A REVIEW ON TEXT TO SPEECH CONVERSION METHODS

[1]Prof. Archana Gondalia, [2]Prof. Namita Patel
[1]Assistant Professor, [2]Assistant Professor
[1]Computer Engineering Department,
[1]Aditya Silver Oak Institute of Technology, Ahmedabad, India

Abstract : Speech is the first important primary need, and the most convenient means of communication between people. The communication among human computer interaction is called human computer interface. This paper gives an overview of major technological perspective and appreciation of the fundamental progress of text to speech conversion and also gives overview technique developed in each stage of classification of speech to text conversion. A comparative study of different technique is done as per stages. This paper concludes with the decision on future direction for developing technique in human computer interface system in different mother tongue and it also discusses the various techniques used in each step of a text recognition process and attempts to analyze an approach for designing an efficient system for speech recognition. However, with modern processes, algorithms, and methods we can process speech signals easily and recognize the text. The objective of this review paper is to recapitulate and match up to different recognition systems as well as approaches for the text to speech conversion and identify research topics and applications which are at the forefront of this exciting and challenging field.

IndexTerms - – **Text-to-Speech conversion (TTS), Speech synthesis, Syllabification, Concatenation, Text Normalization, Text Conversion**.

I. INTRODUCTION

Language is the ability to express one's thoughts by means of a set of signs (text), gestures, and sounds. Text-to-speech (TTS) convention transforms linguistic information stored as data or text into speech. It is most widely used in the audio reading devices for the deaf and dumb people now a deists is one of the major applications of NLP. The NLP module of general TTS conversion system consists of the Pre-processor, text analyzer, contextual analyzer, syntactic prosodic parser, letter to sound module and prosody generator. Synthesized speech can be created by concatenating part of recorded speech which is stored in a database. Speech is often based on concatenation of natural speech that is the units, which are taken from natural speech put together to form a word or sentence.

Concatenative speech synthesis has become very popular in recent years due to its improved sensitivity to unit context over simpler predecessors. Rhythm is an important factor that makes the synthesized speech of a TTS system more natural and understandable. The conversion of text to speech involves many important processes. These processes that can be divided mainly in to three stages such as text analysis, text processing and wave-form generation. Text To Speech synthesis (TTS) is an application to convert the text written in a language into speech. The text to speech conversion system useful for to user to enter text and as output it gets sound. [18]Today there is a wide spread talk about improvement of the human interface to the computer. Because no longer people want to sit and read data from the monitor. Since there is a painstaking effort to be taken, this involves strain to their eyes. In this aspect Speech Synthesis is becoming one of the most important steps towards improving the human interface to the computer.

Here comes the role of the Text To Speech (TTS) engines. Text-To-Speech is a process through which input text is analyzed, processed and "understood", and then the text is rendered as digital audio and then "spoken". It is a small piece of software, which will speak out the text inputted to it, as if reading from a newspaper. There have been many developments found around the world in the development of TTS Engines in various languages like English, French, German etc and even in Hindi. Text-To-Speech (TTS) is a technology that converts a written text into human understandable voice. A TTS synthesizer is a computer based system that can be able to read any text aloud that is given through standard input devices. In general, a TTS system can be broken down into three main parts: a linguistic, a phonetic and an acoustic part. First, an ordinary text is input to the system. A linguistic module converts this text into a phonetic representation. From this representation, the phonetic processing module calculates the speech parameters. Finally, an acoustic module uses these parameters to generate a synthetic speech signal. 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. [18]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. 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. The goal of Text-to-Speech (TTS) synthesis is to convert arbitrary input text to intelligible and natural sounding speech so as to transmit information from a machine to a person. [18]

II. TEXT TO SPEECH PROCESS

The primary goal of Text-to speech (TTS) synthesis is to convert input text into intelligible and natural sounding speech. TTS components comprises two phases [1], the front end which analyses the text, creates possible pronunciations for each word in the context with grapheme to phoneme conversion. The back end generates the speech waveform along with the prosody of the sentence to be spoken. The evaluation of TTS system is based on three characteristics: Accuracy, Intelligibility and naturalness. The HTS system provides the frequency spectrum (Vocal tract), fundamental frequency (vocal source) and duration (Prosody) of speech, which are commonly modeled concurrently by HMMs. Speech waveforms are generated from HMMs themselves based on maximum likelihood criterion [2]. The figure 1 shows architecture of TTS system [3]. The architecture of TTS system is divided into four phases as follows:

