Natural Language Processing Journal 12 (2025) 100171

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

A survey and evaluation of text-to-speech systems for the Tamil language

Ahrane Mahaganapathy *, Kengatharaiyer Sarveswaran

Department of Computer Science, University of Jaffna, Post Box 57, Ramanathan Road, Thirunelvely, Jaffna, Sri Lanka

ARTICLE INFO

Keywords:
TTS
Speech synthesis
Tamil
Benchmark
TTS evaluation

ABSTRACT

This survey provides a comprehensive review of existing Tamil Text-to-Speech (TTS) synthesis systems, synthesis approaches, evaluation approaches, and highlights state-of-the-art approaches and challenges in handling linguistic nuances. Voice-based interfaces are becoming part of life. Therefore, it is import to have an expensive TTS system which can make human experience better. Tamil, with its rich linguistic features and diagnostic nature, presents significant challenges to speech synthesis. In addition to the survey, importantly this work proposes a perceptual evaluation framework which consists of expressiveness, low listening fatigue, and overall quality, in addition to traditional intelligibility and naturalness, dimensions to evaluate better human experience. This study also uses the Comparative Mean Opinion Score (CMOS) for the subjective evaluation instead of the Mean Opinion Score. A dataset for the evaluation was also carefully prepared and six widely used Tamil TTS systems were evaluated using Word Error Rate and the subjective evaluation was done using the proposed evaluation framework with the support of 30 evaluators. The reliability of the subjective evaluation is also assessed using Krippendorff's Alpha. The results indicate the existing systems have significant room for improvement in all perceptual dimensions. The study underscores the need for evaluation datasets and evaluation approaches that cater to subjective perceptual dimensions of speech synthesis for better human experience and lays a foundation for future research and development in Tamil and similar TTS systems.

1. Introduction

Text-to-speech (TTS) systems are essential for human–computer interaction, enabling the conversion of written text into a synthesized spoken format. In recent years, advancements in deep learning have enabled TTS technology to evolve, producing more natural and human-like speech than ever before (Pratap et al., 2024). The TTS process begins with text analysis, where natural language processing (NLP) techniques break down and normalize the input text to ensure correct pronunciation. This includes transforming the text into phonemes (the basic sound units of speech) and processing them with prosody, which adds intonation, stress, and rhythm to make the speech sound more natural (Ahmad and Rashid, 2024).

The versatility of TTS systems allows for their use in various applications. For instance, they serve as vital assistive technologies for the visually impaired, allowing users to listen to screen content via auditory output, significantly improving accessibility (Lăpușteanu et al., 2024; Sharma et al., 2014). Additionally, TTS is widely implemented in commercial services such as virtual assistants such as Hoy (2018), which rely on speech synthesis to provide real-time spoken responses to user queries. Educational tools also benefit from TTS by helping language learners practice pronunciation and gain exposure to appropriate language intonation (Gobinathan et al., 2024). Moreover, TTS plays a crucial role in navigation systems, reading out real-time directions to drivers, and in automated customer service, where it powers interactive voice response (IVR) systems to handle routine inquiries (Inam et al., 2017).

By integrating deep learning models and natural language processing, TTS systems have transformed from basic robotic-sounding voices to highly intelligible and expressive speech, finding widespread application across numerous industries such as audio books, virtual assistants, language learning platforms, accessibility tool for visually impaired and customer service chat bots.

1.1. Significance of Tamil language

Tamil, spoken by more than 86 million people globally (Zeidan, 2020), holds the distinction of being one of the oldest classical languages with a rich literary tradition (Hart, 2000; Zvelebil, 2021). The development of text-to-speech (TTS) systems for Tamil is essential for promoting its usage in the digital age, contributing to the preservation of its cultural heritage (Pubadi et al., 2020).

Predominantly spoken in Tamil Nadu (India) and Sri Lanka, Tamil is also widely used among diaspora communities worldwide. It is the official language of Tamil Nadu, Sri Lanka, and Singapore, and

* Corresponding author.
E-mail addresses: ahrane@univ.jfn.ac.lk (A. Mahaganapathy), sarves@univ.jfn.ac.lk (K. Sarveswaran).

https://doi.org/10.1016/j.nlp.2025.100171
Received 31 December 2024; Received in revised form 24 May 2025; Accepted 20 June 2025

one of India's 22 scheduled languages. Tamil speakers can also be found in countries such as Canada, U.S.A, U.K, Germany, Switzerland, Australia, Malaysia, South Africa, the UAE, and others, showcasing its global reach. Tamil is also taught globally, with standardized teaching methods emerging to support non-native speakers, further emphasizing its cultural and linguistic significance.

Despite its global presence, Tamil has unique challenges. Many diaspora members and migrants to Tamil-speaking regions, such as Tamil Nadu and northern Sri Lanka, as well as older populations in rural areas, may speak Tamil fluently but lack literacy. This gap highlights the importance of TTS systems in providing auditory access to Tamil content such as audio books and creative industry.

1.2. Motivation

Despite its linguistic, cultural, and historical significance, Tamil is considered underrepresented in the field of speech technology research, especially in the development of Text-to-Speech (TTS) systems. The distinctive linguistic characteristics of Tamil are often neglected during TTS system design and evaluation.

Currently, there is no comprehensive up-to-date survey that answers how far speech synthesis has been progressed in Tamil, what are systems cater to Tamil speech synthesis and how far they perform, what are the available datsets for training and evaluation and how a diverse dataset should be for TTS, what are the linguistic challenges of Tamil speech synthesis and how we can identify and evaluate the systems that perform Tamil language speech synthesis. Existing studies primarily concentrate on high-resource languages such as English and Mandarin, offering limited insight into the unique linguistic challenges and practical constraints faced by Tamil TTS development.

There is no standardized subjective evaluation framework specifically for perceptual dimensions of TTS systems. This absence of thorough and systematic analysis obstructs the advancement of high-quality, linguistically nuanced Tamil TTS solutions. Traditional TTS evaluations mostly focus on brief, isolated utterances, overlooking important perceptual factors valued by Tamil-speaking users such as expressiveness, low listening fatigue, and the naturalness of longer-duration speech synthesis.

To bridge this gap, there is a pressing need for a Tamil-specific TTS survey, balanced dataset for evaluation and evaluation framework that integrates these perceptual dimensions, fostering the creation of more inclusive, expressive, and user-centric speech technologies for the Tamil-speaking population.

1.3. Research gaps

Based on the literature survey, this study identifies and aims to address three key research gaps in the domain of Tamil Text-to-Speech (TTS) systems: (1) the absence of a comprehensive and standardized subjective evaluation framework capable of assessing important perceptual dimensions such as intelligibility, naturalness, expressiveness, listening fatigue, and overall quality; (2) the limited exploration of the unique linguistic challenges associated with Tamil speech synthesis; and (3) the lack of a comprehensive survey and standardized evaluation of existing Tamil TTS systems. These gaps collectively contribute to the obstacles in the development of high-quality, contextually appropriate, and practically deployable TTS solutions for Tamil.

While most current evaluation approaches for TTS systems emphasize basic intelligibility and, to some extent, naturalness, they largely ignore other essential perceptual factors such as expressiveness and listening fatigue. This is particularly limiting in applications such as audiobook narration, virtual assistants, or storytelling platforms, where prosodic variation and emotional tone critically affect the user experience. The absence of a multidimensional, standardized evaluation framework makes it difficult to benchmark and compare system performance holistically. This research aims to propose a subjective evaluation framework that addresses this need and supports the advancement of more expressive and user-friendly TTS systems.

The limited exploration of the specific linguistic challenges associated with Tamil speech synthesis presents a significant obstacle in the advancement of Tamil TTS. Tamil phonology involves a range of subtle vowel and consonant distinctions, including retroflex sounds and nasalization, which require precise handling to produce natural-sounding speech. The language's prosodic features, such as stress patterns and intonation contours, further complicate synthesis, especially since these vary across different dialects and speaking contexts. Despite these complexities, existing Tamil TTS research has only marginally addressed such linguistic nuances, often relying on generic or oversimplified language models. This insufficient focus limits the naturalness, intelligibility, and expressiveness of synthesized Tamil speech.

Although Tamil is a language spoken by more than 86 million people in the world, official languages of many countries and several Tamil TTS systems have been proposed over the years, there is no unified and up-to-date survey that systematically evaluates their strengths, limitations, and relative performance. The existing literature is often fragmented, employing various methodologies and inconsistent evaluation metrics, making it difficult to accurately gauge the current state of the field. To address this critical gap, the present study conducts a comprehensive survey and evaluation of Tamil speech synthesis and TTS systems using a unified framework.

1.4. The objectives

The overarching objective of this survey paper is to report approaches for TTS development, comprehensively analyze and evaluate the current state of Tamil Text-to-Speech (TTS) systems, identify key challenges in developing expressive Tamil TTS, and provide insights into future research directions. To achieve this, the following research questions are addressed:

1. What are the approaches used in the TTS systems developed to date?
2. How do existing Tamil TTS systems perform in objective and subjective evaluation metrics?
3. What are the challenges in developing expressive Tamil TTS systems?

