Review on Text to Speech Synthesizer

Satish K. Dutonde[1], Gulab S. Mapari[2], Sagar J. Wagh[3], Dr. Avinash S. Kapse[4]

[1,2,3]*Student, IT Dept., Anuradha Engineering College, Maharashtra, India*

[4]*Head of Dept., IT Dept., Anuradha Engineering College, Maharashtra, India*

ABSTRACT

In this paper we try to explore the various ways of compiling Text-to-speech that have been developed and used a few keyword research researchers and research groups around the world over the past decades. Text-to-Speech Synthesizer software and Hardware are discussed with its features. The most popular smart phone today has it the ability to read text and e-book aloud. Continuous research continues with fluency with the help of one of the important methods i.e., statistical parametric approach to speech synthesis. The main aim of this review paper is to give an overview of the integration of speech into Indian languages, summarizes and compares features of various integration techniques used.

Keywords:- *Text processing, Text-To-Speech (TTS) synthesizer, Speech Enhancement.*

1. INTRODUCTION

Digital speech processing plays an important role in modernity communication research and applications. I the main purpose of speech is communication; it means the transmission of a message between a person and one machine. Text-to-speech system (TTS) converts text into word using speech synthesizer [1]. To be done production of human speech. Computer system used this purpose is called the speech synthesizer, and that is used in both software and hardware form as ARM7 microcontroller that converts Text into Speech as well Speech in Scripture [2]. Text-to-speech (TTS) system. translates standard language text into US & UK English sayings. This synthetic expression cannot be understood by a person with moderate communication skills in English language.

The processes of the text-to-speech system are very different from the production of live human speech. Live human speech production depends on the mechanics of complex liquids depends on changes in lung function and tone of voice lashes [3]. The purpose of the text goes to the speech system is to convert a given text carelessly into a coherent whole spoken waveform. Text processing and speech production the two main parts of the text in the speech system. I The main purpose of the text processing component is to process inserted text and generates the correct sequence of phonemic units. These phonetic units are available by part of speech production or in combination from parameters or by unit selection in large speech corpus [4]. In order to put together a clear voice, of course it is important that the textual processing component produces a proper sequence of corresponding sound units input text negligently. Figure 1 shows the text block diagram in the speech system.

2. TTS SYSTEM MAIN PHASES

2.1 Text processing

A text-to-speech system the input text is first analysed, normalized and transcribed into a phonetic or some other linguistic representation.

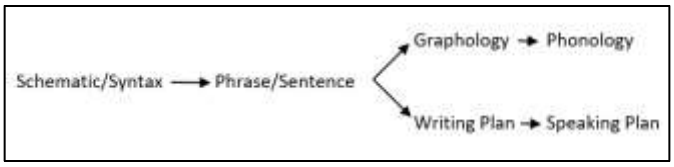

Fig-1 Text Processing Components [5]

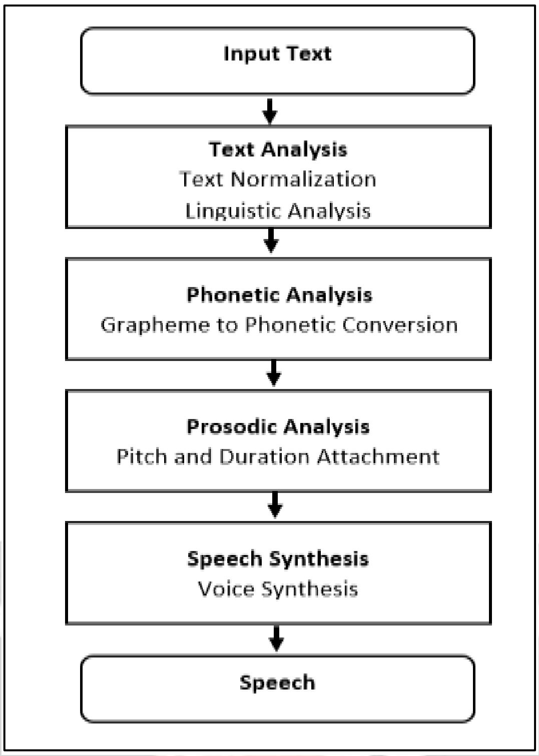

Fig-2 Block diagram of Text to Speech Synthesis

Parts of the text processing are about low quality to process issues such as the separation of sentences and words separation [1] [5] [6].