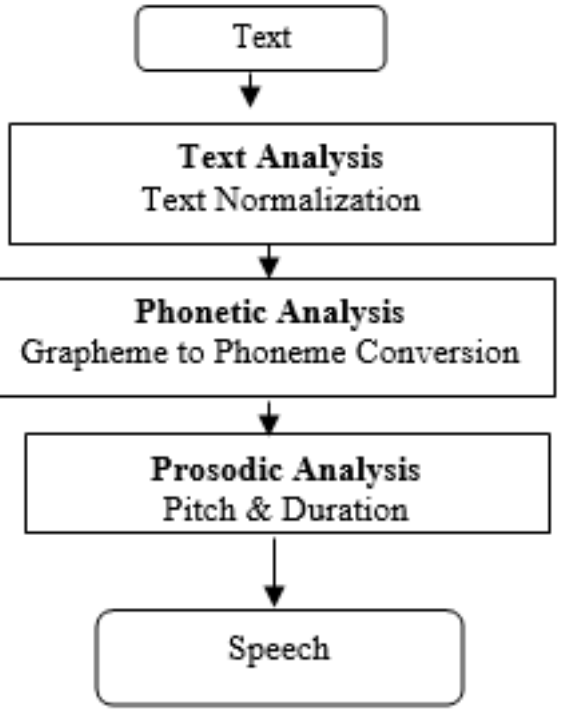

Figure 1Text to speech system architecture

1. Text analysis:

The normalization of the text wherein the numbers and symbols become words and an abbreviation are replaced by their whole words or phrases etc. The linguistic analysis which means syntactic and semantic analysis and aims at understanding the context of the text is also performed in this step. The statistical methods are used to find the most probable meaning of the utterances. This is significant because the pronunciation of a word may depend on its meaning and on the context.

2. Phonetic Analysis:

This converts the orthographical symbols into phonological ones using a phonetic alphabet.

3. Prosodic Analysis:

Prosody contains the rhythm of speech, stress patterns andintonation. The naturalness in speech is attributed to certain properties of the speech signal related to audible changes in pitch, loudness and syllabic length, collectively called prosody. Acoustically, these changes correspond to the variations in the fundamental frequency (F0), amplitude and duration of speech units [1]

4. Speech Synthesis:

This block finally generates the speech signal. This can be achieved either based on parametric representation, in which phoneme realizations are produced by machine, or by selecting speech units from a database. The resulting short units of speech are joined together to produce the final speech signal.

III. LITERATURE SURVEY

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 [6]. T. Dutoit showed that Line Spectral Frequencies (LSF) have good interpolation properties and produce smoother transition than LPC parameters [7].

The speech can be enhanced by using Kalman filter a perceptual post filter concatenated with a standard Kalman filter, it gives the best performance [8]. 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[9] [10] [11]. The two HMMs are used as the post processing of text to speech for spectrum conversion from neutral to expressive speech [12] [13]. 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 [14]. 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 [15] [16]. Table I shows the current status of TTS on Indian Language.

Institute	**Language Cover**	**Synthesis Strategy**	**Unit / Database**	**Text/Speech Segment processing/ Tools**	**Prosody**	**Performance**
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
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
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 implementati on	Unlimited TTS-Average
Tapar University Patiala	Panjabi	Concatenative	Diphone, sub-syllabic	Phonetic segmentation	Prosody rules	Unlimited TTS-better than Average
RIT Islampur (MS)	Kokni	Concatenative	Units	Rules for concatenation	Prosody rules	Limited TTS very poor
CDAC Kolkata	Bengali	Concatenative	Phonemes & Sub – Phonemes (Size 1 MB)	Cool – edit Phonemic/ Segmentation	TDPSOLA /ESNOLA	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
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 SpecificExcellent, Unlimited TTS-Average
IIT Mumbai	Marathi Hindi	Concatenative	Di-phones, Syllables,	Prosody modeling using CART	Prosody rules	TTS-Average
CoE Chennai	Tamil	Concatenative	Diphone	Phonetic segmentation	Prosody rules	TTS Average
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
CDAC Trivendram	Malayalam	Concatenative	Phonemes	Phonemic/ Segmentation	ESNOLA	TTS Good quality
Utkal University Bhuvnesh war	Oriya	Concatenative	Phonemes	Processing of Text parsed in C & V	Prosody rules	TTS-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
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

TABLE NO. 1 CURRENT RESEARCH OF SPEECH SYNTHESIS IN INDIA [4], [12] - [15].

IV. APPLICATION TEXT TO SPEECH

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 affordable for common customers, which makes these systems more suitable for everyday use. Some uses of TTS are described below.

1) Aid to Vocally Handicap:

A hand-held, battery-powered synthetic speech aid can be used by vocally handicapped person to express their words. The device will have especially designed keyboard, which accepts the input, and converts into the required speech within blink of eyes.