1.5. Contributions

One of the primary contributions of this study is the proposal of a multidimensional subjective evaluation framework for Text-to-Speech (TTS) systems. Traditional evaluation approaches often focus narrowly on intelligibility or naturalness, overlooking other important perceptual aspects that affect user experience. This study addresses this gap by extending the Comparative Mean Opinion Score (CMOS) metric to incorporate five perceptual dimensions: intelligibility, naturalness, expressiveness, low listening fatigue, and overall quality. This framework allows for a more user-relevant assessment of synthesized speech, particularly valuable in contexts such as audiobooks, virtual assistants, and education platforms where prosody and emotional tone are essential. To support consistent and linguistically informed evaluation, the study also involves the preparation and use of a balanced evaluation dataset. Building on this foundation, the study conducts a systematic evaluation of six widely used Tamil TTS systems, both commercial and open-source. These include Google TTS, Google Cloud TTS, OpenAI TTS, Bhashini TTS, IIT Madras Indic TTS, and Meta's MMS. Each system is evaluated using a dual approach: objective metrics such as Word Error Rate (WER) to assess transcription accuracy, and the proposed subjective evaluation framework to gauge perceptual quality. This combined analysis offers a clear, comparative picture of the strengths and

weaknesses of each system, providing valuable insights for researchers and developers working on Tamil speech synthesis.

Another important contribution of the study is the in-depth analysis of Tamil-specific linguistic challenges that impact the quality and effectiveness of TTS systems. Tamil exhibits significant dialectal variation, and the gap between its spoken (colloquial) and written (literary) forms known as diglossia creates difficulties in pronunciation modeling. In addition, the language includes phonetic features such as retroflex consonants, nasalization, and gemination, which are often poorly handled in existing systems. These challenges are exacerbated by the limited availability of annotated speech datasets and the lack of dialect-aware models. This study explores these issues comprehensively, highlighting the need for linguistically grounded approaches to TTS development.

In addition to this, the study offers a comprehensive survey of existing Tamil TTS methodologies and systems, covering both classical approaches like concatenative and HMM-based synthesis, and modern neural architectures such as Tacotron2 and FastSpeech2. This survey identifies key methodological trends, persistent limitations, and emerging opportunities, providing a consolidated reference point for future research and development in the field.

The novelty of this study lies in its pioneering extension of the CMOS metric to evaluate five perceptual dimensions critical to user satisfaction. It is also the first to benchmark a diverse set of commercial and open-source Tamil TTS systems using a combined objective-subjective evaluation strategy, drawing on feedback from native Tamil speakers. Moreover, evaluation dataset is made publicly available. This dual-pronged and linguistically informed approach provides a holistic understanding of the current landscape, uncovering systemic limitations particularly in expressive synthesis and setting a new standard for future Tamil TTS research.

1.6. Paper outline

The paper has six sections. Section 1 establishes the significance of Tamil TTS, identifies research gaps, and outlines the objectives and contributions of the article. The background Section 2 provides linguistic and technical foundations of TTS systems, focusing on Tamil phonology and synthesis techniques. Section 3 presents a comprehensive review of existing Tamil TTS systems, spanning both commercial platforms and open-source projects. It discusses the data sources, pre and post-processing steps, and model architectures employed in these systems. Section 4 outlines the evaluation framework and methodology used in this study and reports the results of our evaluation, focusing on both objective (Word Error Rate) and subjective (Comparative Mean Opinion Score) assessments. Results are analyzed across multiple dimensions, including naturalness, intelligibility, expressiveness, low listening fatigue and overall quality. Section 5 explores the core challenges in developing Tamil TTS systems, including the language's diglossic nature, dialectal variations, and contextual phonological complexity. Finally, Section 6 summarizes current limitations in Tamil TTS research and proposes directions for future improvement and development.

2. Background

This section introduces the key concepts required to understand the development of Text-to-Speech (TTS) systems and challenges in developing TTS for Tamil. TTS systems are designed to convert written text into speech by combining principles of linguistic analysis, signal processing, and machine learning. The development of the Tamil TTS system begins with a deep understanding of the language's phonological system. Tamil phonetics and phonology are characterized by distinct vowels, consonants, and features such as retroflex sounds and vowel length distinctions. These elements are significant in capturing Tamil pronunciation (Keane, 2004).

Due to the syllable-centric structure of the Tamil script and the presence of complex context-dependent pronunciation patterns such as allophonic variations and one-to-many mappings between letters, sounds—Grapheme-to-Phoneme (G2P) conversion mechanisms must account for context sensitive mapping to ensure accurate pronunciation modeling in TTS systems (Rajendran and Kumar, 2019).

Beyond pronunciation, prosody and intonation play important role in creating a natural flow in speech. Rhythm, stress, and pitch patterns in Tamil require careful modeling to mimic natural intonations and emotional expressions. Incorporating these aspects enhances the realism and expressiveness of the TTS output (Jalin and Jayakumari, 2017).

Modern advancements in TTS systems, particularly in deep learning, have revolutionized speech synthesis. Models such as Tacotron (Wang et al., 2017) generate mel-spectrograms from input text using sequence-to-sequence learning, which are then converted into audio waveforms using vocoders such as WaveNet (van den Oord et al., 2016), where raw text can be converted into high-quality audio with minimal manual intervention.

Despite these technological advancements, the development of Tamil TTS systems remains challenging. Adapting such architectures for Tamil involves addressing language-specific challenges, including complex grapheme-to-phoneme conversion and prosodic variations, and requires training on large, high-quality Tamil speech datasets. However, the increasing availability of open-source tools and collaborative research platforms offers promising opportunities for the creation of comprehensive datasets. By systematically addressing these challenges, the Tamil TTS systems can achieve high intelligibility and naturalness. These efforts not only make technology more accessible to Tamil-speaking communities but also contribute to preserving and promoting language in the digital era.

2.1. Tamil phonology & challenges

Tamil has a rich phonetic inventory consisting of vowels, and consonants that contribute to distinct sounds. The language is noted for its differentiation between short and long vowels and its use of consonant gemination (Kuno, 1958). Tamil has five primary vowel qualities (/i/, /e/, /a/, /o/, and /u/), each of which has a length distinction, leading to ten phonemic vowels. This contrast in length is significant for distinguishing meanings between words (Kuno, 1958). Tamil's consonant inventory includes stops, nasals, fricatives, and liquids, with a notable distinction between voiced and voiceless obstruents. The language's phonological rules are highly context-sensitive (Krishnamurthy, 1977), and consonant gemination further highlights the importance of phonological context (Balasubramanian and Asher, 1984; Heselwood, 2013). Krishnamurthy (1977) rules also highlight the economy in Tamil's orthographic system, where multiple sound values may be represented by a single symbol, depending on the context. This feature aligns with Zipf's principle of least effort, which minimizes the number of symbols required while retaining phonological richness (Cherry, 1957).

Tamil's phonological system encompasses various factors, such as context-sensitive variations, dialectal influences, and the integration of loanwords, all of which shape its rich phonetic profile (Annamalai, 1979; Keane, 2004; Suseendirarajah, 1993).

A notable feature is the behavior of obstruents (stops and fricatives). In Tamil, voiced obstruents typically occur internally after nasal segments. Conversely, voiceless plosives are predominantly found in word-initial positions, aligned with the pattern outlined by Caldwell's law (Keane, 2004).

The dialectal differences in Tamil and Telugu significantly affect the treatment of certain sounds, such as retroflex consonants, which are articulated with greater retroflexion compared to other Indian languages such as Hindi (Ladefoged and Bhaskararao, 1983). Notably, these retroflex sounds do not appear in word-initial positions within the native Tamil vocabulary which is a characteristic shared among Dravidian languages (Keane, 2004).

Tamil dialects also vary widely in their phonetic characteristics, leading to distinct realizations of sounds in different regions. For example, the Kanniyakumari dialect emphasizes a phonetic and phonological distinction between dental and alveolar stops, such as t̪ and d̪(dental) versus t and d (alveolar) (Keane, 2004). These stops are not only different in articulation but also contrast in meaning, underscoring the diversity of Tamil phonology.

Such dialectal diversity poses challenges for phonetic transcription systems because a single model may struggle to capture the nuanced phonetic variations present across Tamil-speaking regions. Developing an effective transcription framework requires integrating regional phonetic data to ensure both inclusivity and accuracy, enabling the system to reflect Tamil's linguistic spot.

The influence of loanwords, particularly from Indo-Aryan languages like Sanskrit and English, introduces additional complexities to Tamil phonology. In native Tamil words, voiced plosives such as /b/ and /d/ are not typically found in word-initial positions. However, because of borrowing of words from languages such as Sanskrit, Hindi, and English, these sounds are now present in their initial positions (Annamalai, 1979). For example, words like பைபிள் (*baibel*, 'Bible') and டாக்டர் (*daaktar*, 'doctor') begin with /b/ and /d/ respectively, reflecting English influence. Similarly, தர்மம் (*darmam*, 'dharma') and பக்தி (*bakthi*, 'devotion') are borrowed from Sanskrit, introducing initial voiced stops /d/ and /b/ into Tamil's phonological system. The inclusion of phonemes in such loanwords requires TTS systems to be adapted by recognizing these non-native phonetic elements. For example, voiced fricatives such as /v/ and /z/, which are uncommon in the native Tamil phonemic inventory, appear in borrowed terms. The TTS system must accurately identify and represent these sounds to maintain the integrity of loanwords while reflecting the native phonetic characteristics of Tamil.

Overall, these phonological features of Tamil can be challenging factors in modeling TTS speech in Tamil.

2.2. Comparison of Text-to-Speech (TTS) techniques

Methods such as articulatory synthesis, concatenative synthesis, statistical parametric approaches, and neural synthesis have been employed for speech synthesis in studies. Among these, articulatory syntheses model the vocal tract system and articulators to simulate speech sounds, relying on the precise knowledge of human articulation organs (Balyan et al., 2013). Although this method can theoretically produce high-quality speech with adjustable parameters (Jalin and Jayakumari, 2017), its adoption has been limited because of computational complexity and challenges in producing natural-sounding speech. Moreover, recent research activity in this area appears sparse, reflecting their reduced prominence in contemporary speech synthesis studies.

