A Review: Translation of Text to Speech Conversion for Hindi Language

Kaveri Kamble[1], **Ramesh Kagalkar**[2]

[1]Department of Computer Engineering, Dr. D. Y. Patil School of Engineering & Technology,
Savitribai Phule University of Pune, India

[2]Assistant Professor, Department of Computer Engineering, Dr. D. Y. Patil School of Engineering & Technology,
Savitribai Phule University of Pune, India

Abstract: ***In daily life Speech and spoken words have always played a big role in the individual and collective lives of the people. The Speech that represents the spoken form of a language. Speech synthesis is the process of converting message written in text to equivalent message in spoken form .A Text-To-Speech (TTS) synthesizer as a computer-based system that should be able to read text. In this paper, I am explaining single text-to-speech (TTS) system for Indian languages Viz., Hindi to generate speech .This generally involves two steps, text processing and speech generation. A graphical user interface has been designed for converting Hindi text to speech in Java Swings. In India there are different languages are spoken, but each language is the mother tongue of tens of millions of people. The languages and scripts are very different from each other. The grammar and the alphabet words are similar to a large extent. This paper present text-to-speech (TTS) system based on the Concatenative synthesis approach.***

Keywords: Text-to-Speech conversion (TTS), Speech synthesis, Syllabification, Concatenation, Text Normalization, Text Conversion

1. 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 days.TTS 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. 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. 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.

2. Text to Speech System

The main function of text-to-speech (TTS) system is to convert an arbitrary text to a spoken Waveform. This task general**ly** consists of steps, *i.e.*, text analysis, text normalization, text processing, acoustic processing and speech generation. Text analysis part is preprocessing which analyze the input text and organize into manageable list of words, Text normalization is the transformation of text to pronounceable form. The main objective of this process is to identify punctuation marks and pauses between words, Text processing is the conversion of the given text into a sequence of synthesis units, Acoustic processing –the speech will be spoken out the voice characteristic of a person like three types such as Concatenative synthesis, Formanant synthesis, Articultory synthesis whereas speech generation is generation of an acoustic wave form corresponding to each of these units in the sequence.

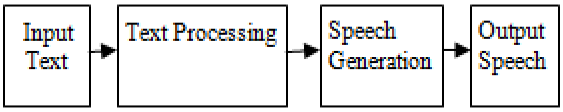

Figure 1: Text to Speech System

3. Literature Survey

Synthesized speech can be created by concatenating part of recorded speech which is stored in a database. The mainly significant qualities of a speech synthesis system are naturalness and Intelligibility [1]. Naturalness expresses how intimately the output sounds like human speech, whereas intelligibility is the easiness with which the output is understood. The main function of text-to-speech (TTS) system is to convert an arbitrary text to a spoken Waveform. TTS is one of the major applications of NLP. TTS Synthesizer is a computer based system that should be understand any text clearly whether it was establish in the computer by an operator or scanned and submitted to an Optical Character Recognition (OCR) system. Single text-to-speech (TTS) system for Indian languages viz.Hindi. text-to-speech (TTS) system based on the Concatenative synthesis approach. This conversion involves text processing and speech generation processes. These processes have the connections to use the linguistic theory, models of speech production, and the acoustic-phonetic characterization of the language. Font Characters Mapping is required for font characters of a Indian language to represent vowel and consonant alphabet.

Text analysis is the task of identifying words in the text [10].Text normalization includes "Token Identification" which is the task of identifying special symbols, numbers and "Token to Words" which convert the identified tokens to words for which there is a well defined method of pronunciation.

In a text-to-speech system, spoken utterances are automatically produced from text [2].Tamil text-to-speech (TTS) system based on the Concatenative synthesis approach. The TTS can be a voice for those people who cannot speak. The TTS system can be useful for the SMSs, Web pages, News, any articles, and the Microsoft office tools and so on. Text-to-speech system based on Concatenative synthesis needs well arranged speech corpus. The quality of synthesized speech waveform depends up on the number of realization of various units present in the speech corpus. A good quality microphone should be used to avoid noise in speech wav file. In text-to-speech, the accuracy of the system is calculated in the ways of naturalness and intelligibility.