2) Source of Learning for Visually Impaired:

Listening is an important skill for people who are blind. Blind individuals rely on their ability to hear or listen to gain information quickly and efficiently. Students use their sense of hearing to gain information from books on tape or CD, but also to assess what is happening around them

3) Talking Books and Toys

Talking book not only teaches how to read but also has more impact on students than text reading. It makes their study more enjoyable and easy. In the same way talking toys are a great source of fun and entertainment for children.

4) Games and Education.

Synthesized speech can also be used in many educational institutions in field of study as well as sports. A teacher can be tired at a point of time but a computer with speech synthesizer can teach whole day with same efficiency and accuracy.

5) Telecommunication and Multimedia.

TTS systems make it possible to access textual information over the telephone. Texts can be large databases which can hardly be read and stored as digitized speech. Queries to such information retrieval systems could be put through the user's voice (with the help of a speech recognizer), or through the telephone keyboard. Synthesized speech may also be used to speak out short text messages in mobile phones.

6) Man-Machine Communication.

Speech synthesis may be used in several kinds of human machine interactions. For example, in warning, alarm systems, clocks and washing machines synthesized speech may be used to give more accurate information of the current situation. Speech signals are far better than that of warning lights or buzzers as it enables to react to the signal more fast if the person is unable to get light due some obstacles.

7) Voice Enabled E-mail.

Voice-enabled e-mail uses voice recognition and speech synthesis technologies to enable users to access their e- mail from any telephone. The subscriber dials a phone number to access a voice portal, then, to collect their e- mail messages, they press a couple of keys and, perhaps, say a phrase like "Get my e-mail." Speech synthesis software converts e-mail text to a voice message, which is played back over the phone. Voice-enabled e-mail is especially useful for mobile workers, because it makes it possible for them to access their messages easily from virtually anywhere (as long as they can get to a phone), without having to invest in expensive equipment such as laptop computers or personal digital assistants

IV. CONCLUSION AND FUTURE WORK

In this paper, we discussed the topics relevant to the development of TTS systems .The text to speech conversion may seem effective and efficient to its users if it produces natural speech and by making several modifications to it. This system is useful for deaf and dumb people to Interact with the other peoples from society. Text to speech synthesis is a critical research and application area in the field of multimedia interfaces. In this paper, a speech synthesis system has been designed and implemented for Hindi Language. A database has been created from the various domain words and syllables. The given text is analyzed and syllabified based on the syllable segmentation rules. The desired speech is produced by the Concatenative speech synthesis approach. Speech synthesis is advantageous for people who are visually handicapped. This paper made a clear and simple overview of working of text to speech system (TTS) in step by step process. The Text to Speech System for Hindi using English Language is able to speak a loud Hindi word which is typed in English. The system read the input data in a natural form. The user types the input string and the system reads it from the database or data store where the words, phones, diphones, triphone are stored. In this paper, we presented the development of existing TTS system by adding spell checker module to it for Hindi language. There are many text to speech systems (TTS) available in the market and also much improvisation is going on in the research area to make the speech more effective, and the natural with stress and the emotions.

REFERENCES

1. **[1]** T. Dutoit, "An Introduction to Text-to -Speech Synthesis", Kluwer Academic Publishers, 1997.
2. Black, A. Zen, H., Tokuda, K. "Statistical Parametric Synthesis", in proc. ICASSP, Honululu, USA,2007.
3. [Online].Available:(2015)http://cdac.in/index.aspx?id=mcst_speech_technology.
4. 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.
5. 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.
6. T. Dutoit, An Introduction to Text-to-Speech Synthesis, Kluwer Academic Publishers, Dordrecht, ISBN 0-7923-4498-7, (1997).
7. 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.
8. Marc Schröder, Expressing Degree of Activation in Synthetic Speech, IEEE Trans. Speech Audio Process, Vol. 14, (2006) pp. 1128 – 1136.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. Shyam S. Agrawal, Development of Resources & Techniques for Processing of Some Indian Languages, in Proc. Invited lecture in C- DAC (Pune, India 2008.).
16. [Online].Available:(2015)http://www.iitm.ac.in/donlab/website_file s/research/Speech/TTS /contents/main.
17. [Online].Available :(2015) http://www.iit.ac.in/ncc2014/tutorials .html.
18. Kaveri Kamble , Ramesh Kagalkar , "A Review: Translation of Text to Speech Conversion for Hindi Language",IJSR,Volume 3 Issue 11, November 2014
19. [Online].Available: (2015)http://w3.org/2006/10/ssml/papers/paper. pdf.
20. [Online].Available :(2015)http://dspace.thapar.edu:8080/dspace/bit stream/1-9/159/8043115.pdf.
21. [Online].Available :(2015)http://dhvani.sourceforge.net.