On the other hand, Neural TTS has emerged as the state-of-the-art approach, utilizing deep learning models to generate highly natural and expressive speech. By utilizing text and speech data pairs, neural models such as WaveNet (van den Oord et al., 2016) and Tacotron (Wang et al., 2017) can effectively produce speech. This paradigm shift towards neural methods has been fueled by their ability to outperform traditional methods, such as concatenative synthesis and statistical parametric approaches, in both flexibility and naturalness (Panda et al., 2020; Kim et al., 2024). However, the neural approach requires significant computational resources and extensive training data, making it more resource intensive than traditional techniques (Panda et al., 2020; Kim et al., 2024).

In summary, while articulatory synthesis remains an academically intriguing but underexplored method, neural TTS has firmly established itself as a leading technology in modern speech synthesis. A detailed comparison of the different text to speech techniques is presented in Table 1 (see Table 1).

2.3. TTS evaluation techniques

Evaluating Text-to-Speech (TTS) systems is essential to determine their effectiveness and quality. Various metrics were employed to assess the performance of TTS systems, each focusing on different aspects of speech as shown in Table 2.

3. Tamil Text-to-Speech (TTS) systems

3.1. Overview of Tamil TTS development approaches

Technological approaches to Tamil TTS have evolved significantly, incorporating advanced neural network architectures and non-autoregressive models. Among the prominent methods are neural TTS systems developed by companies such as ElevenLabs (ElevenLabs, 2024) and Google Cloud (Google, 2024a), which utilize deep learning techniques to generate highly realistic speech. These systems often rely on sequence-to-sequence models that convert text input into audio waveforms, thereby capturing the nuances of human speech through extensive training on diverse datasets.

Another notable model is FastSpeech2, utilized by AI4 Bharat and Bhashini IITM. This approach employs a non-autoregressive architecture, allowing for faster and more efficient speech synthesis. FastSpeech2 enhances the quality of generated speech by leveraging a text-to-mel spectrogram conversion process, which is then transformed into audio using a vocoder. This architecture improves the fluency and expressiveness of the synthesized speech, making it suitable for various applications (Murthy and Umesh, 2023).

Additionally, Meta's Voicebox represents an innovative non-autoregressive model that utilizes FlowMatching techniques. This architecture allows for high-quality voice synthesis by matching the flow of speech patterns without relying on traditional autoregressive methods. Such advancements contribute to a more natural and fluid listening experience (Le et al., 2023).

Moreover, Meta's MMS architecture leverages self-supervised learning techniques to reduce the reliance on large amounts of labeled training data. This approach allows the models to learn effectively from vast unlabeled audio data. The TTS model is built using state of the art neural network architectures that have been fine-tuned to handle the complexities of multilingual speech synthesis (Pratap et al., 2024).

HMM-based speech synthesis continues to be highly relevant, particularly in resource-constrained environments. A language-independent framework for Indian languages — including Tamil — has been proposed to streamline the development of TTS systems across linguistically related languages. By employing a common phone set and question set, this approach facilitates the reuse of monophone models from resource-rich languages like Hindi. As a result, Tamil TTS systems can be developed rapidly without compromising on intelligibility or naturalness. Experimental results show that the Degradation Mean Opinion Scores (DMOS) and Word Error Rates (WER) are comparable to those of language-specific models, highlighting the effectiveness and robustness of the method (Prakash et al., 2014).

The market landscape for the Tamil TTS consists of both commercial providers and open-source initiatives. Major commercial players include Google Cloud, Microsoft Azure, Amazon Polly (Amazon, 2024), and ElevenLabs (ElevenLabs, 2024), all of which offer TTS solutions with voice libraries that feature both male and female voices. These services claim to cater to various Tamil dialects, including Indian Tamil, Sri Lankan Tamil, Malaysian Tamil, and Singaporean Tamil. On the other hand, research initiatives such as AI4Bharat (Kumar et al., 2023), Bhashini IITM (Bhashini, 2024a), and IndicTTS (Prakash and Murthy, 2020) have focused on developing open-source solutions.

Table 1
Comparison of speech synthesis methods.

Method	Inputs	Description	Advantages	Disadvantages
Articulatory Synthesis (Jalin and Jayakumari, 2017)	Vocal resonant frequencies, knowledge of human articulation organs (Balyan et al., 2013)	Models the vocal tract system and articulators (such as lips, tongue, jaw) to simulate speech sounds (Balyan et al., 2013).	Produces high-quality speech when parameters are properly adjusted; allows modification of pitch and articulation (Jalin and Jayakumari, 2017).	Requires expert knowledge of articulation and demands high computational resources to generate natural speech (Jalin and Jayakumari, 2017).
Concatenative Synthesis (Jalin and Jayakumari, 2017; Karthikadevi and Srinivasagan, 2014)	Pre-recorded speech segments (Jalin and Jayakumari, 2017; Karthikadevi and Srinivasagan, 2014)	Pieces together pre-recorded speech to form complete utterances, applying signal processing to smooth the joins (Jalin and Jayakumari, 2017; Karthikadevi and Srinivasagan, 2014).	Delivers natural speech with a large corpus; uses low computational resources (Jalin and Jayakumari, 2017; Karthikadevi and Srinivasagan, 2014).	Requires a large speech corpus; introduces discontinuities may occur at concatenation, points (Jalin and Jayakumari, 2017; Karthikadevi and Srinivasagan, 2014).
Statistical Parametric Methods (Jalin and Jayakumari, 2017; Karthikadevi and Srinivasagan, 2014; Samuel Manoharan, 2022)	Speech data, statistical models (HMMs) (Jalin and Jayakumari, 2017; Karthikadevi and Srinivasagan, 2014; Samuel Manoharan, 2022)	Generates speech parameters (mel-cepstra, F0) from statistical models, then synthesizes the waveform (Jalin and Jayakumari, 2017; Karthikadevi and Srinivasagan, 2014; Samuel Manoharan, 2022).	Supports flexible modeling of speaking styles; reduces storage requirements compared to concatenative methods (Jalin and Jayakumari, 2017; Karthikadevi and Srinivasagan, 2014; Samuel Manoharan, 2022)	Generates synthetic-sounding speech; reduces naturalness compared to other methods (Jalin and Jayakumari, 2017; Karthikadevi and Srinivasagan, 2014; Samuel Manoharan, 2022).
Neural TTS (Panda et al., 2020; van den Oord et al., 2016; Wang et al., 2017; Kim et al., 2024)	Paired text and speech data (Panda et al., 2020; van den Oord et al., 2016; Wang et al., 2017; Kim et al., 2024)	Employs deep learning models (neural networks) to generate speech by learning linguistic and speech patterns from data (Panda et al., 2020; van den Oord et al., 2016; Wang et al., 2017; Kim et al., 2024).	Produces natural and expressive speech; captures linguistic nuances effectively.	Requires large volumes of training data; consumes high computational power during training (Panda et al., 2020; van den Oord et al., 2016; Wang et al., 2017; Kim et al., 2024).

Table 2
Evaluation metrics for text-to-speech systems.

Metric	Description	Our Interpretation
Word Error Rate (WER) (Henriksson, 2023; Murthy and Umesh, 2023; Łajszczak et al., 2024; Hu et al., 2024)	An objective measures that measures the accuracy of synthesized speech by comparing it to a reference transcript at word level (Henriksson, 2023).	Lower WER indicates better performance.
Character Error Rate (CER) (Kumar et al., 2023)	Calculates the character-level errors between synthesized and reference text	Lower CER signifies higher textual accuracy.
Mean Opinion Score (MOS) (Shahid et al., 2016)	A subjective measure where listeners rate the quality of speech on a scale from 1 (bad) to 5 (excellent) (Shahid et al., 2016).	Higher MOS indicates better perceived quality.
Comparative Mean Opinion Score (CMOS) (ITU-T, 1996)	A subjective measure that allows listeners to compare two speech samples and rate them on a scale from −3 (much worse) to +3 (much better) (ITU-T, 1996).	Positive scores indicate preference for one sample over another.
Degradation Mean Opinion Score (DMOS) (Prakash et al., 2014)	Subjective measure of degradation by comparing degraded speech with a reference.	Lower DMOS indicates less perceived quality loss.
Fundamental Frequency Error (F0) (Kumar et al., 2023)	Measures error in pitch contours between synthesized and natural speech.	Lower F0 error reflects better pitch and prosodic quality.
Segmental Intelligibility Test (Shahid et al., 2016)	A subjective measure that tests intelligibility that focuses on the correct identification of consonants in both initial and final positions within monosyllabic words (Shahid et al., 2016).	A higher percentage of correctly identified words indicates better segmental intelligibility.
Sentence Level Intelligibility (SUS Test) (Shahid et al., 2016)	A subjective measure where subjects are asked to transcribe the sentences they hear, which are grammatically correct but lack inherent meaning. This minimizes the possibility of using contextual information (Shahid et al., 2016).	A higher percentage of correctly transcribed words indicates better intelligibility at the sentence level.
Comprehension Test (Shahid et al., 2016)	A subjective measure that assesses the understanding of the underlying message in the synthesized speech. Subjects listen to a paragraph and then answer questions about its content (Shahid et al., 2016).	Higher scoring shows better performance on the comprehension test.
Mel Cepstral Distortion (MCD) (Sailor, 2013; Salesky et al., 2021; Behdad and Gharavian, 2023; Baljekar and Black, 2016)	An objective measure that uses a distortion metric to measure the difference between the Mel-Frequency Cepstral Coefficients (MFCCs) of synthesized and reference speech (Sailor, 2013; Salesky et al., 2021; Behdad and Gharavian, 2023; Baljekar and Black, 2016).	Lower MCD values indicate that the synthesized speech is closer to the reference speech in acoustic properties.