N. Swetha[5] in her paper A blind person cannot also see the length of an input text when starting to listen it with the help of the speech synthesizer, an important feature is to give in advance some information of the text to be read successfully. The synthesizer use to check the document and calculate the estimated duration of reading and speak it to the listener [5]. There are several problems in text pre-processing, such as numerals, abbreviations, and acronyms. Syllable is a part of a word that contains a single vowel sound and that is pronounced as a unit Text-to-Phonetic Conversion The first task faced by any TTS system is the conversion of input text into linguistic representation, usually called text-to-phonetic or grapheme-to phoneme conversion.

Text pre-processing is usually a very complex task and includes several language dependent problems. Digits and numerals must be expanded into full words. The second task is to find correct pronunciation for different contexts in the text. Some words, called homographs, cause the most difficult problems in TTS systems. To Find the correct intonation, stress, and duration from written text is probably the most challenging problem arrived.

Poonam S. Shetake In this paper text-to-speech (TTS) convention transforms linguistic information stored as data or text into speech.TTS systems make it possible to access textual information over the telephone [12].The synthesizer produces speech signals of 16 bits, the sampling rate of which is determined by the sampling rate of the diphone database used. A text to speech (TTS) synthesizer is a system that can read text aloud automatically, which is extracted from Optical Character Recognition (OCR). Diphone which contains the transitions between two phones, has been chosen as the synthesis unit for Concatenative synthesizers [9]. There are about 1500 to 2000 diphones in English, and the diphone mapping for a phoneme string is straightforward

Masatsune Tamura In this paper describes a technique for synthesizing speech with any desired voice. The technique is based on an HMM-based text-to-speech (TTS) system and MLLR adaptation algorithm [16]. In the approach, the HMM-based TTS system is used and the voice characteristics of synthetic speech are changed by transforming HMM parameters of the speech units in the MLLR adaptation framework. To generate speech with an arbitrarily given target speaker’s voice, we adapt the speaker independent models, i.e., average voice models, to the target speaker. In this paper, we described a technique for adapting voice characteristics and prosodic features of HMM-based TTS system to an arbitrarily given target speaker.

Susan r. Hertz a primary factor in determining the choice of synthesis strategy is the rule-writer's linguistic convictions. For example, the proponents of an approach based on demisyllables claim that many of the influences of adjacent sounds on each other are automatically present in the demisyllable[14]. Even more challenging is handling phenomena that cannot easily be captured in rules

Shen Zhang he proposes an approach on emotional audio-visual speech synthesis. Our approach primarily consists of three steps: first we take the text and the target PAD values as input, and employ text-to-speech (TTS) engine to generate neutral speeches. The prosodic word boundaries are automatically predicted by the text analysis module of TTS engine. In our TTS system, maximum entropy (ME) model is used for prosodic word boundary prediction. The expression of human emotion only has gained special attention recently in both audio speech synthesis and talking face animation. To synthesize the dynamic facial expression in continuous speech, the speech acoustic features (e.g., pitch) are taken as important clues to modulate the facial expression on sentence level.

4. System Design and Implementation

From Fig2.The Approach used for implementing Hindi TTS involved basically Three Steps such as Text-Processing, Text Processing, Speech synthesis. In Text Pre-processing basically Text as input which entered manually after that Spelling Checking is important part involved in that. Next is the Text Processing which formed Text standardization and Normalization which is Dictionary based and the Grapheme to Phoneme rule based conversion. Graphemes are usually considered to be the smallest functional units of a written language and Phonemes are the elementary sounds of a language Combination of phonemes gives rise to next higher unit known as syllables which is one of the most important units of a language. A syllable must have a vowel called as its a nucleus, while the presence of consonant is optional. These syllable types are: V, CV, VC, VCC, CVC, CCVC and CVCC; where V and C represent vowel and consonant respectively and The speech can be synthesized by concatenating different pieces of recorded speech from the database. Speech synthesis which involved phonetic rule based Waveform Generation and output as Speech.

