Making Machines Talk In Sanskrit: A systematic exploration Of Text-To-Speech Synthesis For Sanskrit Language

Sabnam Kumari[1], Amita Malik[2]

[1]Computer Science & Engineering Department, Deenbandhu Chhotu Ram University of Science and Technology (DCRUST), Sonipat, Haryana, India, Email: shabnam.schcse@dcrustm.org
[2]Computer Science & Engineering Department, Deenbandhu Chhotu Ram University of Science and Technology (DCRUST), Sonipat, Haryana, India, Email: amitamalik.cse@dcrustm.org

Received: 15.07.2024 Revised: 16.08.2024 Accepted: 09.09.2024

ABSTRACT

Introduction: For more than four thousand years, Sanskrit, one of the most important languages, has served us. Owing to the philological, social, scientific, and pharmacological significances, Sanskrit needs to be preserved. It can only be done, when it will reach to young minds and people will get the resources to learn, read and use this language.

Context: In the last two decades, few researchers have tried to implement Sanskrit Text-to-Speech (STTS) Systems. In this study, we seek to establish the current state-of-art of STTS by comprehensively identifying and analyzing the previous work.

Methods and Procedures: Accordingly, the article represents findings of the Systematic Literature Review (adhered to the model of Kitchenham and Charter) that collected the most relevant literature produced from 2002 to 2024. After the search conducted on 7 databases, the papers were identified as primary studies and were analyzed in detail.

Outcome and Results: The outcome of this systematic literature review (SLR) depicted that there is a lack of Sanskrit Text-to-Speech system (STTS) related work and publications in high-quality journals. It also elucidates the framework for the development and execution of the Sanskrit Text-to-Speech system (STTS).

Conclusion and Implications: This paper examines the possible methodologies, methods, challenges, and limitations of the Sanskrit Text-to-Speech system (STTS) in order to gain a better understanding of the research dynamics in speech synthesis.

Additional Keywords and Phrases: TTS (Text-to-Speech), Systematic review, Natural Language Processing (NLP), Speech Synthesis, Grapheme-to-Phoneme (G2P) conversion, Deep Neural Network (DNN) etc.

Keywords: Sanskrit Text-to-Speech (STTS), Text-to-Speech (TTS), Natural Language Processing (NLP), Grapheme-to-Phoneme (G2P), Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA), Deep Neural Network (DNN), Population, Intervention, Context and Outcome (PICO), Voice Operating Demonstrator (VODOR), Orator VerbisElectris (OVE), MUltichannel Speaking Automaton (MUSA), Non Standard words (NSW), Unit Selection Synthesis (USS), Concatenative Speech Synthesis (CSS), Statistical Parametric Speech Synthesis (SPSS), Hidden Markov Model (HMM), Generative Adversarial Network (GAN), Gaussian mixture models (GMM).

1. INTRODUCTION

Humans' most common mode of communication is speech. Natural languages are used by people to communicate their ideas and thoughts. Natural languages make use of verbal and written elements for this purpose. The written elements correspond to the text and verbal elements correspond to the speech produced. The comprehensible mode of communication is verbal communication. The simplest way to make a synthetic speech is to play pre-recorded samples of authentic speech, such as single words or sentences. The method of transforming any text into conventional, smooth speech in real-time is known as speech synthesis. It is a cross-disciplinary field involving many disciplines such as acoustics (study of mechanical waves), linguistics (study of language), computational linguistics (study of written and spoken language), computer science, mathematics, statistics, signal processing, etc. As a result, Voice Synthesis is a leading-edge information processing technology, particularly for today's intelligent speech interface systems [1].

According to a well-known excerpt of linguist K. Harrison [2], “Languages are the repository of thousands of years of a people's science and art". According to Rachel Dwyer [3], University of London's professor of

Indian Cultures at the School of Oriental and African Studies (SOAS), "You don't open a Hindi film without offering a prayer in Sanskrit to Lord Ganesha for removing obstacles. You don't have a Hindu wedding without it and you don't have a temple without Sanskrit".
According to Oscar Pujol [3], a Sanskrit scholar from Spain, "The third revolution will cover areas like aesthetics, psychology, grammar, computational linguistics, logic, the science of interpretation (mimasa) and study of consciousness". Sanskrit knowledge could be used to tackle some of the challenges that human sciences are facing, such as awareness, consciousness, etc.

1.1. Contributions

- Initially, to determine the importance of the topic, a systematic comparison was conducted to explore text to speech surveys in the literature.
- A comprehensive examination of the current Devanagri-based languages is carried out.
- Some open issues encountered while introducing the Sanskrit Text-to-Speech architecture at various stages.
- By conducting a detailed examination of various Text-to-Speech systems and Sanskrit Text-to-Speech systems, the article intends to fill in the gaps noted in previous studies (STTS).
- Unstructured materials and structured manuscripts released between 2002 and 2024 reveal a slew of primary studies on Sanskrit Text-to-Speech (STTS). These studies can be used by other researchers to further their research.

1.2. Article roadmap

This work covers the history and relevance of Sanskrit as well as Sanskrit research statistics. While discussing the Text-to-Speech system, a comparison is done for various Text-to-Speech systems for Devanagri scripts. The existing systems are compared on the basis of performance parameters. The basics of speech technology and Text-to-Speech system and their architecture are discussed. Finally, the paper examines the obstacles that remain unsolved. Figure 1 depicts the article's timeline.

1.3. Article organization

The following is a breakdown of the entire document: First, the review methodology framed to perform the survey of the literature is explored in section 2. The relevant research questions are framed in section 2.1.2. The background of the Sanskrit Language is discussed in section 3. The global consciousness of Sanskrit followed by its significance is covered in sub-sections of section 3 i.e., 3.1 and 3.2separately. Section 4 summarises the related research, and the methodology adopted for conducting the survey. Section 5 gives a general overview of speech technology. The process of Text-to-Speech Synthesis (TTS) is separated into many phases rather than following a single-phase synthesis rule. Sec. 6 delves into the phases of TTS, as well as its architecture and basic parts. Section 7 speculates on a basic review of several existing synthesis methodologies and procedures as well as the assumptions that underpin them. Their disadvantages and advantages are also addressed.Section 8 delves into the overall discussion of the review. It includes key findings of SLR, its limitations and open issues of the Sanskrit Text-to-Speech system. Section 9 discusses the conclusion and future research avenues.

2. Review methodology

A review methodology is utilized to justify the research gap and highlight the incentive component of completing the survey. The current article's systematic review method is depicted in Figure. 2. The review method is separated into four parts, which are described below, Review Planning, Literature Search, Inclusion/Exclusion Criteria, and Quality Assessment.

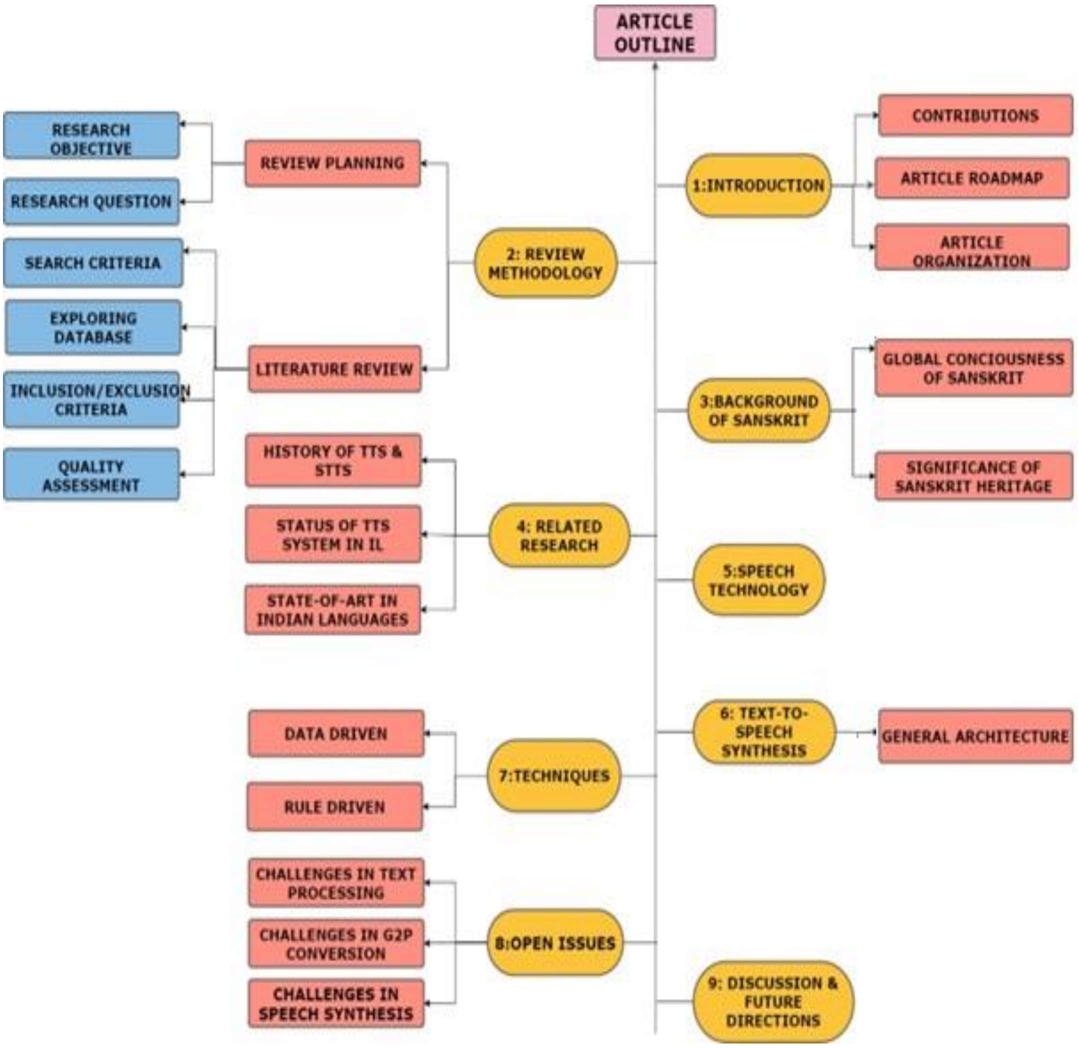

Figure 1: Article Roadmap

2.1. Review Planning

The review planning phase includes the objectives of research and formulating the research questions.