3.2. State-of-the-art techniques

Regarding Tamil TTS systems, researchers have employed various approaches to enhance the speech synthesis quality. One commonly used method is the concatenative approach, which involves concatenating short speech units recorded by human speakers. This approach has shown promising results in producing natural-sounding Tamil speech. Additionally, deep learning techniques, such as WaveNet (van den Oord et al., 2016) and Tacotron (Wang et al., 2017), have led to significant advancements in the performance of Tamil TTS systems by generating more natural intonations and accents.

3.3. Review of available Tamil TTS systems

This comprehensive analysis of TTS systems highlights the diverse methodologies, datasets, and performance metrics used in developing and evaluating TTS solutions, particularly focusing on Tamil and multilingual speech synthesis. The surveyed systems show a broad range of approaches, including concatenative synthesis, neural models such as Tacotron2 and FastSpeech, and advanced generative models such as Glow-TTS and non-autoregressive techniques.

Several models, such as AI4Bharat-TTS and IndicTTS, emphasize leveraging the Indic language datasets to achieve higher intelligibility and naturalness. Meanwhile, efforts such as the bilingual Tamil–English TTS and Thirukkural II focus on specific linguistic challenges, addressing phonetic richness, prosody, and co-articulatory coherence to enhance Tamil speech synthesis. Additionally, techniques like diphone insertion, pitch smoothing, and time-scale modification demonstrate qualitative improvements in speech quality (see Table 3).

Emerging systems from global tech leaders, such as Google Cloud TTS, ElevenLabs, and Voicebox by Meta have extended the potential of TTS with multilingual capabilities. Several TTS platforms claim to cater to multiple language support including Tamil. Listen2it (2024), ResponsiveVoice (2024), and TTSFree (2024) claim to provide various voices for Sri Lankan, Indian, Malaysian, and Singaporean Tamil. However there is no benchmark dataset to assess the dialect wise performance. Even within the same country, multiple regional dialects prevail. Similarly, Speechify (2024) offers different Tamil male and female voices, while Speakatoo (2024) claim to focuses on natural-sounding speech synthesis. Kapwing (2024) and Veed.io (2024) include Indian Tamil male and female voices along with other regional variants. Platforms such as Play.ht (2024) and eSpeak (2024) continue to support multilingual TTS solutions with flexible voice options, while Flite (2024) provides a lightweight alternative for speech synthesis. Furthermore, AI (2024) utilizes neural TTS with large language models to offer over 100 languages and accents, including 27 emotional voices, showcasing advancements in expressive and personalized speech synthesis.

Despite the considerable progress in Tamil Text-to-Speech (TTS) systems, several limitations remain. Firstly, expressive speech synthesis is still underexplored—most models prioritize intelligibility and naturalness, but fall short in conveying nuanced emotions, which are essential for applications such as audiobooks, voiceovers, and storytelling. Secondly, there is limited dialectal coverage, with many models trained predominantly on standard Indian Tamil, often overlooking regional variants such as Sri Lankan, Malaysian, and Singaporean Tamil. This lack of dialectal representation reduces the adaptability and inclusivity of current systems.

Moreover, many models rely heavily on proprietary datasets or the IndicTTS corpus, which lacks phonetic and prosodic diversity across speakers and dialects. This results in a generalization gap, especially when synthesizing speech for underrepresented dialects or speaker types. Gender imbalance in datasets is another concern, as some models are trained only on male or female voices, limiting their range and expressiveness.

Evaluation practices is also challenging as there is no standardized benchmarking framework specifically to Tamil or other Indic languages, making it difficult to assess system performance consistently across models. Lastly, limited transparency in commercial systems regarding training data, model architecture, and evaluation metrics hinders reproducibility and academic comparison, slowing down progress in the open research community.

Addressing these limitations requires a concerted effort toward building more inclusive datasets, creating Tamil-specific evaluation protocols, and advancing research in expressive and dialect-aware TTS synthesis.

3.3.1. Tamil TTS data sources

Government programs focused on digital inclusion, such as Bhashini (2024b) and IndicTTS (2024), are instrumental in generating speech data across multiple languages. These initiatives typically operate under Creative Commons licenses, promoting accessibility and collaboration in research and development. Platforms such as OpenSLR (2023) and the Linguistic Data Consortium (LDC) (Linguistic Data Consortium, 2023) host extensive speech datasets, which are valuable resources for researchers and developers in the field of speech technology. Platforms such as Mozilla Common Voice collect speech data through crowdsourcing, where volunteers contribute recordings. These datasets are typically released under open licenses allowing for their free use and modification (Ardila et al., 2020).

3.3.2. Pre-processing

In the pre-processing phase, several key techniques are employed to prepare raw text for efficient speech synthesis. One such technique is text normalization, which involves converting Tamil text into a format more accessible to the TTS system by removing punctuation, expanding abbreviations, and converting numbers into words (Jalin and Jayakumari, 2017; Bhavi and Kumar, 2015).

Another critical step is phonetic transcription, in which the Tamil text is transformed into its phonetic representation through grapheme-to-phoneme conversion. This ensures more accurate speech generation by mapping Tamil characters to their phonetic sounds. For instance, a syllable-based phonetic transcription was implemented to enhance the accuracy of Tamil speech synthesis and effectively managed a unique set of Tamil phonemes (Karthikadevi and Srinivasagan, 2014).

Because Tamil is a syllable-based language, syllabification, dividing words into syllables is another crucial pre-processing technique. It aids in generating smooth prosody and maintaining rhythmic consistency in the synthesized speech. Samuel Manoharan (2022) employed syllabification in their Tamil TTS system to ensure improved pacing and pronunciation, which is essential for natural-sounding speech output.

3.3.3. Post-processing

In the post-processing phase, the quality of the synthesized speech was significantly enhanced through various methods. Prosody adjustment is a technique in which pitch, duration, and volume are modified to make the speech sound more natural. This process ensures that the speech output is fluid and engaging. Additionally, smoothing techniques are applied to reduce abrupt transitions between sounds, further enhancing the naturalness of the speech output (Arulprakash et al., 2023; Prakash et al., 2023; Sharma et al., 2014).

Emotion recognition and synthesis play a pivotal role in post-processing by adding emotional context to the synthesized speech, making it more expressive and improving listener engagement (Arulprakash et al., 2023). After synthesis, techniques such as noise reduction, voice modulation, and compression are used to refine the overall voice quality. For example, Prakash et al. (2023) utilized FastSpeech2 models to enhance the voice quality of Tamil TTS systems using voice compression and smoothing algorithms.

Tamil TTS systems often employ unit-selection or HMM-based concatenation methods to assemble speech segments, ensuring a natural flow of speech (Jalin and Jayakumari, 2017). Additionally, techniques such as post-emphasis filtering and smoothing are applied to reduce

Table 3
Comparison of text-to-speech models.