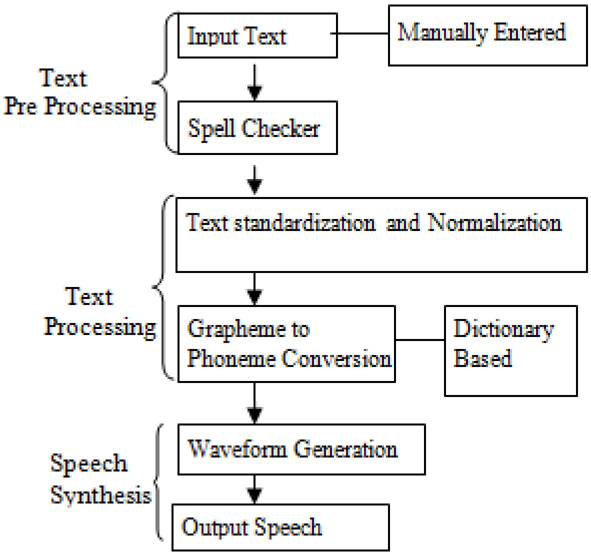

Figure 2: Approach used for implementing Hindi TTS

The general form of Indian the language syllable is as following:

C*VC*,

where C is a consonant, V is vowel and C* indicates the presence of 0 or more consonants. A character in Indian language scripts is close to syllable and can be typically of the following form: C, V, CV, CCV and CVC. There are about 35 consonants and about 18 vowels in Indian languages. Mapping is required for font characters of a Indian language to represent vowel and consonant alphabet.

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

1) **Aid to Vocally Handicapped**
   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.

6. Methodology

The Methodology to be followed for the project is as follows:

1) Concatenative speech synthesis techniques will be used in order to get the naturalness quality in the synthetic speech.
2) The phonemes of the Hindi language can be used as the basic unit for speech synthesis.
3) To design a user interface in which the end user can write editable Hindi text to be converted into speech in the text box
4) Speech database for Hindi Language will be developed by using phoneme.
5) To design a user interface having keyboard of Hindi characters so that the end user can easily type the text in Hindi.
6) The input text will be separated into Hindi Phoneme.
7) Phonemes will be searched in the database and corresponding phonemes sounds will be concatenated to generate synthesized output speech.

7. Conclusion

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 spellchecker 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.

8. Acknowledgment

The authors would like to thank Chairman Groups and Management and the Director/Principal Dr. Uttam Kalwane, Colleague of the Department of Computer Engineering and Colleagues of the Department the D. Y. Patil School of Engineering and Technology, Pune Dist. Pune Maharashtra, India, for their support, suggestions and encouragement.

References

[1] Mrs. S. D. Suryawanshi, Mrs. R. R. Itkarkar, Mr. D. T. Mane, “High Quality Text to Speech Synthesizer using Phonetic Integration” , International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 3, Issue 2, February 2014.

[2] J. Sangeetha,S. Jothilakshmi, S. Sindhuja, V. Ramalingam, “Text to Speech synthesis system for Tamil”, International Conference on Information Systems and Computing (ICISC-2013), INDIA.

[3] Jia *Member, IEEE,* Shen Zhang, Fanbo Meng, Yongxin Wang, and Lianhong Cai, *Member, IEEE,”* Emotional Audio-Visual Speech Synthesis Based on PAD”, IEEE transactions on audio, speech, and language processing, vol. 19, no. 3, march 2011.

[4] Maninder Singh1, Karun Verma2,” Text to Speech Synthesis for numerals into Punjabi language”,

International Journal of Computational Linguistics and Natural Language Processing Vol 2 Issue 7 July 2013 ISSN 2279 – 0756.

[5] n.swetha 2k..anuradha," Text-to-speech conversion", International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 269-278 (2013).

[6] Swati Ahlawat, Rajiv Dahiya," A Novel Approach of Text to Speech Conversion Under Android Environment", IJCSMS International Journal of Computer Science & Management Studies, Vol. 13, Issue 05, July 2013.

[7] D.Sasirekha, E.Chandra," Text to Speech: A Simple Tutorial", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[8] Miss. Priyanka V. Mhamunkar, Mr. Krishna S. Bansode and Prof. Laxman S. Naik," Android Application to get Word Meaning through Voice", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 2, February 2013**.**

[9] H. Hon, A. Acero, X. Huang, J. Liu, and M. Plumpe," Automatic Generation of Synthesis Units for Trainable Text-to-Speech Systems".