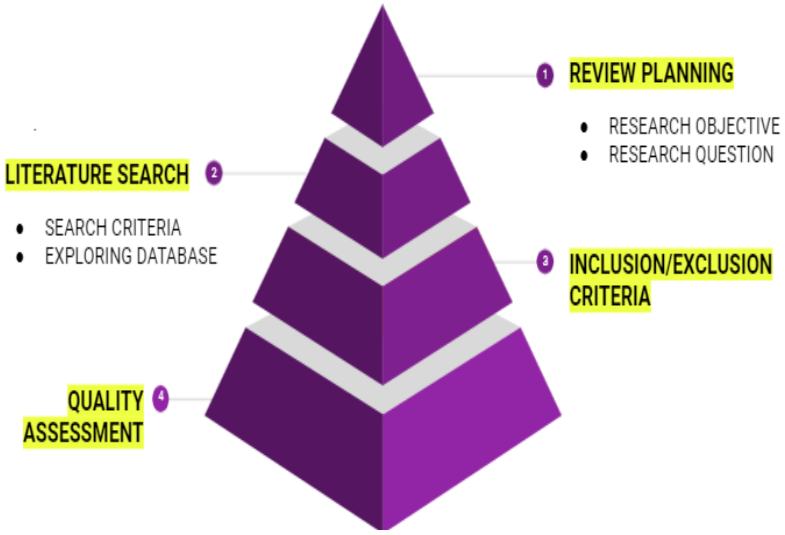

Figure 2: Review Methodology

2.1.1. Research Objective

This paper's main intent is to give a comprehensive and complete review of the literature on Sanskrit Text-to-Speech synthesis. Several critical points have been established to create an effective Sanskrit Text-to-Speech Synthesizer based on this goal.

2.1.2. Research Questions

A series of seven research questions based on the PICO technique (Population, Intervention, Context, and Outcome) [4-5] are constructed to meet our research goal. Table 1 shows the criteria in detail. The following are the research questions:
RQ1: What are the purposes of recent research while developing a Sanskrit Text-to-Speech system? Section 4.1.1 provides an answer to this question.
RQ2: Which existing STTS systems are to be considered while redeveloping an efficient one? Section 4.1.2 provides an answer to this question
RQ3: Which approaches are available in research related to speech synthesis? Section 7quick fixes answer to this question.
RQ4: What variables contribute to a successful Text-to-Speech system? Section 6 provides an answer to this question.
RQ5: How can the effectiveness of a text-to-speech technology be measured?The answer to this question is rescript in Section 6.
RQ6: What are the difficulties in putting a Sanskrit Text-to-Speech system in place? Section 8 provides an answer to this question.
RQ7: What are the current trends and research directions? The answer to this question is rescript in Section 9.

Table 1: PICO criteria

CRITERIA	MEANING	DESCRIPTION
1. Population	It is a field of application.	Literature in Text-to-Speech systems for Sanskrit language
2. Intervention	It covers the approach, framework, technology etc. that addresses a particular issue.	Techniques, tools and approaches for Speech Synthesis and Sanskrit TTS.
3. Context	It is the setting or framework in which the intervention is delivered.	The elements and architecture of Text-to-Speech synthesizer
4. Outcome	It needs to detect elements that are important to researchers.	Challenges and major gaps of research in Sanskrit Text-to-Speech System.

2.2. Literature Search

Literature search includes the search criteria and explorationof the databases.

2.2.1. Search Criteria

The precise keywords were chosen by hand in order to achieve the desired result, which would then help us answer our study questions. As a result, the words for the search string are derived from research questions based on population, intervention, and outcome (PICO). The current article contains literature review of quantitative and qualitative research articles during last 20 years from 2002 to 2021 which are written in English language. The article selection process was guided by Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) [6].The process of selecting articles is depicted in a flow chart given in Figure 3.Some criteria taken into consideration while making the search string are:

- Keywords are derived from research questions, appropriate publications, and books.
- Synonyms and alternate spellings are identified.
- For synonyms and abbreviations, use the Boolean operator OR.
- Connecting unrepeated search phrases with the Boolean operator AND.

A resulting search string is as follows:
("text to speech" OR "TTS" OR "Text-to-Speech Synthesis" OR "speech synthesis") AND ("Sanskrit" OR "Indian languages" OR "Indian language"OR "Devanagri")
Slight adjustments to search string based on the database are done.

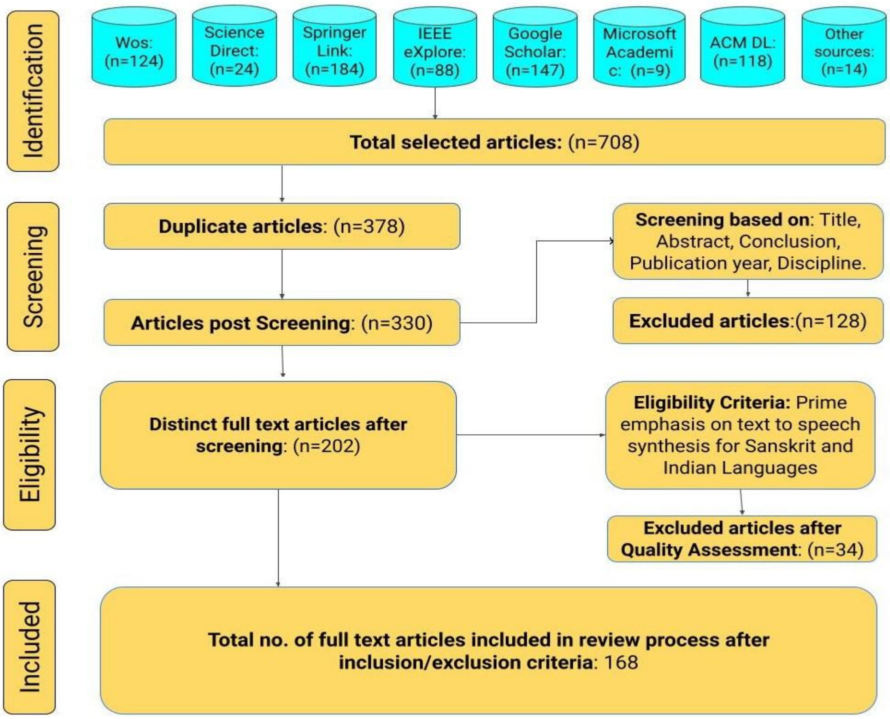
Identification
Wos: (n=124)
Science Direct: (n=24)
Springer Link: (n=184)
IEEE eXplore: (n=88)
Google Scholar: (n=147)
Microsoft Academi c: (n=9)
ACM DL: (n=118)
Other sources: (n=14)
Total selected articles: (n=708)
Screening
Duplicate articles: (n=378)
Screening based on: Title, Abstract, Conclusion, Publication year, Discipline.
Articles post Screening: (n=330)
Excluded articles:(n=128)
Eligibility
Distinct full text articles after screening: (n=202)
Eligibility Criteria: Prime emphasis on text to speech synthesis for Sanskrit and Indian Languages
Excluded articles after Quality Assessment: (n=34)
Included
Total no. of full text articles included in review process after inclusion/exclusion criteria: 168

Figure 3: Selection process of articles

2.2.2. Exploring the Database

Most known databases, such as IEEE eXplore, Web of Science, Science Direct, Google Scholar, Springer Link, ACM DL, and Microsoft Academic, were utilized to search for relevant articles. The distribution of selected papers as per these digital libraries is given in Table 2. These databases included reputed journals, Conferences, Lecture Notes, and Book Chapters. Few preprint versions of the research papers have also been included. Some official websites are also consulted. Table 3 presents the search sources.

Table 2: List of digital libraries

Digital Library	COUNT OF PAPERS	URL
IEEE eXplore	78	http://ieeexplore.ieee.org
Web of Science	124	https://mjl.clarivate.com
Science Direct	24	https://www.sciencedirect.com
Google Scholar	147	https://scholar.google.com
Springer Link	284	https://link.springer.com
ACM DL	118	https://dl.acm.org
Microsoft Academic	9	https://academic.microsoft.com

Table 3: Search sources

Databases	Searched Items	Search applied on
Web of Science	Research articles	Keywords, title, abstract
Science Direct	Research articles, Article-in-press	Keywords, title abstract
Springer Link	Research articles, Book Chapter, Conference proceedings, Lecture Notes	Full text, discipline
ACM DL	Research articles, Book, Book Chapter, Conference Proceedings	Keywords, Title, Author, Abstract
IEEE eXplore	Research article	Full Text
Google Scholar	Research articles, preprints	Keywords, Title, Author
Microsoft Academic	Research articles, Conference Proceedings	Title, Author

2.2.3. Inclusion and Exclusion criteria

The application of inclusion and exclusion criteria establishes the borderline for the literature review process. This phase is a notch on the scale of quality, as the filtered articles are going to write back answers to the research questions. Therefore to avoid selection bias and improve the feasibility of the study, the Inclusion/Exclusion criteria should be agreed upon and formalized. Figure 4 depicts the Inclusion/Exclusion criteria.

Language
Publication date
Inclusion Criteria
Subject
Mode of availablity
Article type
Paper length
Before publication date
Subject Mismatch
Exclusion Criteria
Language Mismatch
Article status

Figure 4: Inclusion/Exclusion Criteria

Inclusion Criteria

IC1: Subject: Papers concentrating on one or more components of the Text-to-Speech synthesis analysis framework for Sanskrit and other Indian languages.
IC2: Publication date: Between 2002 and 2021, research papers, book chapters, and article-in-progress are available in electronic form (both inclusive).
IC3: Article type: Full text and review papers are available to read.
IC4: Mode of availability: Papers with access to digital files.
IC5: Language: English

Exclusion Criteria

EC1: Subject mismatch: Studies that do not focus explicitly on the Text-to-Speech Synthesis and studies that discuss the synthesis of foreign languages are not included.
EC2: Status of articles: Retracted and duplicate articles are excluded.
EC3: Paper length: Very short papers of less than 4 pages are not included.
EC4: Date of publication: Papers published before 2002 are not considered for literature review.
EC5: Language: Paper that is not written in the English language.

2.2.4. Quality assessment

The step of assessing the quality is pertinent in systematic literature review as it verifies and validates the standards of review. We design some set of questions for quality assessment which are as follows:
QA1: Is the objective of the research clearly stated?
QA2: Does the research analyses the gaps that need to be worked upon?
QA3: Does the research performed theoretical and experimental review?
QA4: Does the research talk about the challenges of the problem?
QA5: Whether the research is published in peer-reviewed journals or conference proceedings or indexed book series?
QA6: Is the research precisely and comprehensively accounts for the methodology used?
QA7: Does the research explore the future scope of the work?
Table 4 represents the checklist used to assess the quality of research. The scores for all the articles were calculated by the assessment criteria given in Table 4. The QA1 analyses the purpose of the research. It was answered effectively in 94% of the articles. The QA2 analyzing the research gap was answered efficiently in 45% of the studies. The QA3 gets the measure of review and the response of 58% of papers as appropriate.

