by their musical and duration properties as well as their tone frequency. This 'prosodic phrasing' affects the perceptive of sentences.

Grapheme to phoneme conversion system

Grapheme to phoneme conversion systems plays the role of an essential component in text-to-speech (TTS) systems. These modules operate before the phone sequence is fed into the synthesis routine. Although such systems have emerged with additional challenges as when such conversion modules are implemented with non-inhabitant varieties of languages, such as Indian English. Many existing grapheme to phoneme dictionaries represent American English intonation, but they are not suitable for use in Indian English TTS systems. Therefore, the work has been carried out to modify the existing English grapheme to phoneme dictionary by implementing specific rules for one particular variety of Indian English, namely Assamese English [8].

Modules involved in TTS system

Speech synthesis

A Speech synthesizer is an application system that reads any text, when it is given as input to the computer by the user. It is best suited to TTS or speech synthesis as an automated creation of speech by grapheme to phoneme transcription. A grapheme is the smallest contrastive unit in a written language. It does not carry meaning by itself. Graphemes is composed of alphabets letters, numerical digits, punctuation marks, and the individual symbols of any of the database for writing systems. A phoneme is "the smallest sound producing segment which denotes some meaning to utterances"

Phonetics

Mostly the languages do not have the pronunciation of its written text. To represent the correct and valid pronunciation we require some symbolic representation for the written text. The phonetic is termed as the written symbolic representation of its pronunciation. Every language has a different phonetic alphabet and a different set of possible phonemes and their combinations. The number of phonetic symbols is between 20 and 60 in each language. "A set of phonemes can be defined as the minimum number of symbols needed to describe every possible word in a language". In English there are about 44 phonemes. Due to complexity and different kind of definitions, the number of phonemes in English and most of the other languages cannot be defined exactly.

Phonemes are known to be abstract units and whose pronunciation depends on relative effects, speaker's characteristics, and emotions. During speech, the articulator movements depend on the former and the subsequent phonemes. This causes some variations on how the individual phoneme is pronounced. Another cause for the imperfection of phonetic representation is that speech signal is always continuous and phonetic details are always distinct. The phonetic alphabet consists of vowels and consonants. Vowels are always voiced sounds and they are produced with the vocal cords in vibration and have considerably higher amplitude, while consonants may be either voiced or unvoiced. Vowels are also more stable and easier to analyze and describe acoustically but consonants involve very rapid changes they are more difficult to synthesize properly [9].

Applications of Text-to-Speech System

The application field of TTS is expanding fast whilst the quality of TTS systems is also increasing steadily. Speech synthesis systems are also becoming more reasonable for common customers, which makes these systems more suitable for everyday use[10]. Some applications of TTS are described below.

1) Aid to Vocally Handicapped in the form of a hand-held, battery-powered synthetic speech aid system.
2) Source of Learning for Visually Impaired Listening.
3) Talking Books and Toys Talking book.
4) Games and Education.

5) Telecommunication and Multimedia.

6) Man-Machine Communication. For example, in warning, alarm systems, clocks and washing machines.

7) Voice Enabled E-mail. Voice-enabled e-mail uses voice recognition and speech synthesis technologies.

General Structure of Text-to-Speech System

The Text to Speech Synthesis mechanism involves various steps. As an input, it gets text which is analyzed and converted to its phonetic specifications. Further prosody is generated I and then finally the speech signal. This method of speech generation is further broken into its major modules:

Natural Language Processing (NLP) module: It produces a phonetic transcription of the text read, together with prosody.

Digital Signal Processing module: It transforms the figurative information it receives from NLP into audible and intelligible speech.

Lets see what are the major functions of the NLP module as shown in Fig 4 [11]:

Text Analysis: First the text is segmented into tokens. The token-to-word conversion creates the orthographic form of the token. Lets take an example: in the token “Mr” the orthographic form “Mister” is formed by expansion and “1997” is transformed to “nineteen ninety seven”.

Application of articulation Rules: After the text analysis has been completed, pronunciation rules can be applied. Letters cannot be transformed 1:1 into phonemes because correspondence is not always parallel. In certain environments, a single letter can correspond to either no phoneme (for example, “h” in “caught”) or several phoneme (“m” in “Maximum”). In addition, several letters can correspond to a single phoneme (“ch” in “rich”).

The two applications (rule based and dictionary based for the pronunciation) differ in terms of their sizes. The dictionary-based solution is many times larger than the rules-based solution’s dictionary of exception. However, dictionary-based solutions can be more exact than rule-based solution if they have a large enough phonetic dictionary available.