- Document the structure can be identified by translating the punctuation mark and paragraph formatting.
- Text familiarity handles abbreviations and acronyms. The goal of accustomed to match text e.g., Dr. can be offered as a doctor. The normal routine makes it great output.
- Language analysis includes morphological analysis of proper pronunciation and syntactic analysis to facilitate pronunciation and naming handling incomprehensible things in the text [1] [6].

2.2. Speech generation

The speech generation component processes for generate the speech by using parameters as

- Phonetic analysis focuses on the phone level within each word. Each phone is tagged with information on how to produce a sound and what kind of sound to produce it mean style and emphasis. Grapheme to phoneme conversion Exact pronunciation of all word of the input sentences is determined. Homograph disambiguation Figuring out whether input sentence use the past or present tense version of the word. identifying the tense system of a word is depend on the dictionary [5].
- Prosodic analysis can Analysis of prosody is very important. Because it provides the basis for marking the prosodic effect around our utterance plans i.e., phonological prosodic processing and after to arrive at suitable rendering strategies for the marked prosody i.e., phonetic prosodic processing. There are two approaches available in the prosody [7]. Create an abstract descriptive system which characterizes observations of the behaviour of the parameters of prosody within the acoustic signal and promote the system to a symbolic phonological role. Create a phonological system for input process which eventually result in an acoustic signal jugged by listeners to have a proper prosody.

3. SPEECH SYNTHESIS TECHNIQUES

Symbolic prosody data is used by the synthesizer in produce a speech that uses a certain method. Three key stages of Speech Synthesis Techniques.

3.1 Articulator synthesis

Articulator synthesis aims at computer simulation Neurophysiology and biometrics of speech production. Articulator synthesis uses a mechanical and acoustic model of speech production to integrate speech. This combination produces understandable synthetic speech, but its effect remains the same away from the noise of nature [1] [5].

3.2 Formant integration

In this program the representation of each speech parts is stored on a perimeter basis. Parameter are those of the minimum formant synthesizer for each Holmes part. There is one value for each parameter. This means one expression of the acoustic component [1].

3.3 Concatenative synthesis

Concatenative synthesis is a form of consolidation sound by Concatenative samples of recorded audio called units. The duration of the units is not strictly defined and is possible they vary depending on the implementation, approximately the grade from 10 ms to 10 seconds. Used in compound speech generate user-specific sequences on the website built on recording in another sequence [5].

Concatenative synthesis units

- The phone is a single unit of audio. Speech it is a sequence of such sounds.
- A diphone is defined as a signal from or the centre of the phone or the area of small change inside the phone went to the same location on the next phone.
- Triphone is part of the signal to take in sequence from the middle of the phone completely next to the middle of the third [5].

4. DEVELOPMENT OF TEXT TO SPEECH SYNTHESIZER

Synthetic speech has been a dream of mankind for centuries. Understanding how the systems represent and how development has begun in the current form. This review may provide new researchers with information on further processing. In this paper, the history of integrated speech from the earliest machine experiments to programs that formed the basis of modern synthesizers is discussed.

Producing speech made into curiosity 100 years ago. In those years Gerert of Aurillac invented the first known speech machine. For the next two centuries, inventors such as Albertus Magnus and Roger Bacon built machines known as “talking heads” [7]. However, the first known device to attempt to mimic real human speech was invented by Christian Kratzenstein of St. Petersburg in 1779. This machine can produce five long vowel sounds. Twelve years later, Wolfgang Von Kempelen developed a machine that could produce some vowels and consonants [8] [9].

