International Journal of Advance and Applied Research
www.ijaar.co.in
ISSN – 2347-7075 Impact Factor – 8.141
Peer Reviewed Bi-Monthly
Vol. 6 No. 38 September - October - 2025

A Review of Current Techniques in Text-to-Speech Synthesis

Mahesh S. Gaikwad[1] & Dr. Sanjay T. Wani[2]
[1]*Womens College of Home Science & BCA, Loni, India*
[2]*Womens College of Home Science and BCA, Loni, India*
Corresponding Author – Mahesh S. Gaikwad
DOI - 10.5281/zenodo.17309810

Abstract:

Speech is the most prominent and natural form of communication between human beings. Only a select few who are literate in a given language can access the majority of the digital material available today. The application of NLP provides solutions in the form of natural interfaces, such as TTS, allowing digital content to reach a wider audience and facilitate the exchange of information among different people. A variety of technologies, including text-to-speech technology, have been developed to enhance language learning. The technique of turning text written in natural language into voice is called text-to-speech synthesis. This review paper includes an overview of text-to-speech systems developed for Indian languages. The work done in the speech domain for the Indian language is at the primary stage. The research work for Indian languages is also carried out however, the work is not able to cover the complete phonetic variation of the language.

Keywords: Text-to-speech, Indian languages, Techniques, speech synthesis.

Introduction:

Speech is the most prominent and natural form of communication between humans being. Speech has the ability to express one's thoughts by means of a set of signs, whether acoustic, musical, graphical, gestural. Today's most information in digital form is accessible to a few who can read or understand a particular language. Language technologies can gives solutions in the form of natural interfaces so the digital form content can reach to the masses and facilitate the exchange of information across different people.

Humans have been driven to develop computers with human-like comprehension and speech. In this regard, scientists have worked to create a system for speech signal analysis and categorization. Since, 1960s computer scientists have been researching ways and means to make computer record, interpret and understand text and converted into human speech.

A Text-to-Speech (TTS) converts a raw text into human speech sounds. It can be powerful assistance to communication for visually impaired people and also in telecommunication, industrial and educational applications. Government of India commence development of TTS systems for Indian languages through TTS consortium project under the Ministry of Electronics and Information (MeitY) [6]. So many institutions have been working on speech synthesis such IIIT-H, CDAC- Mumbai, SDAC-Pune and IIT-Madras in various languages. They have

built applications as e-speak, a-speak, Sandesh Pathak.

The amount of work for Indian languages in Speech synthesis has not yet reached to a critical level to be used as real communication tool, as that in other languages of developed countries. For developing a real communication tool the speech synthesis system should continuous synthesis the text into speech like human being.

Methodology:

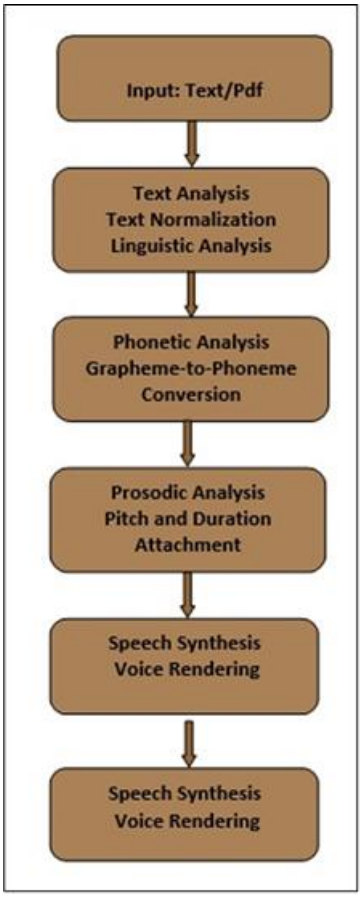

Figure 1: General methodology for Speech Synthesis system.

Techniques of Speech Synthesis:

Articulatory Synthesis:

Articulatory synthesis models the physical articulators such as lips, jaws, tongue, soft palate and so on [15]. In this human speech production system is modelled. It involves simulating the acoustic parts of vocal tract and its dynamic movement. The command parameters are sub-glottal pressure, vocal cord tension and the relative location of different articulatory organs. It produces intelligible synthetic speech but it is far from natural sound and hence not widely used.