TTS Name	Details
DON Lab's Model (Prakash and Murthy, 2020)	**Voice & Acoustic Characteristics:** Male and Female Voices available **Model/Approach Used:** Tacotron2 with WaveGlow **Dataset:** IndicTTS Dataset **Reported Evaluation Score:** MOS: 3.01, MCD: 11.92, F0: 0.35, CER: 0.245 **Review:** The core idea is to build robust TTS models by combining data from multiple languages within the same language family (Indo-Aryan and Dravidian). This approach leverages shared phonotactics across related languages, improving synthesis quality, especially in low-resource scenarios. Limitations: Lower MOS (3.01) and higher MCD (11.92) compared to AI4Bharat-TTS.
AI4Bharat-TTS (Kumar et al., 2023)	**Voice & Acoustic Characteristics:** Trained with male and female voices **Model/Approach Used:** FastPitch with HiFi-GAN V1 **Dataset:** IndicTTS Database (20.33 hours: Male: 10.3 h, Female: 10.03 h) **Reported Evaluation Score:** MOS: 3.84, MCD: 10.57, F0: 0.38, CER: 0.107 **Review:** The paper investigates TTS systems for Indian languages, evaluating various acoustic models and vocoders. FastPitch was identified as the best acoustic model and HiFi-GAN V1 as the best vocoder across 13 languages. The resulting monolingual multi-speaker models significantly outperform existing open-source TTS models in naturalness and intelligibility. The models and code are open-sourced on the Bhashini platform. **Limitations:** The evaluation focused mainly on naturalness (MOS) and intelligibility (CER). Other aspects such as expressive speech, voice cloning, and generalization to unheard speakers were not thoroughly explored.
Vakyansh's Model (Open Speech EkStep, 2021)	**Voice & Acoustic Characteristics:** Not explicitly specified **Model/Approach Used:** Glow-TTS with HiFi-GAN **Dataset:** IndicTTS Dataset **Reported Evaluation Score:** MOS: 2.59, MCD: 7.28, F0: 0.108, CER: 0.134 **Review:** Vakyansh's model uses Glow-TTS with HiFi-GAN, targeting Indian languages via the IndicTTS dataset. While intelligibility is acceptable (CER: 0.134), the model scores low in naturalness (MOS: 2.59) with high MCD. Pitch accuracy is moderate (F0: 0.108). It needs refinement to enhance naturalness. **Limitations:** Lower MOS (2.59) indicates less natural-sounding speech compared to other models.
Bilingual Tamil and English TTS (Ramakrishnan et al., 2010)	**Audio Feature:** Phonetically rich, segmented and annotated at the phone level; varied and natural distribution of prosodic and spectral characteristics of speech sounds; includes smoothing the pitch contour **Model/Approach Used:** Concatenation-based with polyphone as the unit of concatenation; unit selection algorithm and prosody-based unit selection algorithm; grapheme-to-phoneme converter module **Dataset:** CIIL Tamil text corpus, resulting in 1026 sentences; 5 h of Tamil sentences recorded by a male professional Tamil speaker **Reported Evaluation Score:** High mean opinion scores from native Tamil evaluators; Intelligibility of the TTS is quite high and it is also acceptably natural **Review:** The developed TTS engine handles unlimited vocabulary for Tamil and Kannada using a unit selection based framework with phonetic rules and a recorded database **Limitations:** Limited dataset size (1026 sentences, 5 h of recordings). Focuses only on a male speaker, lacking gender diversity. cannot handle numerals, proper nouns, or foreign words, and struggles with intonation for questions/exclamations and natural pausing in long sentences.
Thirukkural II (Prathibha et al., 2002)	**Voice & Acoustic Characteristics:** Voice quality, prosody, intelligibility, co-articulatory coherence, fundamental frequency of voicing, duration, energy, and formant positions **Model/Approach Used:** Concatenative speech synthesis using unit selection, pitch modification using Discrete Cosine Transform (DCT), voice transformation, text normalization, and prosodic rules **Dataset:** Not specified **Reported Evaluation Score:** Claiming to have intelligible and acceptably natural speech , no scores reported **Review:** According to authors, it has the capability to produce different voice types such as female, child, and elderly male. Additionally, it can synthesize many proper nouns derived from other languages. It is also capable of reading digits present in the text. The synthesized speech is claimed to be acceptably natural by incorporating prosodic rules, but no scores reported. **Limitations:** Emotional synthesis is yet to be done.
Google Cloud Text-to-Speech (Google, 2024a)	**Voice & Acoustic Characteristics:** Indian Tamil, Human-like intonation, Near human quality, API available **Model/Approach Used:** DeepMind **Dataset:** Not Specified **Reported Evaluation Score:** Not Specified **Limitations:** Not Specified
Bhashini IITM (Bhashini, 2024a)	**Voice & Acoustic Characteristics:** Male and female voice for Indian Tamil, Indic TTS for Indian Languages **Model/Approach Used:** FastSpeech2 with hybrid segmentation **Dataset:** Indic TTS Dataset **Reported Evaluation Score:** Not Specified **Limitations:** Not Specified
OpenAI TTS (OpenAI, 2024b,a)	**Voice & Acoustic Characteristics:** 6 built-in voices **Model/Approach Used:** tts-1 and tts-1-hd models **Dataset:** Not Specified **Reported Evaluation Score:** Not Specified **Limitations:** Users can only choose from six built-in voices, with limited control over pitch, speed, or prosody. The dataset and training details are not publicly available, limiting transparency and customization.

(*continued on next page*)

Table 3 (*continued*).

TTS Name	Details
Improved Tamil Text-to-Speech Synthesis (Krithiga and Geetha, 2004)	**Voice & Acoustic Characteristics:** Linear Predictive Coefficient (LPC) and residuals, syllable pitch were considered **Model/Approach Used:** Concatenative speech synthesis using syllable units with diphone insertion for smoothing; time scale modification for pitch adjustment **Dataset:** A diphone database of Tamil speech was recorded from a female speaker. Around 1000 diphones (CV–CV combinations) have been created, with a target of 3000 **Reported Evaluation Score:** Qualitative improvements in speech quality and smoothness **Review:** This model employs concatenative speech synthesis using syllable-based units, enhanced by diphone insertion for smoother transitions. It incorporates Linear Predictive Coefficients (LPC), residuals, and syllable-level pitch modifications, contributing to improvements in synthesized speech quality and naturalness. The primary strength lies in its careful attention to pitch and prosodic adjustments through time-scale modification. However, the approach is limited by a relatively small dataset comprising around 1000 diphones recorded exclusively from a single female speaker, limiting generalizability and speaker variability. Future extensions would benefit significantly from expanding the database size, incorporating multiple speakers, and diversifying gender representation to enhance robustness and wider usability. **Limitations:**Limited dataset size (1000 diphones from a female speaker). Focuses only on a female speaker, lacking gender diversity.
Unit Selection and HMM-Based Tamil Speech Synthesizer by SSN (Prakash et al., 2014)	**Voice & Acoustic Characteristics:** Caters to male and female speakers **Model/Approach Used:** Unit Selection and HMM; an Android application has been developed with the Tamil HMM-based Speech Synthesis System (HTS) in a male speaker's voice **Dataset:** Not Specified **Reported Evaluation Score:** Not Specified **Review:** Not Specified **Limitations:** Not Specified
Narakeet (Narakeet, 2024)	**Voice & Acoustic Characteristics:** Different male and female voices **Model/Approach Used:** Not Specified **Dataset:** Not Specified **Reported Evaluation Score:** Not Specified **Limitations:** Not Specified
Dubverse (Dubverse, 2024)	**Voice & Acoustic Characteristics:** Different voices in SL, Indian, Malaysian, Singapore Tamil **Model/Approach Used:** Not Specified **Dataset:** - Proprietary AI Voice Model: NeoDub **Reported Evaluation Score:** Not Specified **Limitations:** Proprietary model and dataset
ElevenLabs (ElevenLabs, 2024)	**Voice & Acoustic Characteristics:** Emotionally rich audio and natural AI voices, Multilingual model (32 languages) **Model/Approach Used:** Eleven multilingual v2, Eleven Flash v2 **Dataset:** Not Specified **Reported Evaluation Score:** Not Specified **Limitations:** Proprietary model with limited transparency on dataset and evaluation metrics.
LOVO AI (AI, 2024)	**Voice & Acoustic Characteristics:** 100+ languages/accents, 27 emotional voices **Model/Approach Used:** Neural TTS with large language models **Dataset:** Not Specified **Reported Evaluation Score:** Not Specified
Micmonster, Listen2it, veed.io, playht (Micmonster, 2024; Listen2it, 2024; Veed.io, 2024; Play.ht, 2024)	**Voice & Acoustic Characteristics:** Sri Lankan, Indian, Malaysian, Singapore Tamil **Model:** Not specified **Dataset:** Not specified **Evaluation:** Not specified
Kapwing (Kapwing, 2024)	**Voice & Acoustic Characteristics:** : Claim to generate multiple accents and regional variations **Model/Approach Used:** Partnered with Elevenlabs **Dataset:** Not Specified **Reported Evaluation Score:** Not Specified
eSpeak (eSpeak, 2024)	**Voice & Acoustic Characteristics:** more than hundred languages and accents **Model/Approach Used:** formant synthesis **Dataset:** No Dataset requires **Reported Evaluation Score:** Not Specified **Reported Evaluation Score:** naturalness is relatively low
Flite (Flite, 2024)	**Voice & Acoustic Characteristics:** worser than festival **Model/Approach Used:** light weight version of festival **Dataset:** Not Specified **Reported Evaluation Score:** Not Specified

sharp transitions in the audio waveform, resulting in more natural and cohesive speech output. Sharma et al. (2014) discussed how these post-processing techniques are instrumental in optimizing the naturalness of Tamil TTS systems.

4. Evaluation of existing available Tamil TTS systems

To evaluate available Tamil TTS systems, we use Word Error Rate (WER) and Comparative Mean Opinion Score (CMOS) as primary metrics. WER provides an objective measure of intelligibility at the word level, which is more relevant than Character Error Rate (CER) for assessing spoken output as it penalizes the error more than character error rate. CMOS is preferred over MOS as it enables more reliable comparisons between systems by reducing listener bias through direct audio pair comparisons.

4.1. Evaluation of existing available Tamil TTS using word error rate (WER)

OpenAI TTS (OpenAI, 2024a,b), Google Cloud TTS (Google, 2024b), Google Text-to-Speech (gTTS) (Google, 2024c), Meta's MMS (Pratap et al., 2024), Bhashini-IITM TTS for Tamil (Bhashini, 2024a), and IIT Madras Indic TTS (Prakash and Murthy, 2020; IIT Madras, 2024) were evaluated using the Word Error Rate (WER).