[10] Lakshmi Sahu and Avinash Dhole,"Hindi &Telugu text-to-Speech Synthesis(TTS) and inter-language text Conversion", International Journal of Scientific and Research Publications, Volume 2, Issue 4, April 2012 1 ISSN 2250-3153.

[11] Ameera Al-Rehili1, Dalal Al-Juhani2, Maha Al-Maimani3 and Munir Ahmed," A Novel approach to convert speech to Text and Vice-Versa and Translate from English to Arabic Language", International Journal of Science and Applied Information Technology, ISSN No. 2278-3083, Volume 1, No.2, May – June 2012.

[12] 1poonam.S.Shetake, 2s.A.Patil, 3p. M Jadhav," Review Of Text To Speech Conversion Methods", Proceedings of 10th IRF International Conference, 01st June-2014, Pune, India, ISBN: 978-93-84209-23-0.

[13] Tapas Kumar Patra, Biplab Patra, Puspanjali Mohapatra," Text to Speech Conversion with Phonematic Concatenation", International Journal of Electronics Communication and Computer Technology (IJECCT) Volume 2 Issue 5 (September 2012).

[14] Susan R. Hertz, James Kadin, And Kevin J. Karplus, Member, Ieee," The Delta Rule Development System for Speech Synthesis from Text", Proceedings Of The IEEE, Vol. 73, No. 11, November 1985.

[15] Sangam P. Borkar, Prof. S. P. Patil," Text To Speech System For Konkani (Goan)Language".

[16] Masatsune Tamura, Takashi Masuko, Keiichi Tokuda, Takao Kobayashi," Text-To-Speech Synthesis With Arbitrary Speaker's Voice From Average Voice", Interdisciplinary Graduate School of Science and Engineering, Euro speech 2001 Scandinavia.

[17] **1**poonam.S.Shetake, 2s.A.Patil, 3p. M Jadhav," Review Of Text To Speech Conversion MethodS**",** International Journal of Industrial Electronics and Electrical Engineering, ISSN: 2347-6982 Volume-2, Issue-8, Aug.-2014.

[18] Théophile K. Dagbaa,*, Charbel Boco," A Text To Speech system for Fon language using Multisyn Algorithm", 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014.

[19] Simon King," An introduction to statistical parametric speech synthesis", *Sadhana*¯ Vol. 36, Part 5, October 2011, pp. 837–852._c Indian Academy of Sciences.

[20] Leija, L.Santiago, S.Alvarado, C., "A System of Text Reading and Translation to Voice for Blind Persons," Engineering in Medicine and Biology Society, 1996.

[21] B. Yegnanarayana, *Senior Member*, *IEEE*, and K. Sri Rama Murty," Event-Based Instantaneous Fundamental Frequency Estimation From Speech Signals",Ieee Transactions On Audio, Speech, And Language Processing, Vol. 17, No. 4, May 2009.

Author Profile

Kaveri S. Kamble Research Scholar Dr. D.Y.Patil School of Engineering & Technology, Pune, Savitribai Phule University of Pune. She received B.E. in Computer Engineering from Computer Department of Sinhgad Institute of technology, Lonavala, Pune from Savitribai Phule Pune University. Currently she is persuing M.E. in computer engineering from Dr. D. Y. Patil School of Engineering & Technology, Savitribai Phule Pune University of Pune.

Ramesh M. Kagalkar was born on June 1st, 1979in Karnataka, India and presently working as a Assistant. Professor, Department of Computer Engineering, Dr. D.Y.Patil School Of Engineering and Technology, Charoli, B.K. via – Lohegaon, Pune, Maharstra, India. He is a Research Scholar in Visveswaraiah Technological University, Belgaum, He had obtained M.Tech (CSE) Degree in 2005 from VTU Belgaum and He received BE (CSE)Degree in 2001 from Gulbarga University, Gulbarga. He is the author of text book Advance Computer Architecture which cover the syllabus of Visveswaraiah Technological University, Belgaum. He has published many research paper in International and international conference.