Formant Synthesis:

This is a rule based synthesis technique, which describes the resonant frequencies of the vocal tract. This method uses source-filter model of language output. The parameters controlling the frequency response of the vocal tract filter and those controlling the source signal are updated at each phoneme. Excitation produced by the root passes through the filter, is qualified by the resonance characterizes of the filter to create language.

Hidden Markov Model (HMM):

The Hidden Markov Model is use statistical models to characterize the sequence of speech spectra and have successfully been applied to speech synthesis systems. This system simultaneously model spectrum, excitation and continuance of speech using content dependent HMMs and generates speech waveforms. HMM creates stochastic model from known utterances and compares the probability that the unknown utterance can be engendered by each model. This approach gives good prosody features with natural sound language.

Concatenative Synthesis:

Concatenative synthesis is the most uncomplicated method to synthesize the speech which is got by concatenating the different sentences, words, syllables, phones, diphones and triphones. These are already stored to get the desired output language. It requires large databases sometimes it is quite impossible to store. This technique produces more natural sound.

There are three subtype of Concatenative synthesis

Unit Selection Speech Synthesis (Corpus Based Synthesis):

Unit selection synthesis uses large database. During database creation, each recorded utterance is segmented into some individual phones, syllables, morphemes, words, phrases and sentences. An index of the units in the speech database is then made based on the segmentation and acoustic parameters such as fundamental frequency, pitch, duration, the state of syllable and previous and next phones. This method provides naturalness in output speech as compared to other techniques.

Diphone Synthesis:

This technique requires fewer databases as compared to the unit selection synthesis. It uses two adjacent phones to make the speech waveform. But this technique suffers through the problem of co-articulation.

Domain Specific Synthesis:

Domain specific synthesis means related to the specific field associated synthesis. In this synthesis database consists of language related to specific line of business and that are used to concatenate to create the speech.

Text Analysis:

Text Normalization:

The first step in all text-to-speech is normalizing the input text into sentence then each sentence has to be divided into a collection of tokens that is in form of words, numbers date and so on. Non-natural language tokens as abbreviations and acronyms transformed to natural language tokens. [14]

Sentence tokenization:

This is the first step of text normalization. In sentence tokenization has some complications because sentences not always terminated by period (.) and sometime sentence terminated by colon (:). To identify sentence boundaries, the input text is divided into tokens separated by whitespace and then any tokens such as !, .,? is selected then decision can be dependent on machine learning as these n tokens indicate in end-of-sentence or not.

Non-Standard Words:

Second step of text normalization is normalizing non-standard words that are numbers, dates, abbreviations or acronyms. These tokens have to be transformed to a sequence of natural words. Abbreviation dictionaries are used because of the intricate and unclear forms of acronyms and abbreviations.

Homograph Resolution:

Third step of text normalization is homograph resolution. A homograph is the words that have same sequence of characters but differ in pronunciation.

Accent:

After text normalization the next step is to find the proper accent for each. The input text can include words or names that cannot be found in the lexicon. The name-accent lexicon is used by so many text-to-speech systems. The accent of unknown words that is not in accent lexicon can be produced through grapheme-to-phoneme conversion method.

Phonetic Analysis (Grapheme-to-Phoneme):

Phonetic alphabets are used in phonetic analysis to translate orthographic symbols into phonological representations.

Prosodic Analysis

Prosody means rhythm of speech, stress patterns and intonation. Naturalness speech is having the certain properties of speech signal that is related with audible changes in pitch, syllabic length and loudness are collectively called as prosody.

Literature Review:

Sai Sawant et al. [1] in this research paper the researcher have used phoneme based concatenative synthesis for English language. This system is implemented with the help of MATLAB map data structure and Matrix operations. In this system researcher phonetically 42 words in English language are recorded and after that phoneme are extracted using PRAAT tools. Extracted phoneme compared with input text phoneme and then concatenated sequentially to reconstruct the desired word. This method is simple and it requires less memory. Speech quality of this synthesis is more naturalness.