Table 4: Quality assessment criteria

S.no	Assessment criteria	Score	Description
QA1	Is objective of the research clearly stated?	+1	If objectives are clearly stated.
		0.5	If objectives are not well stated.
		-1	If objectives are not stated at all.
QA2	Does the research analyses the gaps that need to be worked upon?	+1	Properly analysed and listed the gaps.
		0.5	The gaps are listed in vague manner.
		0	No gap is listed.
QA3	Does the existing research performed theoretical and experimental review for effective contribution?	+1	If constructive contribution to research.
		0.5	If partial contribution to research.
		0	If no contribution to research.
QA4	Does the research talk about the challenges of the problem?	+1	If addresses the major challenges.
		0	If address only one challenge.
		-1	If no challenge is addressed.
QA5	Whether the research is published in peer-reviewed journals, conference proceedings, indexed book series and symposium?	+1	If journal is of Q1 & Q2 rank and conference/symposium is of A rank.
		0.5	If journal is of Q3 rank and conference/symposium is of B rank.
		0	If journal is of Q4 rank and conference/symposium is of C rank.
QA6	Is the research precisely and comprehensively accounts the methodology used?	+1	If methodology is clearly defined with validation and performance parameters.
		0.5	If partial components of methodology are described.
		0	If no methodology is described.
QA7	Does the research explore the future scope of the work?	1	If clearly specifies the scope for further work.
		0	If forthcoming perspective is stated in vague manner.
		-1	If no further work is discussed.

The QA4 evaluates the quality of challenges addressed in the paper and the outcome is 15%. The QA5 checks the rank of publication platforms and came up with a 92% response. The QA6 checks the approaches, dataset, and strategies to be used for implementation and it resulted in 87% response. The QA7 analyses the further perspectives of the work and comes out with a 97% response. Figure 5 depicts the percentage of scores of distinct quality assessment questions of the primary research articles.

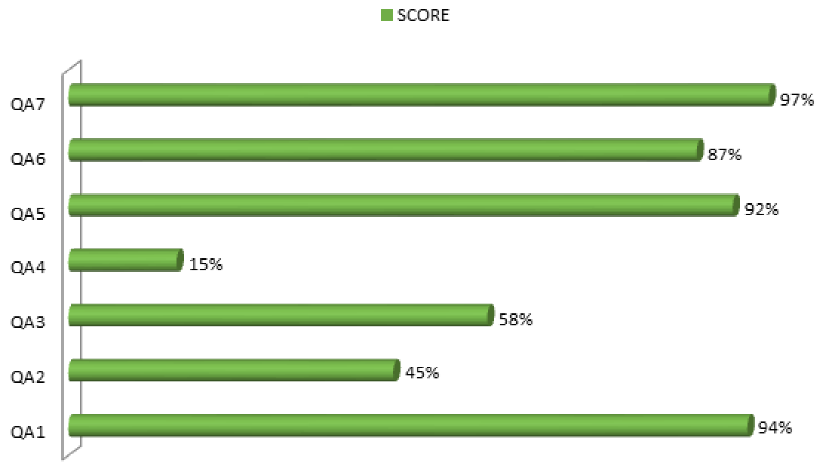

Figure 5: Percentage score for quality assessment questions

3. Background of Sanskrit

Sanskrit is defined by S. Ramkumar [7] as "Sans," which means "whole completion," and "Krit," which means "completed item or work." Amitabh [8] understands that the word Sanskrit is made out of the prefix sam, which means "completely," and the root krit, which means "done." As a result, we can redefine the name Sanskrit as Samskrita, which means "created together in embellished form." Sam (together) and Krita (done) signify "made together in embellished form." It is well-known for its beauty and clarity. It is also acknowledged as "Dev Vani" (god's language) because it is stated that Brahma taught it to the Sages of the celestial bodies. Sanskrit is Hinduism's principal sacred language, and it has been employed throughout Hinduism, Buddhism, and Jainism as a philosophical language [9-10]. Sanskrit, according to Paul [11], is a complete language that lacks any structural ambiguity, both semantically and phonetically.

3.1. Global consciousness of Sanskrit

Sanskrit influenced not only other Indian languages, such as Hindi, which is now one of the country's official languages, and Indo-Aryan languages (Indic languages) like Kannada and Malayalam, but also international languages. According to a survey by the United Nations Organization (UNO), Sanskrit has affected about 97 percent of the world's languages [12]. With the effect of Buddhist teachings in Sanskrit and their translation and diffusion, Nikul Joshi [13] admits the importance of Sanskrit on Sino-Tibetan languages. Telugu is regarded to have a significant lexical similarity to Sanskrit, from which it has borrowed many words.Several Sanskrit loanwords may be found in the vocabulary of many Indian and South Asian languages, including Hindi, Japanese, Tagalog (a Philippine language), Malay (a Malaysian language), and Indonesian [14]. Table 5 shows some examples of Sanskrit-derived words and their meanings in Malay, Indonesian, and Tagalog languages. Sanskrit is the source of several English terms. Table 6 displays some examples of Sanskrit borrowing terms in English. Sanskrit has also had an impact on the Chinese language and many European languages [13][15]. Sanskrit has affected and enriched the languages of Southeast Asian countries such as Thailand, Korea, Mongolia, Kazakhstan, and Uzbekistan.

Table 5: Loanwords of different languages from Sanskrit

Sanskrit	Malay	Tagalog	Indonesian	Hindi	Meaning
Guru	Guru	Guro	Guru	Sikshak	Teacher
Nama	Nama	Nama	Nama	Naam	Known as
Mukha	Muka	Mukha	Muka	Mukh	Face
Sakshi	Saksi	Saksi	Saksi	Saakhi	Witness
Katha	Kata	Katha	Kata	Katha	Fiction
Agama	Agama	Agham	Agama	Dharm	Religious text
Budhhi	Budaya	Budhi	Budi	Budhhi	conscience
Antara	Antara	Antala	Antar	Antar	In mean time

Table 6: Loan words of English from Sanskrit

Sanskrit	English	Meaning
Mala	Mal	Malicious or malnutrition
Pratishat	Percentage	Every hundred
Kafa	Cough	Mucus
Danta	Dental	Teeth
Avtar	Avtara	Icon or figure
Mantra	Mantra	Hymns or slogan
Sam	Same	Identical or equal
Tri	Three	Decimal number 3

3.2. Significance of Sanskrit Language

Puranas are no less than an encyclopedia of knowledge. Among many branches of knowledge that have developed in the modern age is Horticulture, which is related to Plant Science.

Puranas contain significant and relevant information on various aspects of horticulture. Nowadays researchers like D.V. Dwivedi [16-18], R. Joshi [19] are finding new techniques by studying plant taxonomy, plant morphology, plant physiology, plant propagation, medicinal views of plants, etc. as depicted in Sanskrit texts i.e. Puranas, etc.

In agriculture, irrigation plays an important role. In Satapatha Brahmana (a commentary on Sukla Yajurveda, there are numerous hymns pertaining to Parjanya (a water resource) in Vedic Samhitas. Also in Rigveda, various practices are mentioned for water to be used in irrigation. Today, researchers [20] are elucidating new and modern practices of irrigation with the help of Sanskrit heritage.

In Sanskrit texts, Ayurveda is described as the science of life. In the adverse condition of COVID-19, Ayurveda has come up as a savior to millions of people. The recent clinical researches [21]for treating

COVID-19 are based on developing medicines from plants like Ashwagandha [21], etc. researchers are consulting Ayurveda to analyze the various medicinal properties of these plants like alexipharmic (having the quality or nature of an antidote to poison), aphrodisiac, astringent, deobstruent (a medicine that removes obstructions), diuretic (tending to increase the excretion of urine), hypnotic (tending to produce sleep), and sedative (tending to calm, moderate or tranquilize nervousness or excitement), restorative and tonic.

Sanskrit has a significant role in the study of mathematics, science, agriculture, is helpful in speech therapy, and improves imagination and retention power.According to Vaishali etc [22] has widesignificance in both technology and human health. Top universities in Germany (around 14 universities), Australia (around 8 universities), the USA (around 3 universities), the UK (around 2 universities), teach Sanskrit. There is a department at NASA dedicated to studying Sanskrit manuscripts. Sanskrit is, in fact, the only "Unambiguous Spoken language on the planet," according to NASA.

4. Related Research

4.1. History of TTS and Sanskrit TTS

Zoe Handley[23] identified operational context of Text to Speech synthesis as three different roles: reading machine, pronunciation model, and conversational partner. For more than decades, scientists [24] [25] have been working on developing natural voice, human-like speech through Computer-Assisted Language Learning (CALL), Computer-Assisted Pronunciation Training (CAPT), automatic speech recognition and many other models. VOCODER was the first speech processing device developed at Bell Laboratories between 1930 and 1935, and it was used to extract speech's slowly changing acoustic properties. The extracted acoustic features/parameters were used to run the synthesizer. Homer Dudley, a scientist at Bell Labs in New York, created the Voice Operating Demonstrator (VODOR) in 1939-40, an elementary machine producing a synthetic speech that was manipulated like a piano keyboard, but instead of music, it produced a squawking mechanical voice. This work strongly drew strongly the attention of research community towards speech synthesis. A coeval development introduced by Gunnar Fant is the first cascade formant speech synthesizer known as Orator Verbis Electric (OVE I) [26].

At the Electro-Technical Laboratory in Japan, Noriko Umeda created the first broad English text-to-speech technology in the late 1960s. An early prototype of a Speech Synthesis machine, MUSA (MUltichannel Speaking Automaton), was released in 1975. A diphone synthesis process was used. It was one of the earliest TTS systems to work in real-time. It was made out of stand-alone computer hardware and customized software that allowed it to read Italian in a robotic voice that was understandable. DECTalk, a computerized TTS application developed by the Digital Equipment Corporation in Massachusetts, had progressed to the point that the late Stephen Hawking could use a version of it with a keyboard to “talk” in the 1980s. The end result was a set of comprehensible but synthetic-sounding words that the humans still associate with a talking machine. Speech synthesis that was more accurate became more widespread in the early 2000s. Hybrid unit concatenation was the most popular technique discussed. However, starting in 2010, Siri, Apple's debut in the sector of speech recognition that captured the public's attention first. Thanks to decades of research, Artificial Intelligence (AI) powered automated personal assistants added a human touch to the otherwise impersonal world of speech recognition. Following Siri, Microsoft released Cortana, and Amazon released Alexa, kicking off the ongoing competition for supremacy among the tech giants' voice recognition platforms [27]. There has never been a venue for the Sanskrit language. Mr. Diwakar Mishra, a research scholar at JNU in New Delhi, developed the ‘Samvacaka' technique for Sanskrit Text-to-Speech (STTS) Synthesis in 2013. This Sanskrit Text-to-Speech Synthesis model is still functional, although only to a limited extent. Then, in 2018, S. Chandrasekhar created a new Sanskrit Text to Speech technology, which is now unavailable. Figure 6 depicts the entire timeframe.