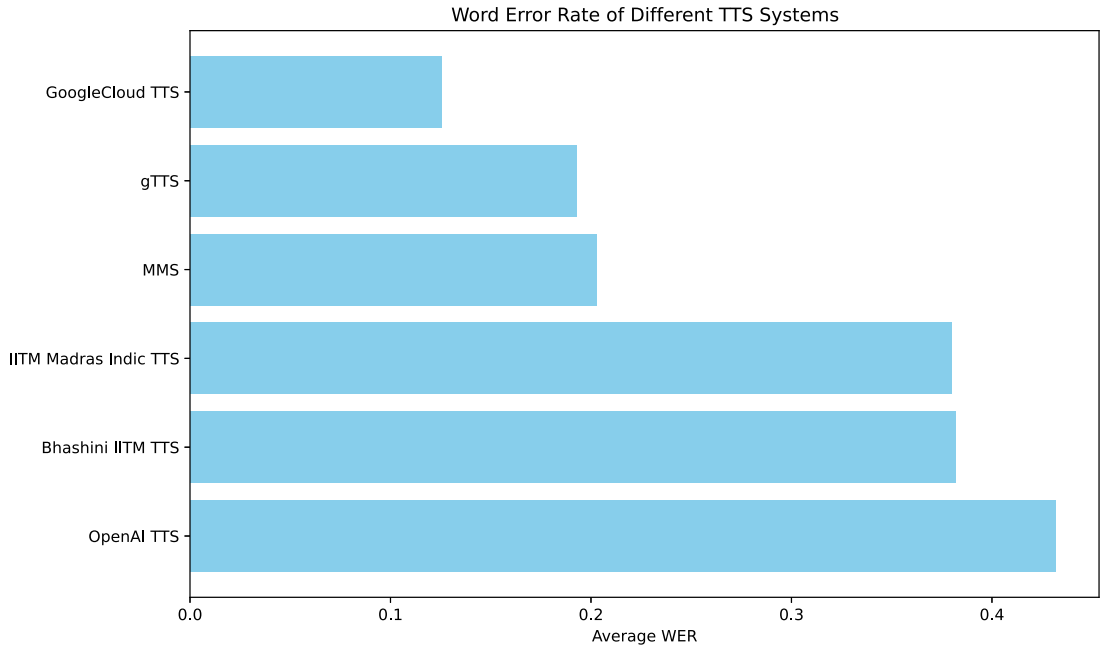

Fig. 1. Word Error Rate (WER) of different TTS systems.

4.1.1. Data collection and creation

The text for this dataset was sourced from *Oru Yogiyin Suyasarithai* (Paramahansa Yogananda, 2024) and *Ezhuna* media articles (Ezhuna, 2024), ensuring a diverse and balanced selection across multiple domains, including business, medical, science, agriculture, and education. This approach prevents the overrepresentation of any single field, making the vocabulary coverage more representative of real-world language use. To further increase the representativeness and coverage, the dataset includes numerical data and code-switched content, both of which are commonly found in everyday speech. Numbers frequently appear in financial, medical, and technical contexts, requires TTS systems to correctly pronounce different formats. Code switched reflects the natural way Tamil speakers integrate English words, making the dataset more realistic for multilingual and code-switched speech synthesis. Additionally, proper names also included. The collected text was synthesized into audio using multiple TTS systems such as Bhashini, IITM, OpenAI, GoogleCloud, MMS, and gTTS.

4.1.2. Methodology

The synthesized audio was transcribed using the Automatic Speech Recognition (ASR) system—Bhashini-IITM ASR (wav2vec) Tamil v1.0 (Ministry of Electronics and Information Technology, India, 2024) and ASR errors referring to the synthesized audio were manually corrected. Subsequently, the original transcription was used for an accurate comparison. Before calculating WER, pre-processing the text included Unicode normalization, punctuation, and extra space removal was carried out. The word Error Rate (WER) was calculated by comparing the transcriptions from TTS systems with the original transcription, using a formula that accounts for substitutions, insertions, and deletions relative to the total reference words.

4.1.3. Interpretation of results

Fig. 1 illustrates the average Word Error Rate (WER) for various Text-to-Speech (TTS) systems, where lower WER values indicate better performance and higher intelligibility. Among the evaluated systems, Google Cloud TTS achieves the lowest WER, demonstrating the highest accuracy and most reliable performance. gTTS and MMS also show strong performance with relatively low WERs, placing them among the top-performing systems. In contrast, OpenAI TTS, IIT Madras Indic TTS, and Bhashini IITM TTS exhibit higher WER values, indicating reduced accuracy and a need for further improvement. Overall, Google Cloud, gTTS, and MMS stand out for their superior intelligibility, while other systems show varying degrees of performance, with OpenAI TTS being the least accurate among the compared models.

4.2. Evaluation of existing Tamil TTS using comparative mean opinion score(CMOS)

The OpenAI TTS (OpenAI, 2024a,b), Google Cloud TTS (Google, 2024b), Google TTS (Google, 2024c), Meta's MMS (Pratap et al., 2024), Bhashini TTS (Bhashini, 2024a), IIT Madras Indic TTS (Prakash and Murthy, 2020; IIT Madras, 2024), and human voices were evaluated by performing user surveys to calculate the Comparative Mean Opinion Score (CMOS). The evaluation compares the perceived quality of voices pairwise across naturalness, intelligibility, expressiveness, low listening fatigue, and overall quality, as described in Table 4.

4.2.1. Speech data collection

Human voices were collected from *Oru Yogiyin Suyasarithai* (Paramahansa Yogananda, 2024) and *Ezhuna* media (Ezhuna, 2024). These voices contain vocabulary from different domains, loan words, and various expressions such as questions, anger, and declarative sentences. TTS voices are synthesized using the text corpus of human speech chosen for evaluation. The dataset includes words and phrases from various domains such as livestock farming, climate, geography, politics, news, agriculture, and business. This diverse vocabulary ensures that TTS systems are evaluated on a broad linguistic spectrum, enhancing their robustness and generalization capabilities. The dataset incorporates sentences exhibiting different emotional tones, such as anger பகவதி, நீ உன்னிடம் வேலை செய்பவருடன் மிக கடுமையாக இருக்கிறாய் *Bagavathi, ni unnidam velai seibavarudan miga kadumayaka irukrai* "Bhagavati, you are being very harsh with the person working under you"., and various sentence structures, including interrogative and declarative forms. This inclusion facilitates the evaluation of TTS systems' ability to generate expressive and prosodically accurate speech.

4.2.2. Survey participants selection

The study involved 30 participants above 18 years of age. Participants were selected from key districts where Tamil spoken in Sri Lanka

Table 4
Evaluation aspects of TTS systems.

Evaluation Aspect	Meaning	Use Case in TTS Evaluation
Naturalness	Naturalness encompasses aspects such as smooth transitions between sounds, appropriate speaking rate, and absence of robotic or distorted artifacts	Used to evaluate whether the TTS output sounds natural and human-like. Essential for applications requiring seamless human interaction (e.g., virtual assistants).
Intelligibility	Assesses how easy it is to understand the words and meaning conveyed by the synthesized speech. Intelligibility depends on factors such as clear articulation, correct pronunciation, and lack of mispronunciations or unintelligible sounds.	Critical for TTS systems used in navigation aids, accessibility tools for the visually impaired, and emergency response systems.
Expressiveness	Evaluates the system's ability to produce the rhythm, stress, and intonation of the synthesized speech. Prosody includes variations in pitch, loudness, timing, and pauses that contribute to the expressiveness and natural flow of the speech	Important for applications such as storytelling, audiobooks, or character voices in gaming where emotional context is crucial.
Listening Fatigue	Listening fatigue means any discomfort caused in your ear and the tiredness caused by listening to the voice for a long period, not due to any device or meaning of the audio	Relevant for scenarios where users need to listen for extended periods, such as online learning platforms, podcasts, audio books or customer service.
Overall Quality	Overall quality of speech defines how well the speech is perceived and understood by listeners, audio quality how natural speech lookalike.	Useful for benchmarking TTS systems in general-purpose applications such as voice assistants, IVR systems, and entertainment platforms.

including Jaffna and Vavuniya (Northern Province), Batticaloa and Ampara (Eastern Province), Kegalle (Central Province), and Gampaha and Colombo (Western Province). This distribution was carefully designed to reflect the major dialectal distinctions identified by Kailainathan (1999), who classifies Tamil dialects in Sri Lanka primarily along geographical and racial (ethnic) lines rather than caste-based divisions. According to Kailainathan, geographical dialects vary between regions such as the Northern, Eastern, Western, and Central provinces, while racial dialects arise from differences between ethnic communities, particularly between Sri Lankan Tamils and Muslims. Tamil-speaking Muslims, who are predominantly located in the Eastern and Western regions, exhibit speech patterns distinct from Tamil-speaking non-Muslim communities. Despite the ethnic and regional diversity, Kailainathan emphasizes that caste-based dialectal variation is not a prominent feature in Sri Lankan Tamil, distinguishing it from Tamil dialects in South India, where caste distinctions in speech such as Brahmin vs. non-Brahmin Tamil are more pronounced (Kailainathan, 1999). A visual representation of the participants' regional origins (districts) is provided in 2, highlighting geographical spread. The participant pool encompassed a representation of a mix of genders (43.33% male and 56.66% female), age groups (ranging from 18 to 65 years), from various professions including linguists and it ensured the evaluation of synthesized speech from multiple perspectives. To ensure informed and reliable feedback, only individuals aged 18 years and above were included in the study. As until 18, according to Sri Lankan education system, participants will be school students. This age criterion was selected based on the assumption that individuals within this range have adequate exposure to Tamil in both formal and informal contexts, making them well-suited to evaluate the synthesized speech. A total of 30 participants were recruited for the study, a sample size aligned with established practices in subjective evaluation studies for TTS systems. Prior research suggests that a participant group of 20–30 individuals is typically sufficient to yield statistically meaningful results while maintaining feasibility in terms of time and resource constraints (Shahid et al., 2016). Participants were recruited through university networks and social media platforms. Invitations to participate in the study were disseminated through email and social media messages, providing potential participants with a brief overview of the study's objectives and evaluation procedures. A total of around 40 individuals were invited to take part in the survey, of whom 30 responded positively and completed the evaluation process.

