ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Review on Text-To-Speech Synthesizer

Suhas R. Mache [1], Manasi R. Baheti [2], C. Namrata Mahender[3]

Research Fellow, Dept. Of Computer Science & IT, Dr. B.A.M. University Aurangabad, India [1]

Asst. Professor, Dept. Of Computer Science & IT, Dr. B.A.M. University Aurangabad, India [2,3]

Abstract: In this paper we try to explore various Text-to-Speech synthesis techniques developed and implemented by several speech synthesis researchers and research groups over the world during the last decades. Few Text-to-Speech synthesizer software and hardware's are mentioned with its features. Today's most popular device smart phone has capabilities to read a text and e-book aloud. A continues research is going on speech smoother with the help of one of the key approaches i.e. statistical parametric approach to speech synthesis. This paper aims to give an overview of speech synthesis in Indian languages, summarizes and compares the characteristics of various synthesis techniques used.

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

I. INTRODUCTION

Digital speech processing plays a vital role in modern speech communication research and applications. The fundamental purpose of speech is communication; it means transmission of message between human and machine. Text to speech system (TTS) converts text into voice using a speech synthesizer [1]. It is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and that is implemented in both software and hardware form like ARM7 microcontroller which converts Text to Speech and Speech to Text [2]. A text-to-speech (TTS) system converts normal language text into US & UK English accents. This synthetic speech cannot be understood by a person with average communication skill in English language.

Text to speech system processes are significantly different from live human speech production. Live human speech production depends on complex fluid mechanics dependent on changes in lung pressure and vocal tract constrictions [3]. The objective of a text to speech system is to convert an arbitrary given text into a corresponding spoken waveform. Text processing and speech generation are two main components of a text to speech system. The objective of the text processing component is to process the given input text and produce appropriate sequence of phonemic units. These phonemic units are realized by the speech generation component either by synthesis from parameters or by selection of a unit from a large speech corpus [4]. For natural sounding speech synthesis, it is essential that the text processing component produce an appropriate sequence of phonemic units corresponding to an arbitrary input text. Fig. 1 shows block diagram of Text to Speech system.

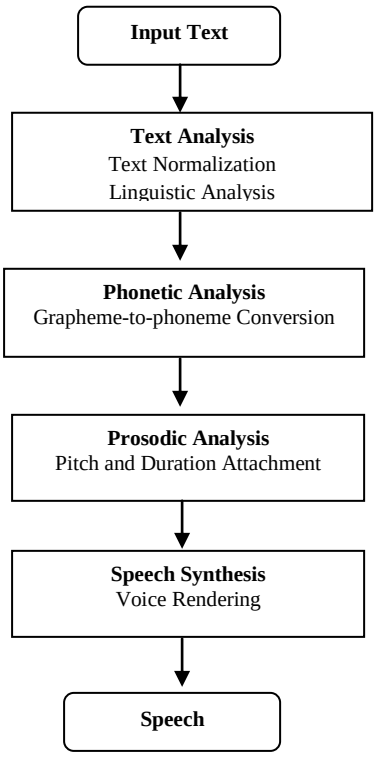

Fig.1 Block diagram of Text-to-Speech Synthesis

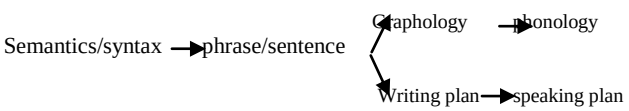

Fig. 2 Text processing components [5]

II. TTS SYSTEM'S MAIN PHASES

1. Text processing
2. Speech generation

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

Text processing components deals with low level processing issues such as sentence segmentation and word segmentation [1][5][6].

a) Document structure detection - The document structure can be detected by interpreting punctuation mark and paragraph formatting.

DOI 10.17148/IJARCCE.2015.4812

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

b) Text normalization - The text normalization handles abbreviation and acronyms. The goal of normalization is to match the text e.g. Dr could be rendered as doctor. Proper normalization makes the good output.

c) Linguistic analysis - Linguistic analysis includes a morphological analysis for proper word pronunciation and syntactic analysis to facilitate accenting and phrasing to handle ambiguities in written text [1] [6].

Speech generation:

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

a) Phonetic analysis- It focuses on the phone level within each word. Each phone is tagged with information about what sound to produce and how to produce it means style and emphasis.

- Grapheme to phoneme conversion: Exact pronunciation of each word of the input sentences is determined.
- Homograph disambiguation: Figuring out whether input sentence use the present tense or past tense version of the word. To identify a word tense system depends on dictionary [5].

b) Prosodic analysis - The analysis of prosody is important because it gives basis for marking prosodic effect around our utterance plans i.e. phonological prosodic processing and later to arrive at suitable rendering strategies for the marked prosody i.e. phonetic prosodic processing. There are two approaches in the prosody [7].