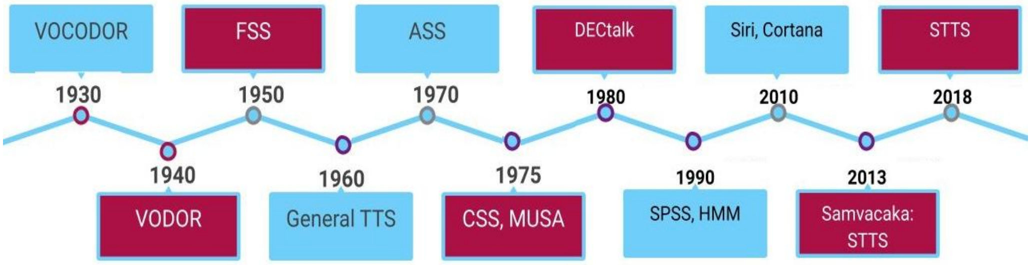

Figure 6: Timeline of Speech Synthesis

4.1.1. Status of TTS systems in Indian Languages

The prehistoric 'Brahmi' writing system gave rise to Indian linguistic scripts. The essential units of the Indian language writing system are 'Aksharas' (Letters), which are syllabic in character and orthographically represent a speech sound. There are 15-18 vowels and 35-38 consonants in all Indian languages. [28]. For each language, the most commonly occurring syllables are rarely more than 300. More than 120 languages in the world use the 'Devanagari writing system. According to the 8th schedule of the Indian Constitution, 22 official languages are recognized. Sanskrit, Hindi, Konkani, Kashmiri, Nepali, Maithili, and Marathi are among India's 22 official languages that use the Devanagari alphabet. Despite sharing a phonetic background Telugu, Tamil, and Kannada have their separate scripts.

TTS for thirteen (13) official languages of India, including Hindi, Assamese, Bengali, Marathi, Odia, Telugu, Rajasthani, Gujarati, Tamil, Bodo, Manipuri, Kannada, and Malayalam is being developed as part of the Consortium Mode Project (CMP), led by Indian Institute of Technology (IIT) Madras, and employs both the open-source FESTIVOX Framework and a cutting-edge HMM-based speech synthesis system (HTS)-based engine. 11 other consortia membersare funded by the TDIL Programme, MEIT, Government of India. Table 7 covers the research activity on Text-to-Speech (TTS) for Indian Languages.

India has a diverse language population. Many of these Indian languages demonstrate orthography and the mother of all these languages is Sanskrit [29]. However, the work for Sanskrit synthesizer is very limited. The major systems developed till now for various Indian Languages are listed in Table 7.

Majority of works done for Indian language TTS system include research conducted by S.P. Panda et al [30], Shiga Yoshinori et al [31], Sheilly Padda [32], Amitoj Singh et al [33], Yishuang Ning et al [34], Sarang Joshi et al [35], Pulkit Sharma et al [36], Sanjay Kakodkar et al [37], NeeteshVashishtha [38], Kinjal v Patel [39], Devyani Kulkarni et al. [40], Jitender pokhariya [41], Chandra SekharamBondu et al. [42], Diwakar Mishra [43-44]. Table 8 concludes all the above-mentioned work.

Table 7: Major systems developed till now on various Indian Languages

System and year of implementation	Implemented at	Methodology used	Language Worked Upon
Acharya (2002)	IIT MADRAS	Rule driven	Telugu
Vani (2004)	HYDERABAD CENTRAL UNIVERSITY (HCU)	MBROLA, Diphone Synthesis	Telugu
LRTC (2010)	IIIT Hyderabad	Festvox, Data driven	Hindi, Sanskrit
Samvacaka (2013)	JNU, New Delhi	Festvox, Data driven	Sanskrit
Thirukkural&Vaachaka	IISC BANGALORE		Tamil and Kannada
RC-ILTS-ORIYA	UTKAL UNIVERSITY, ORISSA	Concatenation technique	Oriya
TTS Hindi	C-DAC, NOIDA	Concatenation technique	Hindi
TTSConsortium (for 13 Indian languages)	IITM (Indian Institute of Technology, Madras)	Statistical parametric approach, HMM	Telugu, Gujarati, Assamese, Bodo, Bengali, Malayalam, Hindi, Kannada, Odia, Manipuri, Rajasthani, Marathi, and Tamil

Table 8: Literature Review on Sanskrit TTS

Year & publication	Author	Title	Summary
2020, Springer [30]	Panda S.P., Nayak A.K.	A survey on speech synthesis techniques in Indian Languages	Addressed the characteristics of Indian languages, challenges and techniques of TTS.
2020, Springer [31]	Shiga Yoshinori et al.	Text-to-Speech Synthesis	Explained the workings of a cutting-edge TTS system and provided a quick overview of other traditional speech synthesis technologies, including their benefits and drawbacks.

2020, IRJET[32]	Padda Sheilly et al.	Review of various Text-to-Speech Synthesis Methodologies	Discussed various ideas to recognize the text character and convert it into speech signal along with some applications of TTS systems.
2019, Springer [33]	Singh Amitoj, Kadyan Virender	ASRoIL: A Comprehensive Survey for Automatic Speech Recognition of Indian Languages	Concluded that most of studies use HMM-GMM classifier and HTK tool kit.
2019, Applied Science (MDPI) [34]	Ning Yishuang, Sheng He	A Review of Deep Learning Based Speech Synthesis	The difficulty of deep learning for speech synthesis was discussed, as well as several promising research directions..
2019, Helix [35]	Joshi Sarang, bairagi Vinayak K.	Recent Trends in Text-to-Speech Synthesis of Indian Languages	Prosody is important in producing natural speech, according to the researchers. As a discipline, the study of Indian language prosody is still in its early stage.
2018, Elsevier [36]	Sharma Pulkit, Abrol Vinayak	Reducing Footprint of Unit Selection based Text-to-Speech System using Compressed Sensing and Sparse Representation	Compressed sensing measures and the coefficient of Sparse Representation are used to compress the vocabulary learned for each phonemes..
2018 IJCA [37]	Kakodkar Sanjay, Borkar Samarth	Acoustics Speech Processing of Sanskrit Language	Using k-NN and SVM algorithms prepared the system which worked for customized database only with less accuracy.
2017, IJACEE [38]	VashishthaNeetesh	Implementation of Sanskrit Linguistics in Artificial Intelligence Programming	Expressed role of Sanskrit language and its capability in AI programming for medium of interaction.
2017, 4ICMRP [39]	Kinjal patel	Sanskrit: Some Insights as a Computer Programming Language	Presented some technical strengths of Sanskrit as a programming language.
2016, IJRET [40]	Kulkarni Devyani et al.	HTK Based Speech Recognition Systems for Indian Regional Languages: A Review	Gave a quick review of HTK toolkit-based automatic voice recognition systems for Indian regional languages.
2014, IJERT [41]	Pokhariya Jitender, Mathur Sanjay	Sanskrit Speech Recognition using Hidden Markov Model Toolkit	Build a LINUX based Sanskrit speech recognition system with 50 utterances only.
2014, IJCNWMC[42]	C. Bondu , S. Rama Krishna	An Approach for Grapheme to Phoneme Alignment for Sanskrit TTS	Discusses important issues relating to Grapheme to Phoneme alignment for Sanskrit text written in Devanagari script.
2013, IEEE Conference [43]	Mishra Diwakar, Bali Kalika	Syllabification and Stress Assignment in Phonetic Sanskrit Text	The presented Grapheme to Phoneme (G2P) converter programme translates Devanagari UTF-8 Sanskrit text into its phonetic representation with set syllable boundaries and syllable stress values.
2011, Springer [44]	Mishra Diwakar, Bali Kalika	Challenges in Developing a TTS for Sanskrit	At JNU's Special Centre for Sanskrit Studies, Diwakar Mishra presented ongoing research on the Sanskrit TTS system known as "Samvacaka." It emphasizes on the development of various TTS System modules as

			well as potential problems in this field.

4.1.2. State-of-the-art Sanskrit TTS

What distinguishes a Sanskrit Text to Speech System from those for other Indian languages Text-to-Speech System is not its methodology, but its language-specific aspects. Before preparing speech database and linguistic analysis tools, the basic requirement is the study of the so-called 'correct' Sanskrit pronunciation (shuddhauccharana), which is one of the major concerns of the Sanskrit community. Sanskrit is not the first language of most of the people who use it, so even they do not use it as speech, therefore, the study of their speech samples could not give authentic conclusions. Many researchers have worked in this direction with limited features implemented [43-47]. Few researchers have put their efforts and developed working application interface for Sanskrit are (Table 9 concludes Sanskrit Text-to-Speech systems):

- Acharya (2002) [45] is a Sanskrit TTS done at IIT Madras. It also synthesized Tamil, Telugu, Hindi and Sanskrit.
  Limitations: It is not compatible with Unicode. Also, it is not accessible at present.
- Vani – An Indian Language TTS Synthesizer (2004) [46] is a speech synthesis project in IIT Bombay. It works for Hindi, and it reads text encoded in iTrans. Limitations: Only limited Sanskrit text is also considered. It doesn't work for Non-Standard Words (NSW).
- LTRC, IIIT Hyderabad (2010) [47] is a project of IIIT Hyderabad. It is based on Hindi TTS and considered sentences of normal conversation. It is a prototype of a Sanskrit speech synthesis system, not a full version TTS synthesizer. Limitations: It does not claim that proper care has been taken for the coverage of phones in different contexts while preparing the data. It does not deal with Non-Standard Words (NSW).
- Samvacaka - Sanskrit vacaka (2013) [43-44] at JNU, Delhi. Diwakar developed his own corpora and used the Festival Speech Synthesis framework using the unit selection method. Limitations: It does not read Vedic text with accent marks. It cannot recite the verses (slokas). The modules to predict prosody – pause, phone duration, and pitch – have not been developed. Uncovered category of NSW: Roman numerals, Fractions, Ratio, URL, Number range, Percentage, Alphanumeric String. Text Normalizer is poor for blind people.
- Hear2Read (2016) [48] is s mobile phone application based on Andriod. Limitations: Prosody prediction is missing. Visarga is not heard clearly. No chanting of hymns. Poor voice quality. The spelling of copied text gets changed and gives wrong results. It works only on Android OS.

Table 9: Working interface (system) for Sanskrit TTS

System and year of implementation	Implemented at	Methodology used	Working now or not?
Acharya (2002)	IIT Madras	MBROLA	Not working
Vani (2004)	III Bombay	MBROLA, Diphone segmentation approach	Not working
LRTC (2010)	IIIT Hyderabad	FestVox, Data driven approach	Not working
Samvacaka (2013)	JNU, New Delhi	FestVox, Data driven approach	Not working
Hear to Read (2016)	Sri Aurobindo University	Festival, F-lite.	Working as Android application

5. Speech Technology

Speech is the most fundamental and important aspect of human communication. An individual can divulge their emotions, feelings, thoughts or we can say the state of mind through speech. The main purpose of this study is to develop such machine with an understandable, comprehensive, accurate, and natural-sounding voice that can transfer knowledge to end-users in the voice of their choice, in Sanskrit language and accent. Sanskrit has three accents i.e. udātta (raised or high pitch), the anudātta (not raised or low pitch) and the svarita (sounded or combination of low and high tone) [49]. The research in the domain of Sanskrit speech synthesis spans a wide range of disciplines, including phonetics (generation of voice and insight) to morphology (the pronunciation: phonemes, stress, intonation, duration) and syntax (parts of speech: pauses, rate of speaking, emphasis) to speech processing [50]. Thus, the intrinsic next move is to make a machine to talk in Sanskrit just like human beings. Speech Technology consists of two key components: Speech Synthesis and Speech Recognition. Figure 7 shows the categorization of Speech Technology [51].