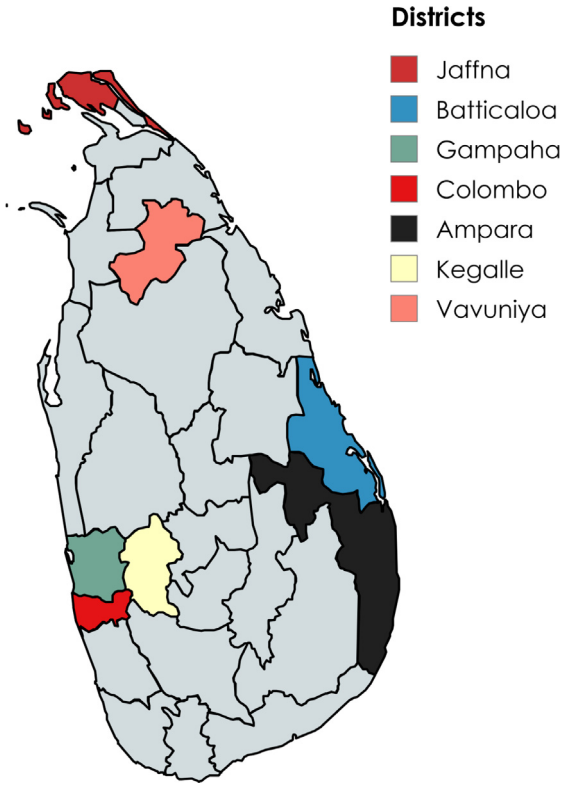

Fig. 2. Major Tamil speaking regions in Sri Lanka from where evaluators are selected.

4.2.3. Survey methodology

To ensure unbiased results, the speech data both TTS and human voices were shuffled and presented to participants without labeling, thus keeping the identity of the voices anonymous. Before participating, respondents provided demographic information, including age, gender, native language (Tamil), previous experience with TTS technologies, and how frequently they engaged with recorded or automated speech, such as audiobooks or virtual assistants. The participants were asked to rate given voices on multiple dimensions by pairwise comparison.

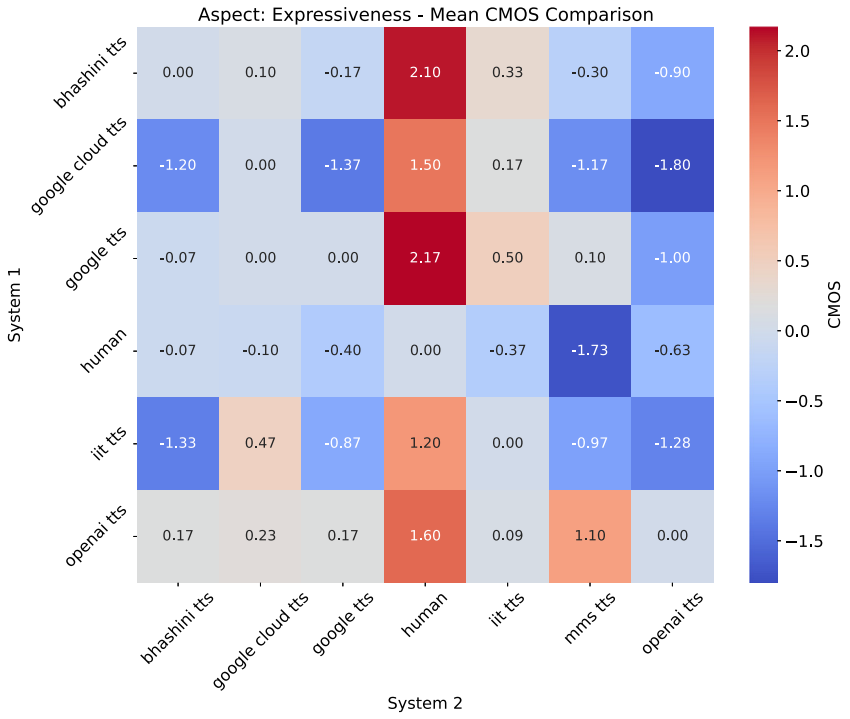

Fig. 3. CMOS between tested audios on Expressiveness.

Each sub-section of the survey denotes an aspect of the speech that we assess. Participants were tasked with rating the voices based on these dimensions, using a scale from −3 to +3. This scale indicates their preferences: −3 indicates a strong preference for the first voice, 0 indicates no preference, and +3 indicates a strong preference for the second voice.

4.2.4. Interpretation of results

The scores for each compared audio pair were averaged among all participants to calculate the mean CMOS for each criterion, providing a comprehensive view of the preferences between the voices. The CMOS revealed the overall tendency of participants towards specific voices in each pair based on various assessed criteria.

Fig. 3 presents the CMOS for tested systems and the human speech on the aspect of expressiveness, where the System 2 is Comparandum (The entity being compared) and the System 1 is Comparatum (The entity it is compared to).

The rows represent System 1, and the columns represent System 2. The intensity of the color indicates the mean CMOS scores:warmer tones (red) suggest that System 1 is preferred and cooler tones (blue) suggest that System 2 is preferred. The neutral values were near-white. Each cell displays the mean CMOS Score, providing precise quantitative details for the comparison. Human voices remained significantly better. The IIT Madras Indic TTS performs slightly lower than human in terms of expressiveness. MMS and Google Cloud shows moderately lower performance than human in expressiveness. OpenAI and Google cloud moderately lower in expressiveness than human. Google and Bhashini performs relatively low in expressiveness compared to human speech. Fig. 4 presents the CMOS for tested systems and the human speech on the aspect of listening fatigue. This heatmap reveals human speech outperforms or performs equally in most comparisons. Only Human vs Google TTS resulted Google TTS perform good in one instance. Fig. 5 illustrates the heatmap for mean CMOS Scores on Intelligibility and highlights notable trends across various TTS systems. Human speech outperforms all the TTS systems in intelligibility. OpenAI is relatively less intelligible compared to other systems.

Fig. 6 illustrates CMOS between tested items on Naturalness. The heatmap for mean CMOS Scores on Naturalness highlights that human speech almost outperforms all the TTS evaluated in naturalness. IIT Madras Indic TTS performs comparatively higher than the other TTS systems. OpenAI performs lesser comparatively to other TTS systems.

Fig. 7 illustrates CMOS between tested items on Overall Quality. "Human" generally outperforms most synthetic systems. Generally OpenAI TTS performs less compared to all the TTS systems.

4.2.5. Inter-rater agreement score calculation

In subjective evaluations like the CMOS, inter-rater agreement is important to ensure the reliability of human ratings, especially when assessing perceptual qualities that are prone to individual interpretation. We computed inter-rater agreement measuring consistency both globally and across specific perceptual dimensions. While global evaluation provides an overall reliability snapshot of the evaluation process, per-aspect evaluations values help identify which attributes are consistently interpreted and which, may require improved rater training. Thus, inter-rater agreement analysis not only validates the results but also guides the refinement of evaluation protocols for future studies.

To evaluate the reliability of the subjective evaluation metric CMOS, we considered several inter-rater reliability measures: Percentage Agreement, Cohen's Kappa (κ) (Cohen, 1960), Fleiss' κ (Fleiss, 1971), the Intraclass Correlation Coefficient (ICC) (Shrout and Fleiss, 1979), and Krippendorff's Alpha (Krippendorff, 2004).

Percentage Agreement offers a straightforward computation of the proportion of identical ratings between raters. However, it fails to adjust for chance agreement, making it unreliable for scientific evaluations and often overestimating the degree of consensus among raters (Gwet, 2014).

Cohen's κ is a widely used statistic for measuring agreement between two raters assigning categorical labels. It adjusts for chance agreement and is relatively simple to interpret. Nonetheless, its applicability is limited to two raters and cannot be extended directly to multi-rater setups without averaging multiple pairwise agreements. Furthermore, Cohen's κ treats labels as nominal, meaning it considers only whether raters agree or disagree, not how far apart the disagreements are. This is problematic for ordinal data such as CMOS, where disagreement between −1 and −2 is less severe than between −3 and

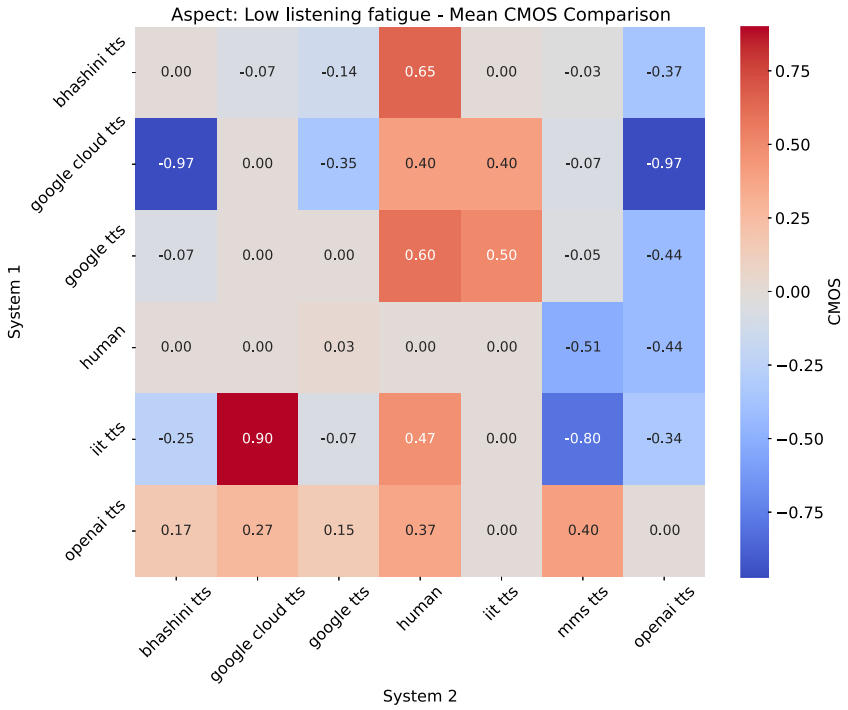

Fig. 4. CMOS between tested audios on Listening Fatigue.