- Create an abstract descriptive system which characterizes observations of the behavior of the parameters of prosody within the acoustic signal (fundamental frequency movement, intensity changes and duration movement) and promote the system to a symbolic phonological role.
- Create a phonological system which can be used to input to process which eventually result in an acoustic signal jugged by listeners to have a proper prosody.

III. SPEECH SYNTHESIS TECHNIQUES

Symbolic prosody data is used by the synthesizer to generate speech using a specific method. There are three main categories of speech synthesis techniques.

i. Articulator synthesis
ii. Formant synthesis
iii. Concatenative synthesis

i. Articulator synthesis

Articulator synthesis aims to simulate computationally the neurophysiology and biometrics of speech production. Articulator synthesis uses mechanical and acoustic model of speech production to synthesize speech.This synthesis produce intelligible synthetic speech, but its output is still far from natural sound [1] [5].

ii. Formant synthesis

In this system representation of individual speech segments are stored on a parametric basis. The parametric are those of low level Holmes formant synthesizer for each segment. There is a single value for each parameter. This means single acoustic segment speech [1].

iii. Concatenative synthesis

Concatenative synthesis is a technique for synthesizing sound by Concatenative samples of recorded sound called units. The duration of units is not strictly defined and may vary according to the implementation, roughly in the range of 10 ms up to 10 seconds. It is used in speech synthesis to generate user specific sequence of sound from a database built from recording of other sequences [5].

Units for Concatenative synthesis

a. Phone: Phone is single unit of sound. The speech is a sequence of such sounds.

b. Diphone: A diphone is defined as the signal from either midpoint of a phone or point of least change within the phone to the similar point in the next phone.

c. Triphone: A Triphone is a section of the signal taking in a sequence going from middle of a phone completely through the next one to the middle of a third [5].

IV. DEVELOPMENT OF TEXT-TO-SPEECH SYNTHESIZER

Artificial speech has been a dream of the humankind for centuries. To understand how they represent systems and how the development has taken shape in the present form. This review may provide new researchers with insights for further processing. In this paper, the history of synthesized speech from the first mechanical efforts to systems that form the basis for today's high-quality synthesizers is discussed.

Generating synthetic speech has been a curiosity for the past 100 years. Around those years Gerbert of Aurillac created the first known mechanical talking machine. For the next two centuries, inventors like Albertus Magnus and Roger Bacon created machines know as “talking heads” [7]. However, the first known machine that tried to mimic real human speech was developed by Christian Kratzenstein of St. Petersburg in1779. This machine could produce five long vowel sounds. Twelve years later, Wolfgang Von Kempelen developed a machine that could produce vowels and some consonants [8] [9].

The first full TTS system was developed in the late 1960’s. Since then, there have been many advances in the accuracy and the quality of TTS systems. Companies like IBM, Microsoft, and Bell Labs have developed both free and commercially available systems. Following are the some products of Text-to-Speech synthesizer.

A. MITalk

In 1976, Allen, Hunnicutt, and Klatt developed the MITalk at MIT available in English language. This TTS used different levels to convert text to synthesized speech. In the first level, abbreviations, numbers, and symbols were transformed into words. Then, using a 12,000 morph (prefixes, roots, and suffixes) lexicon, words were converted to their phonetic equivalent. Words not in the lexicon were converted to phonemes by using rules [10].

B. DEC Talk

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Digital Equipment Corporation DEC Talk was based on Klattalk system it is available in American English, German and Spanish. The DEC Talk system later became commercially available in 1983. The system is capable to say most proper names, e-mail and URL addresses and supports a customized pronunciation dictionary. It has also punctuation control for pauses, pitch, and stress and the voice control commands are inserted in a text file which is used by DEC talk software applications. The speaking rate is adjustable between 75 to 650 words per minute [11].

C. Festival

The Festival is multi-lingual TTS currently available in English British and American, Welsh and Spanish. This system is developed by Alan Black and Paul Taylor. The system is written in C++ and supports residual excited LPC and PSOLA methods and MBROLA database. With LPC method, the residuals and LPC coefficients are used as control parameters. It uses letter-to-sound rules and a large lexicon for TTS conversion. Speech synthesis is accomplished using unit-selection concatenation of diphones. Festival offers a general framework for building speech synthesis systems. [12].