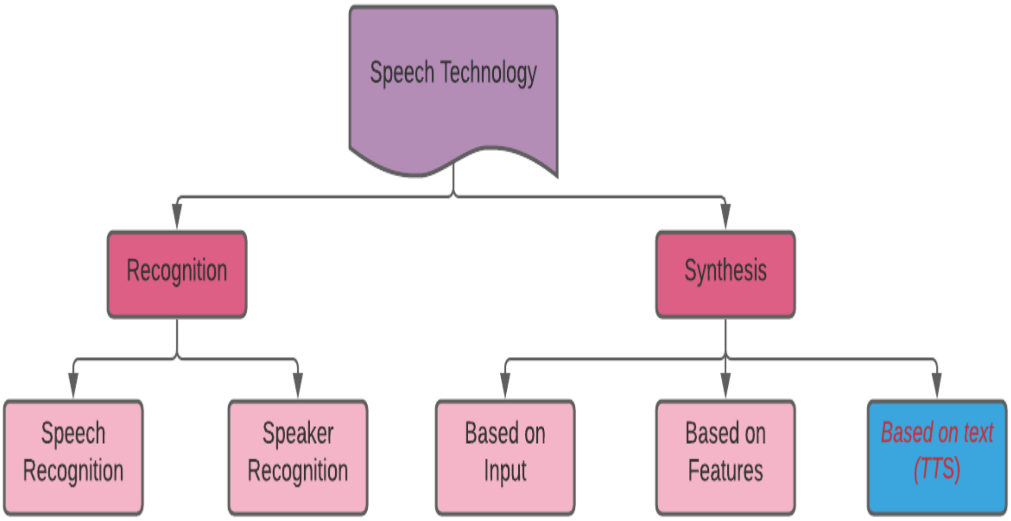

Figure 7: Components of Speech Technology

Recognition is the process of transforming speech signals into words/phrases with specific order by the implementation of an algorithm and executing as a program. It incorporates the machine accepting the individual's speech and illustrating what has been said. Recognition is based on speech recognition and speaker recognition.

Synthesis is the process of a computer producing an acoustic voice signal. It is an artificial simulation of an individual's speech by a machine. Concatenation of waveforms is used to convert written information into spoken speech [52].Synthesis is done on basis of input, features and text.

6. Text-to-Speech

A Text-to-Speech (TTS) system works by giving the machine a Sanskrit text to utilize as a background for generating utterance acoustic speech parameters. The automatic examination of the structure of Sanskrit words into their morphological parts is used in text-to-speech synthesis [53]. Any Sanskrit text can be transcribed into speech by mixing the pronunciations of those sub-word units according to letter- and morph-to-sound principles. Because the developed system can handle open-ended Sanskrit text, it's suited for large-scale applications including reading text aloud to blind users, reciting mantras and shlokas for those who don't know Sanskrit, and reading electronic mail over the phone, among others [54]. The quality of synthesized speech is required for the Sanskrit TTS system to be validated. Three metrics are commonly used to quantify efficacy and excellence: intelligibility (the ability of the listener to grasp the message), comprehensiveness (contextual intelligibility), and naturalness (the quality of the generated speech is closer to actual human speech) [55-57].

Intelligibility: It relates to the system's ability to produce audio and the ease with which the speech can be comprehended. How properly can the listener reconstruct the speaker's speech? [55] It disseminates the first-rate audio generated. Is it cleaner? Is it listenable?

Comprehensiveness: It refers to the ability of the listener to understand the message. Alternatively, we can say that it measures the degree of received messages being understood. How well a listener is able to interpret the meaning of speech? Comprehensiveness can be restated as contextual intelligibility.

Naturalness: It refers to a belief that the synthesized speech is moreclose to a human-generated speech. It represents the first-rate of the created speech. Does it come out as emotionless? Is the speech timed, structured, and pronounced correctly??

In addition to intelligibility, comprehensiveness and naturalness, there are a few other evaluation criteria that influence the quality of a text-to-speech system. Expressivity, speaker resemblance, likeability are included among the criteria. The ability to modify one's speaking style is referred to as expressivity. [58]. Is the speaker able to express happiness, sadness, or surprise feelings? The appeal perceived by the listener is known as likeability. Is it pleasant? Is it delightful? Is it miffy or annoying? [59]. Speaker similarity refers to the fact that the voice signal provides information about the speaker's anatomy, physiology, linguistic experience, and mental state at all levels of the message production process, allowing the speaker's identity to be heavily reflected through speech. [60]. Table 10 gives the description of criteria to evaluate the effectiveness of Text-to-Speech Systems.

Table 10: Criteria of evaluation of Text-to-Speech Systems.

CRITERIA	DESCRIPTION
Intelligibility	Is it listenable? Is it clean?
Comprehensiveness	How well a listener is able to interpret meaning of speech?
Naturalness	Does it sound like humans?
Expressivity	Is it express the speaker's feelings e.g. happiness, surprise etc.?
Likeability	Is it pleasant? Is it annoying?
Speaker Similarity	Expression of individuality.

6.1. General Architecture of TTS

The general architecture is a blueprint consists of general components and sub-components developed that will perform collectively to implement the overall system. The architecture of the Text-To-Speech System involves three main phases: Text Processing, Grapheme to Phoneme Conversion, and Speech Synthesis.

6.1.1. Text Processing:

The first component of the Text-To-Speech System is text processing. Its major task is to take any text as input and convert it into a format that may be used for further linguistic processing.

a) Text Normalization

The text normalization process, as the name implies, converts the input text into a format that is acceptable for the context. Tokenization module is another name for it. Sentence tokenization and nonstandard word substitution (NWS) are the two main tasks of the Text Normalization Module. Sentence tokenization is a method of segmenting input material into sentences by exploiting the punctuation in the text. Text strings that must be turned into words for speech, such as numbers, acronyms, and abbreviations, are dealt with using nonstandard word substitution.

b) Phonological analysis:

The acoustic analysis includes the phonological examination of the normalized text obtained from the text analysis step. It involves grapheme-to-phoneme conversion, which converts orthographical symbols to phonological symbols. In TTS systems designed for languages where words with the same spelling are read differently depending on their semantics, homograph disambiguation is critical.

6.1.2. Grapheme to Phoneme Conversion

Grapheme to phoneme conversion refers to the process of generating pronunciation for words based on their written form. Every word is converted to its phonetic representation, which includes syllable boundaries and stress information. It takes standard words as input and converts them into a corresponding sequence of phones

6.1.3. Speech Synthesis

The task of speech synthesis includes the generation of an acoustic waveform, reducing various signal level discontinuities, incorporating prosodic features at signal level [38].

a) Prosodic analysis:

Prosodic analysis entails examining several prosodic qualities and the voice synthesis technique will provide synthesized speech with certain features. It is concerned with the physical characteristics, physiological production, acoustic features, auditory perception, and neurophysiological status of speech sounds or signs. [61]. Prosody is a subset of human conversation having suprasegmental characteristics. Since prosody is a suprasegmental feature, language units(character, phoneme, word etc)become important for prosody generation. Masashi Aso et al [62] haved discussed about capturing and prediction of suprasegmental features.

b) Speech Generation

It produces acoustic output for given text input. It is an artificial simulation of an individual's speech by a machine. Concatenation of waveforms is used to convert written information into spoken speech. It entails utilizing a speech synthesis approach to construct the final speech waveform with respect to the phonological features identified during the analysis phase. Figure 8 depicts the levels of text-to-speech synthesis, with specifics on each phase discussed further below [28]:

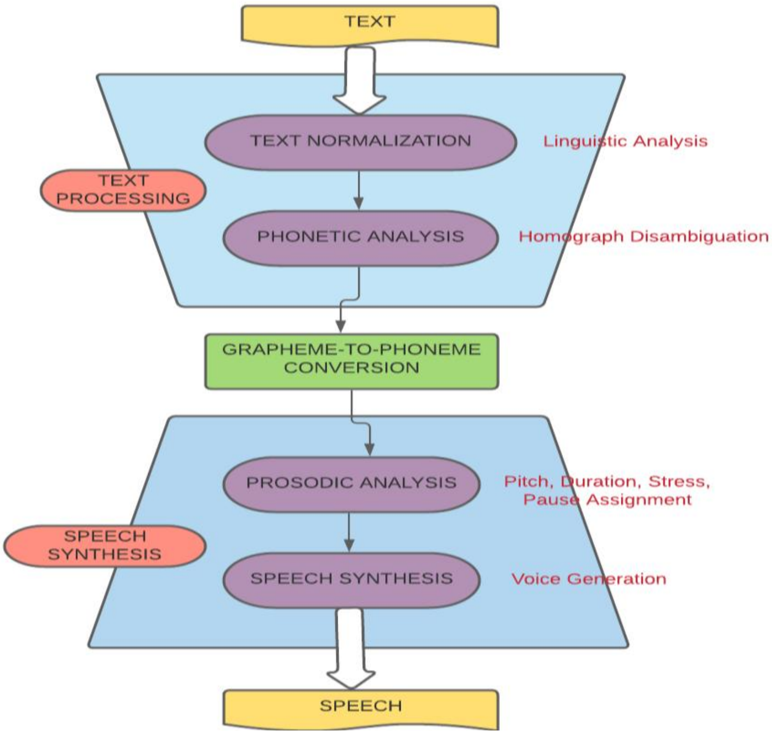

Figure 8: General Architecture of TTS

7. Techniques for speech synthesis

The techniques of speech synthesis have been carved up into two: Rule driven speech synthesis and data-driven speech synthesis. The principle of rule-driven methods is to simulate the speech voice according to the rules deduced from the articulation process or acoustic process. It includes Articulatory Speech Synthesis and Formant Synthesis. The principle of data-driven methods is to generate the speech by recorded speech data or by the statistical parameters got from speech data [63-65]. It includes concatenative and parametric Speech synthesis. Figure 9 shows the complete categorization of the speech synthesis techniques.