The first complete TTS program was launched in the late 1960s. Since then, there have been many improvements in the accuracy and quality of TTS systems. Companies such as IBM, Microsoft, and Bell Labs have developed free and commercially available systems.

The following are some Text-to-Speech synthesizer products.

- **MITalk**

  In 1976, Allen, Hunnicutt, and Klatt founded the MITalk MITalk in English. The TTS used different levels to translate text into integrated speech. At first level, abbreviations, numbers, and symbols were converted into words. Then, with the help of 12,000 morph (beginning, roots, and suffixes) dictionaries, the words are changed to phonetic. Non-dictionary words are converted into phonemes by rules [10].

- **DEC Talk**

  Digital Equipment Corporation Talk was based on the Klattalk program available in USA English and Spanish. The DEC Talk system was later commercially available in 1983. The system is able to specify multiple correct words, email and URL addresses and supports a custom pronunciation dictionary. It also controls punctuation marks, height, and pressure and voice control commands are inserted into a text file used by DEC speaking software applications. Speech rate is adjustable between 75 and 650 words per minute [11].

- **Festival**

The festival has many TTS languages currently available in British and American English, Welsh and Spanish. The program was created by Alan Black and Paul Taylor. The program is written in C ++ and supports the remaining LPC and PSOLA modes and the MBROLA website. In the LPC method, the residues and coefficient of LPC are used as control parameters. It uses audio-recording rules and a large lexicon of TTS conversion. Speech integration is done using the optional unit to integrate diphones. The festival provides a common framework for building programs to integrate speech [12].

- **AT&T VOICEBUILDER**

AT&T VOICEBUILDER only supports the English language. It provides a new tool for researchers and staff who want to make their voices integrated into a high-quality text-to-speech marketing program without the need to install, edit, or manage speech processing software and tools. Used as a web service in the AT&T Speech Mash up Portal. The program records and verifies users' comments, processes them to create a voice for action and provides a web service API to make voice available to real-time applications using a cloud-based processing platform. All procedures are automatic to avoid human intervention [4].

- **Text-to-Speech Program in Indian Languages**

IIT Hyderabad has developed a comprehensive framework for Hindi and Telugu languages to produce text processing modules and language resources that can be expanded to all Indian languages with minimal effort and time. Anand Arokia et al. how to apply basic language knowledge, acoustic data and machine learning techniques. Their efforts in this regard are supportive especially font identification, font-to-Akshara fonts, and Aksharas naming rules and customization [13].

- **WHISTLER**

Whistler is a text to Speech training program available in English that automatic learning model parameters on a corpus. The speech engine is based on Concatenative synthesis and the training process in Hidden Markov models. The list of word-for-word compilation unit is automatically generated from the label less web site using the Whisper speech recognition system. The speech recognition unit detects speech and automatically labels parts of speech. This approach not only improved the natural environment but also reduced the time required to create a new voice and made the synthetic speech resemble that of the first speaker [14].

- **Mobile-based TTS**

Now many day-to-day mobile phone manufacturers offer text-to-speech space. It is very useful for the visually impaired person; they can hear text from a cell phone screen and even read e-books. Google has launched Android-based TTS mobile in English, Spanish, French and Italian etc. [15]. SVOX corporation has launched its TTS-based Android phone. Can read English text, e-book and can translate speech into another language [16]. C-DAC Mumbai has developed TTS based on Android phone in Marathi and Oriya languages [17].

- **Conversion of text to speech using OCR**

This system consists of a portable camera, computer device and speaker. First read the picture text, pay attention to the letters, numbers and symbols. The reconstructed text is converted into speech [18].