D. AT&T VOICEBUILDER

The AT&T VOICEBUILDER supports only English language. It provides a new tool to researchers and practitioners who want to have their own voices synthesized by a high-quality commercial-grade text-to-speech system without the need to install, configure, or manage speech processing software and equipment. It is implemented as a web service on the AT&T Speech Mash up Portal. The system records and validates users' utterances, processes them to build a synthetic voice and provides a web service API to make the voice available to real-time applications through a scalable cloud-based processing platform. All the procedures are automated to avoid human intervention [4].

E. Text-To-Speech System in Indian Languages

IIIT Hyderabad has build a generic framework for Hindi and Telugu languages for generating text processing modules and linguistic resources which could be extended to all of the Indian languages with minimal efforts and time. Anand Arokia et al. approach is to make use of minimal language information, acoustic data and machine learning techniques. Their efforts in this direction support mainly for font identification, Font-to-Akshara conversion, and pronunciation rules for Aksharas and text normalization [13].

F. WHISTLER

Whistler is a trainable Text-to-Speech (TTS) system available in English language that automatically learns the model parameters from a corpus. The speech engine is based on Concatenative synthesis and the training procedure on Hidden Markov Models. The speech synthesis unit inventory for each individual voice is constructed automatically from unlabeled speech database using the Whisper speech recognition system. The speech recognition unit recognizes the speech and automatically labels the speech segments. This method not only improved naturalness but also decreased the time required to create a new voice and made the synthetic speech similar to the original speech donor. [14].

G. Mobile phone based TTS

Now a day's many mobile phone manufacturer companies provide Text to Speech facility. It is very useful for visually disabled person; they can hear the text that appears on a mobile phone screen even for reading e-books. The Google have introduced TTS for Android based mobile phone available in English, Spanish, French and Italian etc. [15]. SVOX corporation introduced its Android phone based TTS. It can read a English text, e-book and can translate a speech to another language speech [16]. C-DAC Mumbai has developed Android phone based TTS for Marathi and Oriya language [17].

H. Text to speech conversion using OCR

This system is consisting of a portable camera, a computing device and speaker. First it reads image document, recognize the characters, numerals and symbols. A reconstructed text is converted into speech [18].

V. SMOOTHING METHODS FOR TEXT-TO-SPEECH SYNTHESIS

In unit selection based Concatenative speech synthesis, joint cost also known as Concatenative cost, which measures how well two units can be joined together. After units are concatenated, most system attempts three join cost function and three smoothing methods such as No smoothing, linear smoothing and Kalman filter based smoothing [19]. T. Dutoit showed that Line Spectral Frequencies (LSF) have good interpolation properties and produce smoother transition than LPC parameters [20].

The speech can be enhanced by using Kalman filter a perceptual post filter concatenated with a standard Kalman filter, it gives the best performance [21]. A text-to-speech system produces neutral speech, it can be converted into emotional speech by modifying the pitch counter (F0) of stressed words by using Gaussian normalization technique [22] [23] [24]. The two HMMs are used as the post processing of text to speech for spectrum conversion from neutral to expressive speech [25] [26]. The expressive speech for storytelling application is generated by applying a set of prosodic rules for converting neutral speech produced by TTS system into storytelling speech with modification in Pitch, Intensity, Tempo and Duration [27]. Multiple acoustic models are often combined in statistical parametric speech synthesis. The combination of multiple acoustic models HMMs and Gaussian gives significant improvement in the quality of the synthesized speech [26] [28]. Table I shows the current status of TTS on Indian Languages.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