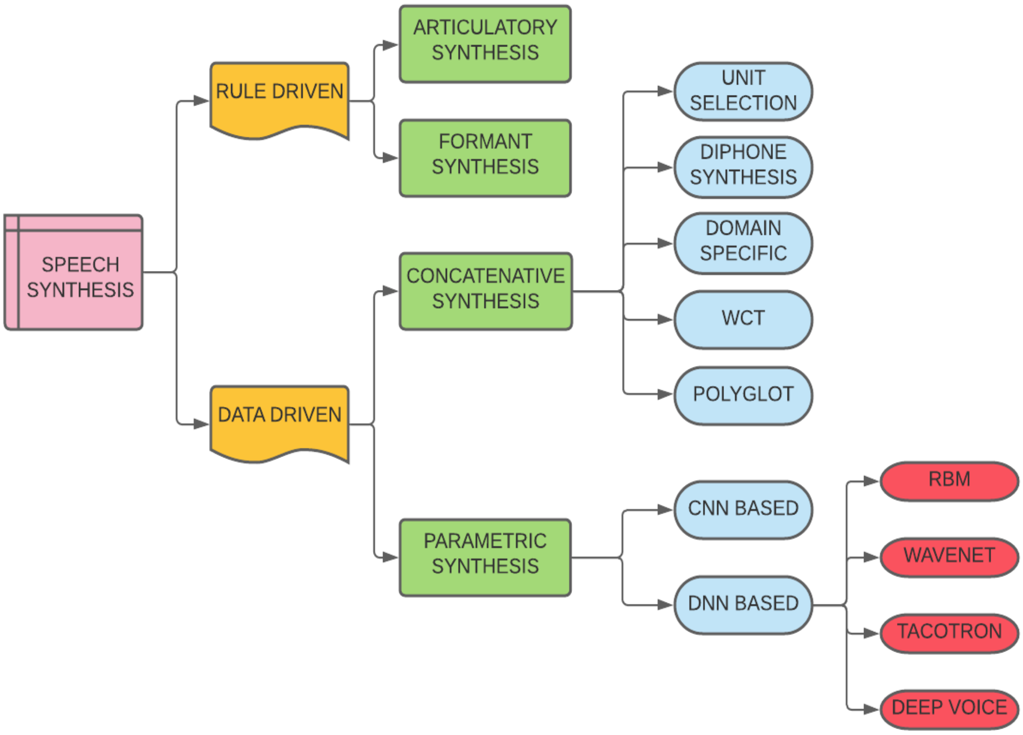

Figure 9: Approaches of Speech Synthesis

7.1. Rule Driven

7.1.1. Articulatory Speech Synthesis

The position of the speech articulators, such as the lips, jaw, and tongue, is used to approximate the vocal tract (VT) form, and airflow through the VT representation is simulated to synthesis speech in Articulatory Speech Synthesis (ASS) [66]. It simulates a human's natural voice creation process. By building a framework, T. Ngo [67] investigated the impact of speech and acoustical parameters on speech perception in a crowded place. P. Birkholz [68] discussed the use of articulatory synthesis for managing secondary prosodic aspects via rules at the word level for manipulating prosodic features such as vocal tract length, nasality, and articulatory precision. Zhigang Yin [69] concluded the main problems of the Articulatory Speech Synthesis approach. It is very difficult to acquire data for designing the articulatory model and to control the model. Although the synthesis quality of this approach is not very good, this approach is very flexible. It allows the developers to adjust the parameters of the model precisely and to change the synthesized sounds easily.

7.1.2. Formant Speech Synthesis (FSS)

The formant Speech synthesis technique is a simulation of the acoustic process. This method is rule-based, and it generates speech segments by generating fake signals that closely resemble the formant structure and other spectral aspects of natural speech. [70]. Kaur G [71] defined the formant synthesis as the ‘synthesis by the rule’. The source-filter model of speech generation is used in Formant Synthesis. In this paradigm, speech is produced by two types of sound sources: voiced sound (such as vowels) and unvoiced sound (such as consonants) (most consonants). The vocal tract model then modifies the resulting voiced and unvoiced sounds, which are then radiated by the radiation model. The most famous formant synthesizer is the Klatch synthesizer which is developed by Dennis Klatch in 1980. The fundamental benefit of formant synthesis is its small size and flexibility (generated speech output is intelligible and can be modified easily to obtain different voice and emotion characteristics). They also require little memory and can run on compact, power-saving devices.

7.2. Data-Driven

7.2.1. Concatenative Speech Synthesis (CSS)

This method is based on corpora and employs a dataset of peer-recorded speech samples (words, phonemes, syllables, allophones, di-phones, tri-phones, and half-syllables) to generate verbal output by concatenating appropriate speech components based on the set of input. Some techniques of concatenative synthesis are Diphone Synthesis, Domain-Specific, Syllable based, Waveform Concatenation Technique (WCT), Multi-Lingual, Polyglot Synthesis, etc. [72-73]. The foremost benefit of CSS is that utterance is synthesized by concatenating several natural speech segments which produce intelligible and natural-sounding synthetic speech. As far as the downsides of concatenative are concerned, it is limited to one speaker and one voice, also there are the discontinuities of the unit boundaries and the artificial feeling of the prosody. To overcome the downsides other variants of CSS like Unit Selection Synthesis (USS) came into existence.

7.2.2. Statistical Parametric Speech Synthesis (SPSS)

Parametric models are used to synthesize the speech derived from homophones. It uses statistical parameters instead of speech data to synthesize the speech. The model is said to be statistical parametric because speech is specified by parameters that are defined using statistics (variance, mean, etc.). The excitation parameters and spectral parameters of the speech are extracted from the speech corpora by using a set of generative models e.g. Deep Neural Network (DNN), Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Generative Adversarial Network (GAN), etc. While performing synthesis, the parameters corresponding to different speech units are estimated in the text using the generative models [74-75]. The voice variations can be generated by reflecting changes in parameters. Major advantage includes SPSS as a highly flexible technique with respect to vocal sway and speech style. In addition to this, it overcomes the degradation of speech quality and the memory requirement to store the parameters is less.

Contrast and comparison of all the approaches are given in table 11.

Table 11: Comparison of speech synthesis techniques [76-81]

Technique	Method Used	Advantages	Disadvantages
Articulatory Speech Synthesis	A Mathematical model for generation of	a) No speech database is required b) Produces natural synthesized speech	a) Difficult to obtain parameters for modelling human speech

Technique	human voice	c) Produces highly intelligible speech	production
Formant Speech Synthesis Technique	A methodology based on Rules	a) There is no need for a speech corpus. b) Good in producing non-nasal and fricative sounds c) The user can control all acoustic characteristics	a) Difficult to manage timing of source and filter parameters b) Produces artificial robotic sound
Concatenative			
Unit Selection technique	Speech units are stored using a corpus-based technique.	a) Natural-sounding speech is generated. b) Known for picking segments from a certain location, resulting in less-than-ideal synthesis.	a) A huge speech database is required. b) Process of synthesis is slow c) High development time and cost
Diphone Speech Synthesis Technique	Diphone units are stored using a corpus-based technique.	a) Works for small size database b) The limited vocabulary is needed	a) Need to identify diphone units for adding new language. b) Doesn't works well with language having inconsequence with pronunciation rules
Domain Specific Speech Synthesis	Corpus Based Technique storing speech word segments	a) Highly natural quality speech is produced b) Simple to implement	a) Limited words may be produced
Syllable based Speech Synthesis	Syllable units are stored using a corpus-based technique.	Natural speech segments are produced	In order to add new language, you'll need to determine syllable units.
WCT Speech Synthesis Technique	Fraction-based waveform concatenation technique	a) Overall naturalness is on the average b) Fewer storage space c) Overhead computation is reduced	For the smooth concatenation, dynamic fraction duration evaluation is required.
Multi-Lingual Technique	Corpus-based technique	Able to generate speech in different languages from a single speech corpus	It is tough to create a speech repository.
Polyglot Speech Synthesis	Corpus-based technique	Able to produce voice in multiple languages	Language switching is difficult
Statistical Parametric SpeechSynthesis			
HMM Based Speech Synthesis	Statistical technique	a) It is possible to model vocal characteristics and create new voices. b) A modest amount of speech data can be used to synthesis emotional speech.	a) Quality of speech is lower than unit selection b) Conversion of prosodic features is difficult
DNN Based Speech Synthesis Technique	Deep learning	a) Supports context- based variation in acoustic modelling b) Noise robustness	c) Need large training data d) High computational cost
RBM Speech Synthesis Technique	Deep Neural Network	a) DBN can be obtained by training and stacking several layers of Restricted Boltzmann Machines (RBM) to provide better computational power b) Can encode high order correlations	a) Computationally expensive b) Suffer from segmentation problem of training data
Wavenet	Deep Neural	a) Can produce high quality	a) Require long training

	Network	speech waveforms b) Faster than RNN c) High naturalness of speech	stage b) Front-end errors will have an impact on the synthesis effect.
Tacotron	Deep Neural Network	a) Speech synthesis model with complete end-to-end functionality b) Has the ability to generate high-quality voice waveforms c) considers character feature	Cost to train the model is high
CNN	Deep Neural Network	Training the model is achieved at high speed by embedding transfer learning.	Quality of speech may be downgraded.
Deep Voice	Deep Neural Network	a) Infused small networks that brisklydepict speech waveforms in real time. b) considers both character and phoneme feature c) combines both wavenet and tacotron features	a) Trade-off between speech synthesis speed and speech quality. b) Not end to end synthesis

8. DISCUSSION

The goal of this Systematic Literature Review (SLR) is to provide an understanding of recent applications and analogous challenges in the Sanskrit Text-to-Speech System. In nutshell, we first and foremost discuss the most relevant study conclusions, inadequacies of the studied techniques, as well as unresolved challenges from our SLR findings. The flaws and inadequacies in the studied work, as well as the repercussions of evaluating the Sanskrit Text-to-Speech System, are one of the major issues.

8.1. Key findings of SLR

Some of the most notable findings from the SLR analysis are as follows:

- Analysed that Sanskrit is suitable for computational linguistics and it is been considered the most suitable language for Artificial Intelligence.
- Identified speech synthesis approaches/ technologies that considerably perform well in achieving the efficiency of the system.
- Several open issues in this field that are still not sufficiently discussed and solved.

8.2. SLR limitations

In this article, we attempted to address research questions in order to compile a summary of current literature related to the Sanskrit text to speech systems. Because the systematic review employed the PICO technique and was confined to papers discovered on these websites, it may have missed important articles from other websites.