Sneha C. Madre et al. [3] in this paper review shows that the researcher has used OCR for extracting text from image and then recognized character is converted into audio using MATLAB. This system is easy but problem where inevitable like making templates of every character with every font. Researcher describe system is specially used for visually challenged people.

Pravin M. Ghate et al. [4] this paper researcher has used unit selection speech synthesis for Marathi language. Researcher has uses combinations of units like Syllable, Words and Barakhadi as databases. In this synthesis deeply focused on words corpus. Researcher evaluated that speech quality is more than 80 per cent naturalness.

Rupinderdeep Kaur et al. [5] this paper review shows that the researcher has developed speech synthesis for Punjabi language using general approach such as Text processing, Linguistic analysis, Prosodic prediction and Waveform generation. In this paper researcher has described general architecture of TTS and different waveform generation method i.e Formant synthesis, Concatenative synthesis and Statistical parameter synthesis.

Itunuoluwa Isewon et al. [8] in this paper researcher used NLP and digital signal processing for designing speech synthesis in English Nigerian language. Researcher describe that natural language processing module produces a phonetic transcription of text read together with prosody and then Digital signal processing module transforms symbolic information from NLP into audible speech . Here author create the GUI based application for speech synthesis.

Sangramsing Kayte et al. [9] in this paper researcher has used Hidden Markov Model based speech synthesis for Marathi language. Researcher describes that routes of speech parameters are generated from the trained Hidden Markov Models. This synthesis consists of two parts as Training and Synthesis. In training part, spectrum and excitation parameters are extracted from the annotated speech database and converted to a sequence of observed feature vectors which is modelled by a respective sequence of HMM. Here spectrum part represented by met-cepstral coefficients and delta-delta coefficient. Delta-delta coefficient and LogF0 denote the excitation portion. In synthesis part, input text is converted into sequence of contextual label then as per the label sequence HMM sequence is constructed by concatenating context dependent HMM. After this met-cepstral coefficient and F0 routes are generated by using the parameter generation algorithm. Speech waveform is synthesized directly from the generated mel-cepstral coefficients and F0 values by using MLSA filter. In this speech synthesis, speech database developed by IIT-Hyderabad. It consists of 12 type of 1000 sentence for training.

Sangramsing N. Kayte et al. [11] in this paper researcher deals with a corpus driven Marathi TTS system based on concatenative synthesis. For implementation of TTS in

Marathi the MATLAB 2014 has been used. This synthesis consists of first text analysis performed then sentence splitting i.e. Syllabification based on Linguistic rules. After that waveform concatenation process, required syllables are retrieved from speech corpus based on text analysis and arranged to produce the speech. During the concatenation process of speech unit, there will be glitches in the joint. These are removed with the help of waveform smoothing process. For the spectrum smoothing, use time scale modification method and PRAAT software used to calculate duration value for each syllable. Also linear predictive coding used for representing the spectral envelop of digital signal of speech.

Nilesh Fal Dessai et al. [13] in these researchers have used Concatenative synthesis method for Konkani language. The quality of generated speech depends on the unit size. Researcher describe that word synthesis is better than di-phone or phoneme synthesis. The naturalness and intelligibility of any concatenated voice synthesis system are crucial features. They have performed listeners test, it is observed that the developed system performs better than eSpeak in terms of uniqueness, naturalness and intelligibility. When we use small size of unit the quality of speech not good but coverage will be more. When we increase size of unit it increases the quality of synthesized speech but we can't cover the whole language.

Sangramsing Kayte et al. [15] this paper review shows that the researcher has used Concatenative speech synthesis for Marathi language. They had only worked on Marathi language. In this paper researcher deeply focused on Di-phone. Researchers use the Diagnostic Rhyme Test (DRT) to compare two synthesizers: the Festival TTS and the MARY TTS system. The naturalness of the synthesized speech for both the synthesizers needs to be enhanced. The naturalness results are not so good because di-phone database contains only one instance of each speech unit. So researcher says that di-phone based speech synthesis quality is not good.

Nikisha Jariwala et al. [18] in this research paper researcher have used concatenative synthesis method using MATLAB tool for Gujrathi language. This system is mostly developed for visually impaired people. For speech synthesis researcher developed syllable corpus.