5. SOME WAYS TO COMBIN TEXT-TO-SPEECH

In the selection of a unit based on Concatenative expression synthesis, the shared cost is also known as the Concatenative cost, which measures how well the two units can be combined. After the units are assembled, most of the system tests three times the cost of operation and three slide modes such as No slide, line slide and Kalman filter-based slider [19]. T. Dutoit has shown that Line Spectral Frequencies (LSF) has better translation features and produces a smoother transition than LPC parameters [20].

Conversation can be enhanced by using the Kalman filter visual interface filter combined with the standard Kalman filter, providing excellent performance [21]. The text-to-speech translation system produces intermediate speech, which can be converted into a heart-to-heart speech by adjusting the voice count (F0) of accented words using the usual Gaussian method [22] [23] [24]. converting the spectrum from natural to expressive speech there are two HHMs used [25] [26]. Explicit storytelling of the storytelling app is generated using a set of prosodic rules for converting a neutral speech generated by the TTS program into a storytelling dialogue with Pitch, Intensity, Tempo and Duration. Many acoustic models are often integrated into a combination of parameter mathematical expression [27]. The combination of multiple HMM and Gaussian acoustic models provides significant improvements in integrated speech quality [28].

6. CONCLUSION

The test-to-speech synthesizer is gradually evolved from a few decades ago to find the current situation. Three basic Ways to combine speech are Articulator, Formant and Concatenative synthesis used in various synthesizers. Many new applications are being upgraded, but they are Understandable and the comprehension of passive speech has not yet been reached acceptable level. Even in India some research organizations also work on translating text into speech in regional languages such as Marathi, Hindi, Telugu, Punjabi, Kannada etc. But all these programs do not repeat natural human speech. There is a great width in between integration development to achieve this high quality natural and emotional aspect.

7. ACKNOWLEDGMENT

We would like to sincerely acknowledge the uncourageous efforts of Information Technology Department of Anuradha Engineering College, Chikhli. Our heartfelt thanks to faculty members who helped us in prepare review paper and give direction with their precious suggestions & rich experience.

8. REFERENCES

[1]. Archana Balyan, S.S. Agrwal and Amita Dev, Speech Synthesis: Review, IJERT, ISSN 2278-0181 Vol. 2 (2013) p. 57 – 75.

[2]. D.D. Pande, M. Praveen Kumar, A Smart Device for People with Disabilities using ARM7, IJERT, ISSN 2278-0181 Vol.3(2014) p. 614 – 618.

[3]. J.O. Onaolap, F.E. Idachaba, J. Badejo, T. Odu and O.I. Adu, in Proc. of the World Congress on Engineering, (London, UK. 2014).

[4]. Alistair Conkie, Thomas Okken, Yeon-Jun Kim, Giuseppe Di Fabbrizio, Building Text-To-Speech Voices in the Cloud, in Proc. AT&T Labs Research, Park Avenue, Florham Park, NJ- USA).

[5]. Mark Tatham and Katherine Morton, Developments in Speech Synthesis (John Wiley & Sons, Ltd. ISBN: 0-470-85538-X, 2005).

[6]. AIndumati and Dr. E. Chandra, Speech processing –An Overview, Int. J. of Engg. Sci. and Tech., Vol. 4, (2012) p. 2853-2860.

[7]. Mattingly I. G., Speech Synthesis for Phonetic and Phonological Models, T.A. Sebeok (Ed.) Current Trends in Linguistics, Vol. 12, (1974) p. 2451-2487.

[8]. Klatt Dennis, Review of Text-to-Speech Conversion for English, J. of the Acoustical Soc. of America, Vol. 3, (1987) p. 737-793.

[9]. Schroeder M., A Brief History of Synthetic Speech, J. Speech Communication, Vol. 13, (1993) p. 231-237.

[10]. Allen, John, Hunnicutt, Sharon, and Dennis Klatt, Text to Speech, The MITTALK System (Cambridge: Cambridge University Press, 1987).