8.3. Open issues

In the analysis of previously implemented Sanskrit Text-to-Speech systems; opportunities and prospects for further research were inferred. The primary research difficulties for building a generic framework for an efficient Sanskrit Text-to-Speech System are discussed in this part. The following are the key challenges to accomplish a Text-to-Speech synthesis [82-86]:

8.3.1. Text analysis

The task of text analysis includes text normalization, detecting spelling errors, and syllabification [87-90]. Some specific challenges at the text analysis level are:

a) Character Encoding Identification: Text is stored on computers using some character encoding schemes e.g., ASCII (American Standard Code for Information Interchange), ISCII (Indian Standard Code for Information Interchange), Unicode (UTF-8, UTF-16, UTF-32), etc. Character encoding schemes map internal computer numbers to characters in the text. A single code point in the Unicode character set can be mapped to different byte sequences, depending on the type of encoding used for the text. Thus, the challenge for Text-to-Speech system is to figure out and process the character encoding scheme being used.

b) Text Normalization: Text pre-processing is also called Text Normalization. The input text may contain numbers, abbreviations, acronyms, dates, time, and special characters. Ambiguities in text related to date format, numbers (cardinal, ordinal, roman), and abbreviations that need expansion into phonic representation make it complex.

c) Sentence splitting: This task includes the segmentation of the input text into a list of sentences. Numerous algorithms in Text-to-Speech synthesis choose one sentence at first glance. Owing to the fact, linguistic units comparatively smaller than a sentence are substantially impacted by their adjacent units which in turn make its execution difficult [91-93]. The major concern is to determine how a text should be divided into sentences for further processing. Thus, we need to find such an algorithm that offers easy processing in less time.

d) Identifying the word boundaries at text level: The sequence of characters delimited by white spaces, or whitespace and punctuation is identified as a word. This is the basic definition of 'word' used for generating the list of words from an input sentence [94]. Sanskrit has very limited punctuation, thus it becomes a major challenge to identify the word boundaries.

8.3.2. Grapheme-to-Phoneme (G2P) Conversion

The general approach of transforming a grapheme into a phoneme produces a phonic notation (visual representation) from the textual input (letters or words) [95-97]. The orthography of letters or words can be named as a grapheme sequence (or graphemes); the phonetic form can be named as a phoneme sequence (or phonemes). The effective output of the whole Text-to-Speech system relies on the supremacy (the grapheme-phoneme string alignment, word error rate, and phoneme error rate) of the pronunciation module. Consequently, the precision and validity of Grapheme-to-Phoneme (G2P) conversion become a substantial task.

8.3.3. Speech Synthesis

Signal concatenation: Signal Concatenation stage is responsible for producing acoustic output for given text input [98]. A number of different approaches are available for generating acoustic signals. The challenge is to select a model like Hidden Markov Model (HMM), Support Vector Machine (SVM), Deep Neural networks (DNN) and its variants like Convolutional Neural Networks (CNN), Deep Convolutional Neural Networks (DCNN), Recurrent Neural Networks (RNN), Generative Adversarial Networks (GANs), Deep Belief Network (DBN), Gaussian Mixture Models (GMM), etc. taking the relationships between the enunciated inputs and the phonic features into consideration.

9. Conclusion and future research avenues

Sanskrit is a language that needs to be preserved and it can only be done, when it will reach to young minds and people will get the resources to learn, read and use this language. The country will never require following any educational culture, medical invention, or space research from any other country. But due to limited resources to learn, read and write in the Sanskrit language and less importance given by the educational curriculum, people don't provide enough attention to it. The survey discusses various techniques of Speech synthesis and also a comparison is done between all major techniques. No efficient Sanskrit Text-to-Speech System is working in the current scenario.

9.1. Future Research avenues

After the study of this paper, the researchers can implement the open challenges in their research problems. Developing a Text-to-Speech for a new low resource language requires language-specific modules. The researchers can pay attention to some new technologies of Deep Neural Network (DNN) and variants of DNN after examining the merits and demerits of DNN. There are many open areas for improvement in terms of implementing supra-segmental features.

We need to develop a Sanskrit Text-to-Speech System with prosody prediction modules- pause, phone duration, pitch, intonation, etc. implemented within to give naturalness to synthesized voice. Also, the Text-to-Speech System for Vedic Sanskrit i.e., chanting of hymns with accent markers has not been developed so far. The system for Vedic Sanskrit should perform Verse Recitation in proper prosody. To make the Sanskrit Text-to-Speech System more acceptable, it must deal with the uncovered category of Non-Standard Words (NSW) like Roman numerals, Fractions, Ratio, URL, Number range, Percentage, Alphanumeric String, etc. When existing Non-Standard Words (NSW) belong to more than one category then there should be an Ambiguity Resolution Mechanism. The system should remove the stress of reading for various categories of people e.g., person with a learning disability, a person with literacy issues, etc. Text-to-Speech System should be developed with a Text Normalizer that is suitable explicitly for blind people.

REFERENCES

[1] Xu, S.H. (2007). Study on HMM-Based Chinese Speech Synthesis; Beijing University of Posts and Telecommunications: Beijing, China, 2007.

[2] K. David Harrison. (2007). "When Languages Die: The Extinction Of The World's Languages And The Erosion Of Human Knowledge". When Languages Die: The Extinction Of The World's Languages And The Erosion Of Human Knowledge.

[3] Khan Faizal (2020). After Yoga, it's now the turn of Sanskrit to take the global state - The Financial Express.

[4] Higgins J. et al. (2011). Cochrane Handbook for Systematic Reviews of Interventions. Vol. 4. John Wiley & Sons.

[5] B. Kitchenham and S. Charters. (2007). "Guidelines for performing Systematic Literature Reviews in SE," Guidel. Perform. Syst. Lit. Rev. SE. pp. 1–44.

[6] D. Moher. et. al. (2009). "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement (Chinese edition)", J. Chin. Integrative Med., vol. 7, no. 9, pp. 889-896.

[7] Ramkumar, S. (2020). Research Productivity through the Lens of Doctoral Guidance: A Study of Sanskrit Universities in India. Journal of Scientometric Research. 9(1). 10.5530/jscires.9.1.3.

[8] Dwivedi A.V. (2018) Sanskrit (Saṃskṛt). In: Hinduism and Tribal Religions. Encyclopaedia of Indian Religions. Springer, Dordrecht.

[9] W.D. Whitney. (1884). "The Study of Hindu Grammar and the Study of Sanskrit" is an article from The American Journal of Philology, Volume 5.

[10] Dwivedi, R. (2015). Psycholinguistics and Sanskrit: Is Devabhasha Mother of Psycholinguistics?. Psycholinguistics, (18 (2)), 181-185.

[11] Paul S. (2019). Who killed Sanskrit?. Deccan Herald. 10. https://www. deccanherald.com/opinion/main-article/who-killed-sanskrit-756464.html

[12] Book: Science of the Mystical: An Eye Opener to the Scientific Substratum in Spirituality. Dr. Sree Ranjani Sudhakar. Giri trading agency.

[13] Joshi, N. (2016). Sanskrit. World History Encyclopedia.

[14] Uni, K. (2017). Shared Arabic and Sanskrit loanwords beneficial for teaching Malay vocabulary to Nepali speakers. Pertanika Journal of Social Sciences and Humanities. 25. 1199-1212.

[15] Werner, K. (1987). THE INDO-EUROPEANS AND THE INDO-ARYANS: THE PHILOLOGICAL, ARCHAEOLOGICAL AND HISTORICAL CONTEXT. Annals of the Bhandarkar Oriental Research Institute, 68(1/4), 491-523.

[16] Dwivedi, Dhananjay Vasudeo. (2018). Plant Propagation as Described in Sanskrit Texts. SAMBODHI.

[17] Dwivedi, Dhananjay. (2019). Plant morphology as depicted in Sanskrit texts.

[18] Dwivedi, D. V. (2021). Cultural and Scientific Evaluation of Nārikela (Coconut) in Indian Perspective. Jahnavi Sanskrit E-Journal. 11(I) (44), 58–67.

[19] Joshi, R. &Dharmadhikari, T. &Bedekar, Vijay. (2008). The Phonemic Approach for Sanskrit Text. Sanskrit Computational Linguistics, First and Second International Symposia Rocquencourt, France. 5402. 417-424.

[20] Dwivedi, D. V. (2017). Concept of Irrigation as Depicted in Sanskrit Texts. संस्कृतविमर्शः, 12, 71–90.

[21] Dwivedi, D. V. (2020). Scientific and Medicinal Evaluation of Aśvagandhā (Withaniasomnifera (Linn.) Dunal). Journal of the Oriental Institute.

[22] Vaishali Kherdekar. (2024). Importance of Sanskrit Language in Natural Language Processing and Machine Translation: A Review. International Journal of Intelligent Systems and Applications in Engineering, 12(3), 3515–3519.

[23] Zöe Handley. Is text-to-speech synthesis ready for use in computer-assisted language learning?. Speech Communication. 2009.Pages 906-919. ISSN 0167-6393.

[24] Daniel Korzekwa, Jaime Lorenzo-Trueba, Thomas Drugman, BozenaKostek. 2022. Computer-assisted pronunciation training—Speech synthesis is almost all you need, Speech Communication.

[25] Suhani, & Dev, Amita & Bansal, Poonam. (2023). CTC-Based End-to-End Speech Recognition for Low Resource Language Sanskrit. 1-5.

[26] Cohen, Philip & Oviatt, Sharon. (1995). The Role of Voice input for Human-Machine Communication. Proceedings of National Academy of Science, PNAS-1995, 1073, USA.

[27] Christogiannis, C., et. Al. (2002). Construction of the acoustic inventory for a Greek text-to-speech concatenative synthesis system, In Proc: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. II929-932.

[28] E. Veera Raghavendra, S. Desai, B. Yegnanarayana, A. W. Black and K. Prahallad, "Global syllable set for building speech synthesis in Indian languages," 2008 IEEE Spoken Language Technology Workshop, 2008, pp. 49-52.

[29] Chaudhari, P. R., Gangurde, P. C., & Kulkarni, N. L. (2015). Study of methodologies for utilizing Sanskrit in computational linguistics. International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE), 1

[30] Panda, S.P., Nayak, A.K. & Rai, S.C. (2020). A survey on speech synthesis techniques in Indian languages. Multimedia Systems 26, 453–478.

[31] Shiga Yoshinori et al. (2020). Text to Speech. Speech-to-Speech Translation, Springer Briefsin Computer Science.

[32] Padda Sheilly et al. (2020). Review of various Text-to-Speech Synthesis Methodologies. International Research Journal of Engineering and Technology (IRJET). ISSN: 2395-0072.

[33] Singh, A., Kadyan, V., Kumar, M. et al. (202) ASRoIL: a comprehensive survey for automatic speech recognition of Indian languages. Artificial Intelligence Rev 53, 3673–3704.

[34] Ning Y., He Sheng Wu Zhiyong, Xing chunxiao. (2019). A Review of Deep Learning Based Speech Synthesis. Appl. Sci., 9, 4050.

[35] Joshi, Sarang &Bairagi, Vinayak. (2019). Recent Trends in Text-to-Speech Synthesis of Indian Languages. HELIX. 9. 4931-4936.

[36] Sharma, Pulkit &Abrol, Vinayak & Nivedita, & Sao, Anil. (2018). Reducing footprint of unit selection based text-to-speech system using compressed sensing and sparse representation. Computer Speech & Language.

[37] G., Sujay & Borkar, Samarth. (2018). Acoustics Speech Processing of Sanskrit Language. International Journal of Computer Applications. 180. 27-32.