TABLE I **CURRENT RESEARCH SCENARIO ON INDIAN LANGUAGES**

Institute	**Language Cover**	**Synthesis Strategy**	**Unit / Database**	**Text/Speech Segment processing/ Tools**	**Prosody**	**Performance**
CEERI, Delhi	Hindi Bengali (partly)	Formant (Klatt – type) Synthesis	Syllables & Phonemes (Parameter Data Base	Manual (Rules for smoothing) Parsing rules for Syllabification	Manual + Some rules	Copy Synthesis Excellent Unlimited TTS-Average
TIFR Mumbai	Hindi Bengali Marathi Indian English (Partly)	Format (Klatt type) Synthesis	Phonemes and other units	Automatic parsing rules for phonemization, Rules for smoothing prosody	Prosody rules	Unlimited TTS – More than Average
IIIT (Hyd.)	Hindi Telugu Other languages	Concatenative	Data base in required languages as per festival norms	For unit as per festival system As per requirements of Festival System	Prosody studies in required language done and implemented	Unlimited TTS- better than Average
HCU (Hyd.)	Telugu	Concatenative	Diphone	MBROLA based	Prosody rules	Unlimited TTS-Average
IIT, Chennai	Hindi Tamil	Concatenative diphone synthesis (1400 diphones)	Syllabus (Mainly)	Automatic segmentation using group delay functions for unit selection Festival System	Pitch tracks determined and implementation	Unlimited TTS-Average
CoE Chennai	Tamil	Concatenative	Diphone	Phonetic segmentation	Prosody rules	TTS Average
IIT Mumbai	Marathi Hindi	Concatenative	Di-phones, Syllables,	Prosody modeling using CART	Prosody rules	TTS-Average
IIT Delhi	Hindi	Concatenative	Unit selection	Rule and corpora based method	Prosody rules	TTS-Average
CDAC, Mumbai	Marathi Odia	Concatenative	Unit selection	Festival based speech synthesis	Prosody rules	TTS-Average
CDAC, Pune	Hindi, Indian English	Concatenative	Phonemes, other units	Festival based speech synthesis	Prosody rules	TTS-Average
CDAC, Noida	Hindi	Concatenative	Multi form units, Diphones Syllables, frequent words, phrases etc.	Parsing for syllables, Statistical processing of text for formation of phonetically rich sentences and other units	Study of intonation patterns,	Domain Specific- Excellent, Unlimited TTS-Average
CDAC Kolkata	Bengali	Concatenative	Phonemes & Sub – Phonemes (Size 1 MB)	Cool – edit Phonemic/ Segmentation	TDPSOLA /ESNOLA	Unlimited TTS-Average

DOI 10.17148/IJARCCE.2015.4812

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

CDAC Trivend-ram	Malayalam	Concatenative	Phonemes	Phonemic/ Segmentation	ESNOLA	TTS Good quality
Bhrigus Software Ltd. Hyd	Hindi, Telugu & Others	Concatenative	Phonemes, Using Festival requirements	Fest VOX tools Festival	Intonation using (CART for Prosody modeling)	Unlimited TTS-Average
Prologix Software, Lucknow	Hindi	Concatenative	Di-phone data base	Festival based-Fest VOX tools	-	Unlimited TTS-better than Average
Webel Mediatronics, Kolkata	Bengali, Hindi	Formant type	Phonemes (Parameters of phonemes)	Rules for concatenation and smoothing of parameters Text processing rules	Intonation rules being implemented .	Unlimited TTS.- Less than Average
RIT Islampur (MS)	Kokni	Concatenative	Units	Rules for concatenation	Prosody rules	Limited TTS very poor
Utkal University Bhuvnesh war	Oriya	Concatenative	Phonemes	Processing of Text parsed in C & V	Prosody rules	TTS-Average
Tapar University Patiala	Panjabi	Concatenative	Diphone, sub-syllabic	Phonetic segmentation	Prosody rules	Unlimited TTS-better than Average
IISC Bangalore	Bengali Hindi Gujarati Kannada Malayalam Marathi Oriya Punjabi Tamil Telugu Pashto	Concatenative	Phones of C & V, syllables	Phonetic transcription	Prosody rules	Unlimited TTS-better than Average

TABLE NO. 1 CURRENT RESEARCH OF SPEECH SYNTHESIS IN INDIA [17], [29] - [34].

VI. CONCLUSION

Test-to-Speech synthesizer has steadily developed from last few decades to gain the current shape. The three basic methods of speech synthesis are Articulator, Formant and Concatenative synthesis used in various synthesizers. Many new applications are developed, but intelligibility and comprehensibility of synthetic speech has not reached the acceptable level. Even in India some research organizations are also working on Text-to-Speech in regional languages like Marathi, Hindi, Telugu, Punjabi, Kannada etc. But all these systems do not replicate the human natural speech. There is huge scope in improvement in the synthesis to achieve a high level of natural and emotion aspect.

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] A. Indumati 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.

DOI 10.17148/IJARCCE.2015.4812

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

[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/2105862/how-to-get-started-with-google-text-to-speech.html.
[16] [Online].Available:(2015)https://svoxbilevoices.wordpress.com/page/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.
[29] Shyam S. Agrawal, Development of Resources & Techniques for Processing of Some Indian Languages, in Proc. Invited lecture in C-DAC (Pune, India 2008.).
[30] [Online].Available:(2015)http://www.iitm.ac.in/donlab/website_files/research/Speech/TTS /contents/main.
[31] [Online].Available :(2015) http://www.iit.ac.in/ncc2014/tutorials.html.
[32] [Online].Available: (2015)http://w3.org/2006/10/ssml/papers/paper.pdf.
[33] [Online].Available :(2015)http://dspace.thapar.edu:8080/dspace/bitstream/1-9/159/8043115.pdf.
[34] [Online].Available :(2015)http://dhvani.sourceforge.net.