Amit Kumar Jha et al. [19] Researcher have used concatenative synthesis method for Maithili language. Maithili language is syllabic in nature. Researcher develops the speech corpus as a syllable. For speech enhanced naturalness purpose author record and store most frequently occurring words. The quality of synthesized speech in terms of intelligibility and naturalness is evaluated to approximately 84 percent. The speech corpus consist of 930 syllable (C * V) in total. Each position has 300 syllables and 10 independent vowels. 930 units of speech data is built from all three positions i.e. initial, middle and final to account for maximum possible phonetic coverage.

S.D. Shirbahadurkar et al. [21] in this research paper researcher describe TTS system using concatenative synthesis method for Marathi language with unit selection speech database. Researcher developed Marathi speech synthesizer using different choice of units as word, phonemes and syllables as a database. The quality of this synthesized speech in terms of intelligibility and naturalness is evaluated to approximately 81 percentages.

Soumitra Das et al. [24] in this researcher proposed speech synthesizer for Marathi language using Syllabic approach. Syllabic

approach is dividing a word into syllable units for natural speaking. In this technique word divided into syllabic cluster for that first find out the correct structure of word. Accuracy of speech using this approach is good.

Saadin Oyucu et al. [25] in this research paper researcher developed speech synthesizer using deep learning for Turkish language. This system handles complex input text such as image and video. He developed Turkish corpus and proposed a Tacotron 2 + HiFi-GAN structure for TTS system. Speech qualify of this TTS system is best as per MOS score 4.49.

Conclusion:

Text-to-speech is the application of NLP. This is a best tool for learning. The work carried out in the speech domain for English and other European languages have achieved an accuracy of more than 85% - 90% synthesized rate. The speech synthesizer and their various techniques that have been examined in this paper. We determined that a large number of speech synthesizing studies deal with foreign languages as compared with Indian languages. A TTS system with several speech synthesis techniques together produces a higher quality result.

References:

1. Sai Sawant and Mangesh Deshpande. English Text to Speech Synthesizer Using Concatenation Technique. Springer Nature Singapore Pte Ltd. 2018 M. Singh et al. (Eds.): ICACDS 2018, CCIS 905, pp. 471–480, 2018.
2. Er.Sheilly Padda, Er. Nidhi, Ms. Rupinderdeep Kaur. A Step towards Making an Effective Text to speech Conversion System. International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 Vol. 2, Issue 2, Mar-Apr 2012, pp.1242-1244
3. Sneha.C.Madre and Prof.S.B.Gundre. OCR Based Image Text To Speech Conversion Using MATLAB. Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS 2018) IEEE Xplore Compliant Part Number: CFP18K74-ART; ISBN: 978-1-5386-2842-3
4. Pravin M Ghate*1 and S.D.Shirbhadurkar. SPEECH SYNTHESIS USING SYLLABLE FOR MARATHI LANGUAGE. INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY ISSN: 2277-9655 [Ghate * et al., 7(1): January, 2018]
5. Rupinderdeep Kaur, Mr. R.K. Sharma and Mr. Parteek Kumar. BUILDING A TEXT-TO-SPEECH SYSTEM FOR PUNJABI LANGUAGE. Natarajan Meghanathan et al. (Eds): ACSIT, SIPM, FCST, CoNeCo, CMIT - pp. 71–87, 2017.
6. TDIL TTS, “Indian Language Technology Proliferation and Deployment Centre: Text-to-speech,” available: http://tdil-dc.in/.
7. Text to Speech Conversion with Phonematic Concatenation. Tapas Kumar Patra, Biplab Patra and Puspanjali Mohapatra. International Journal of Electronics Communication and Computer Technology (IJECCT) Volume 2 Issue 5 (September 2012)
8. Itunuoluwa Isewon, Jelili Oyelade and Olufunke Oladipupo. Design and Implementation of Text To Speech Conversion for Visually Impaired People. International Journal of Applied

Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 7– No. 2, April 2014

9. Sangramsing Kayte , Monica Mundada and Dr. Charansing Kayte. A Marathi Hidden-Markov Model Based Speech Synthesis System. IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov - Dec. 2015), PP 34-39 e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
10. M. Z. Rashad, Hazem M. El-Bakry, Islam R. Isma'il and Nikos Mastorakis. An Overview of Text-To-Speech Synthesis Techniques. ISSN: 1792-4316 ISBN: 978-960-474-207-3 July 2010
11. Sangramsing N. Kayte, Monica Mundada , Dr. Charansing N. Kayte , Dr.Bharti Gawali. Artificially Generated of Concatenative Syllable based Text to Speech Synthesis System for Marathi. IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. II (Nov -Dec. 2015), PP 44-49 e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
12. Pravin M. Ghate and Suresh D. Shirbahadurkar, "Syllable-Based Concatenative Speech Synthesis for Marathi Language". Springer Nature Singapore Pte Ltd. 2019 S. Fong et al. (eds.), Information and Communication Technology for Competitive Strategies, Lecture Notes in Networks and Systems 40, pp 615-624
13. Nilesh Fal Dessai, Gaurav Naik, Jyoti Pawar. : Development of Konkani TTS System using Concatenative Synthesis. International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) 2016 pp-344-348
14. Sai Geeta1 and B.L. Muralidhara.: Syllable as the Basic Unit for Kannada Speech Synthesis. IEEE 978-1-5090-4442-9/17, 2017
15. Sangramsing Kayte , Monica Mundada , Dr. Charansing Kayte. : Di-phone-Based Concatenative Speech Synthesis Systems for Marathi Language. IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 5, Ver. I (Sep - Oct. 2015), PP 76-81 e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
16. Sangramsing Kayte, Kavita Waghmare, Dr. Bharti Gawali.: Marathi Speech Synthesis: A review, International Journal on Recent and Innovation Trends in Computing and Communication, ISSN: 2321-8169-3708 – 3711 June 2015
17. Sangramsing Kayte , Monica Mundada , Jayesh Gujrathi. : Hidden Markov Model based Speech Synthesis: A Review, International Journal of Computer Applications (0975 – 8887) Volume 130 – No.3, November 2015
18. Nikisha Jariwala and Bankim Patel. : A System for the Conversion of Digital Gujarati Text-to-Speech for Visually Impaired People., Springer Nature Singapore Pte Ltd. 2018 S. S. Agrawal et al. (eds.), Speech and Language Processing for Human-Machine Communications, Advances in Intelligent Systems and Computing.
19. Amit Kumar Jha, Piyush Pratap Singh, Pankaj Dwivedi. : Maithili Text-to-Speech System, IEEE 978-1-7281-2472-8/19 2019
20. Suhas R. Mache, Manasi R. Baheti, C. Namrata Mahender, Review on Text-To-Speech Synthesizer, IJARCCE, ISSN (Online) 2278-1021, ISSN (Print)

2319 5940 Vol. 4, Issue 8, August (2015) p. 54 – 59

21. S.D.Shirbahadurkar, D S.Bormane, R.L.Kazi, Subjective and Spectrogram Analysis of Speech Synthesizer for Marathi TTS Using Concatenative Synthesis, International Conference on Recent Trends in Information, Telecommunication and Computing, 978-0-7695-3975-1/10 (2010)
22. Jayesh Tanna, Vijay Savani, Amit Degada, Text-To-Speech (TTS) Conversion System for Gujarati Language, JoEDT, ISSN: 2229-6980 (2013) 1-10, STM Journals 2013.
23. [G. D. Ramteke1 and R. J. Ramteke, Text-To-Speech Synthesizer for English, Hindi and Marathi Spoken Signals, British Journal of Applied Science & Technology 15(3): 1-16, ISSN: 2231-0843, March 2016
24. Das, S. (2023, April 27). Syllabic Speech Synthesis for Marathi Language. 2023 1st International Conference on Cognitive Computing and Engineering Education (ICCCEE). https://doi.org/10.1109/icccee55951.2023.10424461
25. Oyucu, S. (2023, April 18). A Novel End-to-End Turkish Text-to-Speech (TTS) System via Deep Learning. Electronics, 12(8), 1900. https://doi.org/10.3390/electronics12081900