[38] NeeteshVashishtha. (2017). Implementation of Sanskrit Linguistics in Artificial Intelligence Programming. International Journal of Advances in Computer and Electronics EngineeringVolume: 02 Issue: 02, pp-17.

[39] Patel K.V. (2017). Sanskrit: Some Insights as a Computer Programming Language. 4th International Conference on Multidisciplinary Research & Practice (4ICMRP-2017). pp 179.

[40] Kulkarni Devyani et al. (2016). HTK Based Speech Recognition Systems for Indian Regional Languages: A Review. International journal of engineering research and technology, ISSN: 2395-0072.

[41] Pokhariya, J.S., & Mathur, S. (2014). Sanskrit Speech Recognition using Hidden Markov Model Toolkit. International journal of engineering research and technology, 3.

[42] Chandra SekharamBondu and Rama Krishna S. (2014), "An Approach for Grapheme to Phoneme Alignment for Sanskrit TTS", International Journal of Computer Networking, Wireless and Mobile Communications (IJCNWMC), Vol.4, Issue 2, Apri-2014, 93-100.

[43] Mishra, Diwakar & Bali, Kalika & Jha, Girish. (2013). Syllabification and Stress Assignment in Phonetic Sanskrit Text. ICSDA.

[44] Mishra, Diwakar, Bali, Kalika and Jha, Girish Nath. (2011). Challenges in Development of Sanskrit Speech Synthesis. Proc. ICISIL 2011, CCIS-139, Springer, Heidelberg.

[45] Acharya. (2002). Project by IIT. Madras. Multilingual Computing for Literacy and Education.

[46] Jain, Harsh, Kanade, Varun and Desikan, Kartik. (2004). Vani - An Indian Language Text to speech Synthesizer, Final Stage Report'. Dept. of Computer Science and Engineering, IIT, Mumbai.

[47] https://ltrc.iiit.ac.in/

[48] https://hear2read.org/

[49] Whitney, W. (1869). On the Nature and Designation of the Accent in Sanskrit. Transactions of the American Philological Association (1869-1896), 1, 20-45.

[50] Panda, S. P., & Nayak, A. K. (2015). An efficient model for text-tospeech synthesis in Indian languages. International Journal of Speech Technology, 18(3), 305–315.

[51] Panda, S. P., & Nayak, A. K. (2016). Modified Rule-based concatenative technique for intelligible speech synthesis in Indian languages. Advanced Science Letters, 22(2), 557–563.

[52] Bhangale, K. B., &Mohanaprasad, K. (2021). A review on speech processing using machine learning paradigm. International Journal of Speech Technology, 24(2), 367-388.

[53] Jha, D., Jha, R., & Varshney, V. (2014). Natural Language Processing and Sanskrit. International Journal of Computer Engineering & Technology, 5(10), 57-63

[54] Hustad K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. Journal of speech, language, and hearing research: JSLHR, 51(3), 562–573.

[55] Alonso Martin F. et. al. (2020). Four-Features: Evaluation of Text to Speech Systems for Three Social Robots. Electronics. 9(2):267.

[56] Nick Campbell. (2007). Evaluation of speech synthesis: From Reading Machines to Talking Machines. In Evaluation of Text and Speech Systems. Pages 29–64.

[57] Wesley Mattheyses, Werner Verhelst. 2015. Audiovisual speech synthesis: An overview of the state-of-the-art. Speech Communication. Pages 182-217.

[58] Mark Huckvale et. Al. (2007). How Is Individuality Expressed in Voice? An Introduction to Speech Production and Description for Speaker Classification. In Speaker classification I, Pages 1–20.

[59] Gobinda G Chowdhury. (2003). Natural language processing. In Annual Review of Information Science and Technology, 37, Pages 51–89.

[60] Gujarathi, Priyanka & Patil, Sandip. (2021). Review on Unit Selection-Based Concatenation Approach in Text-to-Speech Synthesis System. 10.1007/978-981-33-6691-6_22.

[61] Simon King. (2010). A beginners guide to statistical parametric speech synthesis. In A tutorial on HMM speech synthesis.

[62] Masashi Aso, Shinnosuke Takamichi, NorihiroTakamune, Hiroshi Saruwatari. 2020. Acoustic model-based subword tokenization and prosodic-context extraction without language knowledge for text-to-speech synthesis. Speech Communication. Volume 125. Pages 53-60

[63] Heiga Zen, Keiichi Tokuda, and Alan W. Black. (2009). Review: Statistical parametric speech synthesis. Speech Communication. 51(11). 1039–1064.

[64] Rolf Carlson, Björn Granström. 2005. Data-driven multimodal synthesis. Speech Communication. Pages 182-193.

[65] Pravin Bhaskar Ramteke, Shashidhar G. Koolagudi. 2019. Phoneme boundary detection from speech: A rule based approach. Speech Communication. Pages 1-17.

[66] D. Qinsheng. et. al. (2011). Articulatory Speech Synthesis: A Survey. 14th IEEE International Conference on Computational Science and Engineering, Dalian, 2011, pp. 539-542.

[67] Ngo, T., Akagi, M., Birkholz, P. (2020). Effect of articulatory and acoustic features on the intelligibility of speech in noise: an articulatory synthesis study. Speech Communication. 117, 13-20, 2020.

[68] Brizkhols, P., Lucia M., Xu Y. (2017). Manipulation of prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis. Computation of Speech Language.

[69] Yin, Zhigang. (2017). A Simplified Overview of TTS Techniques. DEStech Transactions on Computer Science and Engineering.

[70] Stevens, K.N. (2002). Toward formant synthesis with articulatory controls. In Proc: IEEE Workshop on Speech Synthesis, pp. 67-72

[71] G. Kaur and P. Singh. (2019). Formant Text-to-Speech Synthesis Using Artificial Neural Networks. Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), Gangtok, India. pp. 1-6.

[72] Panda, Soumya & Nayak, Ajit. (2020). Spectral Smoothening Based Waveform Concatenation Technique for Speech Quality Enhancement in Text-to-Speech Systems. 10.1007/978-981-15-1081-6_36.

[73] Matej Rojc, Zdravko Kačič. 2007. Time and space-efficient architecture for a corpus-based text-to-speech synthesis system. Speech Communication. Volume 49. Issue 3. Pages 230-249.

[74] Panda, Soumya & Nayak, Ajit & Patnaik, Srikanta. (2015). Text-to-speech synthesis with an Indian language perspective. International Journal of Grid and Utility Computing, Inderscience Publisher, 6. 170.

[75] H. Ze, A. Senior and M. Schuster., (2013). Statistical parametric speech synthesis using deep neural networks. IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 7962-7966.

[76] Book:

[77] Voice Communication between Humans and Machines, David B. Roe and Jay G. Wilpon, Editors. National Academy of Sciences.

[78] Tomoki Koriyama and Takao Kobayashi. 2019. Statistical Parametric Speech Synthesis Using Deep Gaussian Processes. IEEE/ACM Trans. Audio, Speech and Lang. Proc. 27, 5 (May 2019), 948–959Ping, Wei & Peng. et. al. (2017). Deep Voice 3: 2000-Speaker Neural Text-to-Speech.

[79] Heiga Zen, Keiichi Tokuda, Alan W. Black. 2009. Statistical parametric speech synthesis. Speech Communication. Volume 51, Issue 11. Pages 1039-1064.

[80] Delic, V., et. al. (2019). Speech technology progress based on new machine learning paradigm. Computational intelligence and neuroscience, 2019.

[81] Yolchuyeva, S., Németh, G., &Gyires-Tóth, B. (2019). Grapheme-to-Phoneme Conversion with Convolutional Neural Networks. Applied Sciences, 9, 1143

[82] Bali, K., & Das, A. (2006) A Note on Interpreting Text for Indian Language TTS.

[83] Z. Ling et al. (2015). Deep Learning for Acoustic Modeling in Parametric Speech Generation: A systematic review of existing techniques and future trends. IEEE Signal Processing Magazine, vol. 32, no. 3, pp. 35-52

[84] S. Rao, et al. (2015). TTS Evaluation: Double-ended Objective Quality Measures. Proceeding of IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), pp. 1-6

[85] Kentaro Mitsui, Tomoki Koriyama, Hiroshi Saruwatari. 2021. Deep Gaussian process based multi-speaker speech synthesis with latent speaker representation. Speech Communication. Volume 132. Pages 132-145.

[86] Kuligowska, Karolina &Kisielewicz, Paweł&Włodarz, Aleksandra. (2018). Speech synthesis systems: Disadvantages and limitations. International Journal of Engineering and Technology (UAE). 7. 234-239. 10.14419/ijet.v7i2.28.12933.

[87] Burkhardt, Felix & Reichel, Uwe. (2016). A Taxonomy of Specific Problem Classes in Text-to-Speech Synthesis: Comparing Commercial and Open Source Performance.

[88] Joshi, H., Bhatt, A., & Patel, H. (2013). Transliterated search using syllabification approach. In Forum for information retrieval evaluation.

[89] Pandey, P. (2014). Akshara-to-sound rules for Hindi. Writing Systems Research, 6(1), 54-72.

[90] K. Vythelingum, Y. Estève and O. Rosee, "Error detection of grapheme-to-phoneme conversion in text-to-speech synthesis using speech signal and lexical context," 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2017, pp. 692-697, doi: 10.1109/ASRU.2017.8269004.

[91] Hellwig, O. (2016). Detecting sentence boundaries in Sanskrit texts. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 288-297).

[92] Hellwig, Oliver. (2015). "Using Recurrent Neural Networks for joint compound splitting and Sandhi resolution in Sanskrit." In 4th Biennial Workshop on Less-Resourced Languages.

[93] Hellwig, O., &Nehrdich, S. (2018). Sanskrit word segmentation using character-level recurrent and convolutional neural networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 2754-2763).

[94] Reddy, V., Krishna, A., Sharma, V. D., Gupta, P., & Goyal, P. (2018). Building a word segmenter for sanskrit overnight. arXiv preprint arXiv:1802.06185.

[95] Yolchuyeva S, Németh G, Gyires-Tóth B. Grapheme-to-Phoneme Conversion with Convolutional Neural Networks. Applied Sciences. 2019; 9(6):1143.

[96] Hadj Ali, I., Mnasri, Z. &Lachiri, Z. DNN-based grapheme-to-phoneme conversion for Arabic text-to-speech synthesis. Int J Speech Technol 23, 569–584 (2020).

[97] Kumar, C.S., Govind, D., Nijil, C., &Narwaria, M. (2006). Grapheme to Phone Conversion for Hindi.

[98] J. Schroeter et al. A perspective on the next challenges for TTS research. Proceedings of 2002 IEEE Workshop on Speech Synthesis. 2002. pp. 211-214.

[99] http://sanskrit.jnu.ac.in/samvacaka/index.jsp

[100] Holla, S. et al. (2022). End-to-End Speech Recognition for Low Resource Language Sanskrit using Self-Supervised Learning. 148-152.