Prosody Generation: After the pronunciation has been specified, the prosody is generated. The degree of naturalness of a TTS system is dependent on prosodic factors like intonation modelling, amplitude modelling and duration modelling (including the duration of sound and the duration of pauses, which determines the length of the syllable and the tempos of the speech.

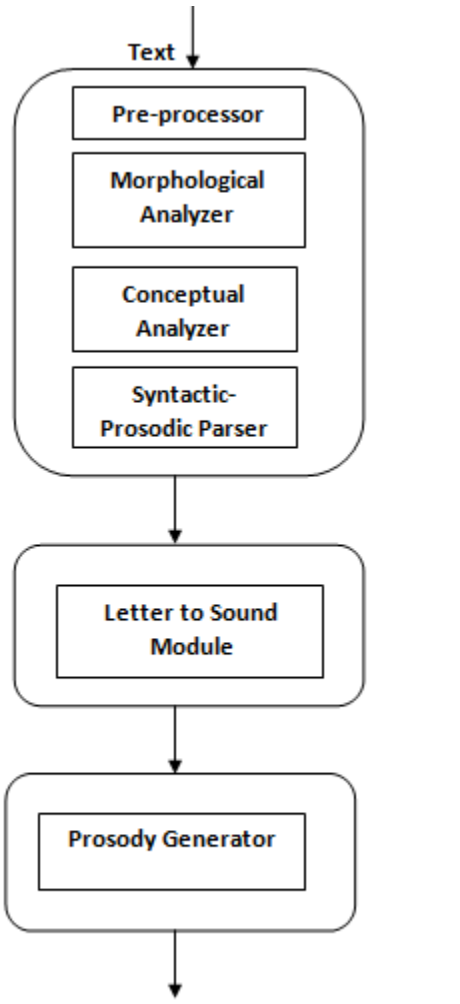

Fig 4: Functions of NLP module of TTS System

The output of the NLP module is passed to the DSP module[12]. This is where the actual synthesis of the speech signal happens. In concatenative synthesis the selection and linking of speech segments take place. For individual sounds the best option (where several appropriate options are available) are selected from a database and concatenated.

Conclusion

Text-to-Speech device can convert the text input into sound with a performance that is high enough and a readability tolerance, with the less average time of processing These types of portable devices, does not require internet connection, and can be used independently by people even who are physically impaired. Through the different technologies and method, we can make editing process of various books or web pages easier. In this paper, we have discussed various methodologies by which a general text can be converted to speech.

References

1. Text to Speech Conversion System using OCR Jisha Gopinath, Aravind S, Pooja Chandran, Saranya S S, www.ijetae.com,Certified Journal, Volume 5, Issue 1, January 2015
2. http://www.voicerss.org/tts/
3. http://www.comsys.net/technology/speechframe/text-to-speech-tts.h
4. TEXT TO SPEECH CONVERSION MODULE Hussain Rangoonwala1, Vishal Kaushik , P Mohith and Dhanalakshmi Samiappan International Journal of Pure and Applied Mathematics Volume 115 No. 6 2017, 389-395.
5. Text to Speech Conversion S. Venkateswarlu1,D.B.K. Kamesh , J. K. R. Sastry and Radhika Rani Indian Journal of Science and Technology, October 2016
6. Text To Speech Conversion Using Different Speech Synthesis Hay Mar Htun, Theingi Zin, Hla Myo Tun INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 4, ISSUE 07, JULY 2015
7. Sumeer Mittal , MrNavdeep Singh Sethi and Sanjeev KumarSharma,—Partof Speech Tagging of Punjabi Language using N Gram Model‖,International Journal of Computer Applications (0975 – 8887) ,Volume 100– No.19, August 2014.
8. 2016 IEEE Region 10 Conference (TENCON) 09 February 2017 Singapore, Singapore
9. International Journal of Innovative Research in Computer and Communication Engineering Implementation of Text to Speech Conversion Technique Mohd Bilal Ganai , Er jyoti Arora
10. A Review: Translation of Text to Speech Conversion for Hindi Language Kaveri Kamble1 , Ramesh Kagalkar2 International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064
11. Design and Implementation of Text To Speech Conversion for Visually Impaired People, Itunuoluwa Isewon et. al , International Journal of Applied Information Systems (IJAIS), Foundation of Computer Science FCS, New York, USA Volume 7– No. 2, April 2014.
12. Text-to-speech technology: In Linguatec Language Technology Website. Retrieved February 21, 2014, from http://www.linguatec.net/products/tts/information/techno logy