[11]. Sami Lemmetty, Review of Speech Synthesis Technology (Helsinki University of Technology, 1999).

[12]. Black A. and P. Taylor. The Festival Speech Synthesis System: system documentation, (Human Commu. Research Centre, Uni. of Edinburgh, Scotland 1997).

[13]. Anand Arokia et al., Text Processing for Text-to-Speech Systems in Indian Languages, Proc. in 36th ISCA Workshop on Speech Synthesis, (Bonn, Germany, August 2007) pp.22-24.

[14]. Xuedong Huang, Alex Acero, Jim Adcock, Hsiao-Wuen Hon, John Goldsmith, Jingsong Liu and Mike Plumpe, WHISTLER: A TRAINABLE TEXT-TO-SPEECH SYSTEM, in Proc. Microsoft Corporation, (Washington USA).

[15]. [Online].Available:(2015)http://www.greenbolt.com/article/210586 2/how-to-get-started-with-google-text-to-speech.html.

[16]. [Online].Available:(2015)https://svoxbilevoices.wordpress.com/pag e/2.

[17]. [Online].Available:(2015)http://cdac.in/index.aspx?id=mcst_speech _technology.

[18]. Jisha Gopinath, Aravind S, Pooja Chandran and Saranya S.S., Text to Speech Conversion System using OCR, IJETAE, ISSN 2250- 2459, Vol.5 (2015) pp.389-395.

[19]. Jithendra Vepa and Simon King, Subjective Evaluation of Join Cost and Smoothing Methods for Unit Selection Speech Synthesis, IEEE Trans. Speech Audio Process, Vol. 14, (2006) pp. 1763 – 1771.

[20]. T. Dutoit, An Introduction to Text-to-Speech Synthesis, Kluwer Academic Publishers, Dordrecht, ISBN 0-7923-4498-7, (1997).

[21]. Ning Ma. and Rafik A. Goubran, Speech Enhancement Using a Masking Threshold Constrained Kalman Filter and Its Heuristic Implementations, IEEE Trans. Speech Audio Process, Vol. 14, (2006) pp.19-32.

[22]. Marc Schröder, Expressing Degree of Activation in Synthetic Speech, IEEE Trans. Speech Audio Process, Vol. 14, (2006) pp. 1128 – 1136.

[23]. Rohit Deo and Pallavi Deshpande, Neutral to Emotional Speech Conversion by Pitch Counter Modification for Marathi, IJERT, ISSN 2278-0181 Vol. 3 (2014) pp. 2228-2231.

[24]. Kai Yu and Steve Young, Continuous F0 Modeling for HMM Based Statistical Parametric Speech Synthesis, IEEE Trans. Speech Audio Process, Vol. 19, (2011) pp. 1071 – 1079.

[25]. Chung-Hsien Wu, Chi-Chun Hsia, Te-Hsien Liu, and Jhing-Fa Wang, Voice Conversion Using Duration-Embedded Bi-HMMs for Expressive Speech Synthesis, IEEE Trans. Speech Audio Process, Vol. 14, (2006) pp. 1109 – 1116.

[26]. Zhen-Hua Ling, Korin Richmond, Junichi Yamagishi, and Ren-Hua Wang, Integrating Articulatory Features Into HMM-Based Parametric Speech Synthesis, IEEE Trans. Speech Audio Process, Vol. 17, (2009) pp. 1171 – 1185.

[27]. Mariët Theune, Koen Meeks, Dirk Heylen, and Roeland Ordelman, Generating Expressive Speech for Storytelling Applications, IEEE Trans. Speech Audio Process, Vol. 14, (2006) pp. 1137 – 1144.

[28]. Heiga Zen, Mark J. F. Gales, Yoshihiko Nankaku, and Keiichi Tokuda, Product of Experts for Statistical Parametric Speech Synthesis, IEEE Trans. Speech Audio Process,Vol. 20, (2012) pp. 794 – 805.