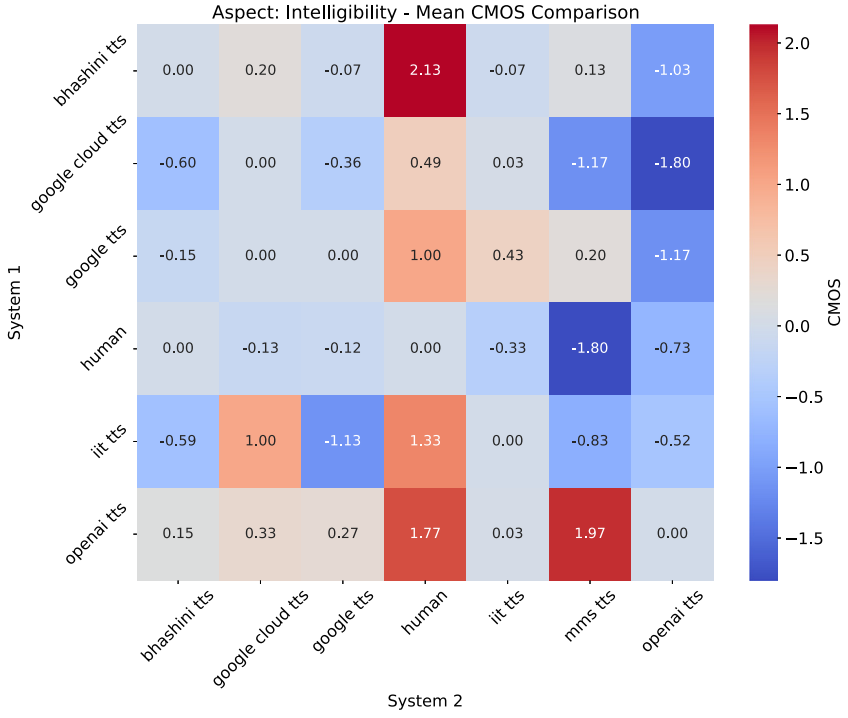

Fig. 5. CMOS between tested audios on Intelligibility.

3. Nominal treatment ignores this gradation, thus misrepresenting the severity of disagreement (Carletta, 1996).

Fleiss'κ generalizes Cohen's κ to handle more than two raters and is suited for nominal-scale data like Cohen's κ, Fleiss' κ does not account for the ordered nature of ordinal data, limiting its appropriateness for CMOS-style evaluations (Fleiss, 1971).

The Intraclass Correlation Coefficient (ICC) is better suited for continuous or interval-scale ratings and is not appropriate for ordinal data, which are common in subjective evaluations like CMOS (Shrout and Fleiss, 1979).

Therefore, we used Krippendorff's alpha, a flexible and robust metric of inter-rater agreement. Krippendorff's Alpha supports any number of raters, accommodates missing data, and allows for various data types including ordinal, which makes it especially suitable for the CMOS scale (Krippendorff, 2004; Artstein and Poesio, 2008). In our evaluation, Krippendorff's Alpha was computed overall global and per perceptual aspect, enabling an accurate measurement of inter-rater

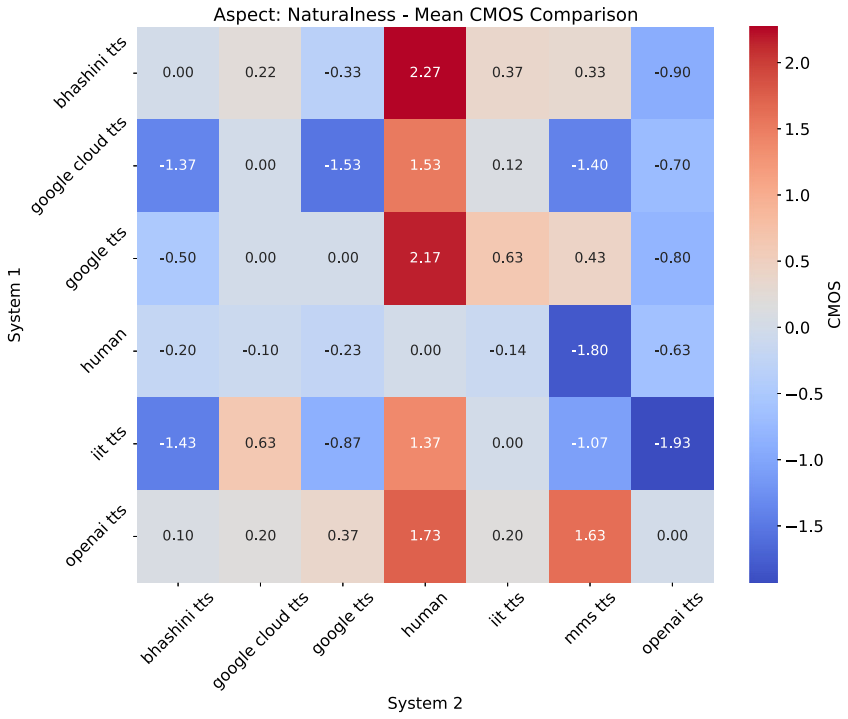

Fig. 6. CMOS between tested audios on Naturalness.

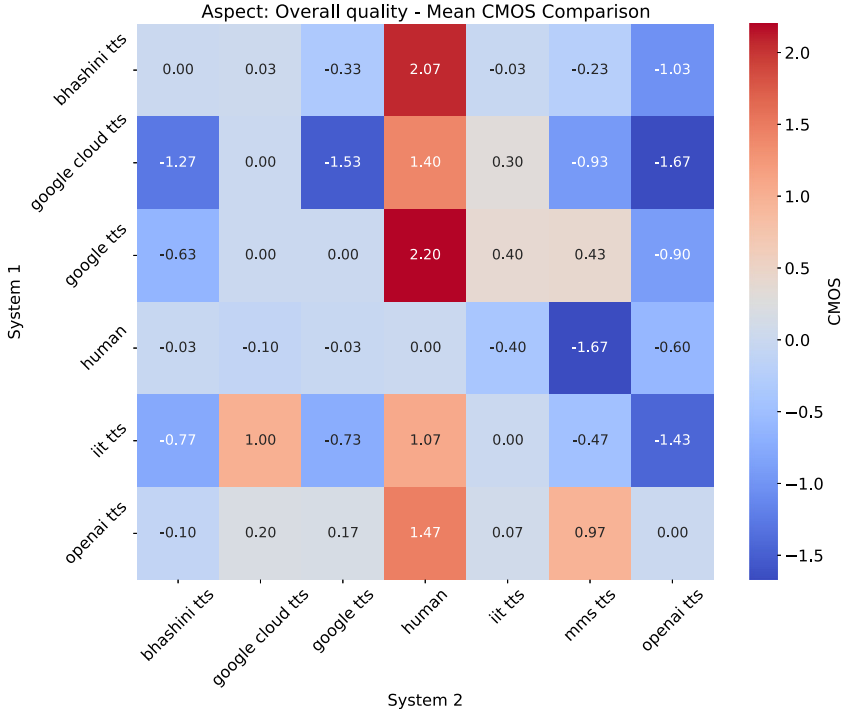

Fig. 7. CMOS between tested audios on Overall Quality.

consistency in a multi-rater, ordinal-rating environment. The overall global agreement provides a single summary measure of consistency across all perceptual aspects evaluated. It helps to assess the general reliability of the entire rating process rather than focusing on individual aspects alone. Considering the global Alpha is important because it reflects how much raters agree overall, indicating the trustworthiness of the subjective evaluation as a whole. By calculating Alpha for each perceptual aspect, we can assess which qualities were consistently perceived and which might require clearer definitions or better rater training. High Alpha values indicate strong agreement and reinforce the validity of the conclusions drawn about TTS system performance. Conversely, low agreement flags aspects that may need clearer definitions. Thus, Krippendorff's Alpha serves not only as a reliability check but also as a diagnostic tool to improve future evaluation design and system development.

The calculation of Krippendorff's Alpha involves comparing the observed disagreement among raters calculated using Eq. (1) with the disagreement expected by chance calculated using (2), using a distance

Table 5
Krippendorff's alpha for each perceptual aspect in TTS evaluation.

Perceptual Aspect	Krippendorff's Alpha
Expressiveness	0.5799
Intelligibility	0.5899
Listening Fatigue (Low)	0.1226
Naturalness	0.5545
Overall Quality	0.5220

Global Krippendorff's Alpha: 0.4848.

function appropriate to the scale of the data. For ordinal ratings, we employed a squared difference function to penalize larger rating discrepancies more heavily. A coincidence matrix was constructed to capture how often rating pairs co-occurred across raters, and Alpha was computed using all valid pairings. This procedure was applied both per aspect, to assess consistency in individual perceptual dimensions, and globally, by aggregating all ratings into a single analysis (Krippendorff, 2004) using Python library krippendorff (Castro, 2017).

Observed Disagreement (D_o)

The observed disagreement is computed by averaging the pairwise distances across all voice pairs (items):

$$D_o = \frac{\sum_i \sum_c \sum_{c'} N_{i,cc'} \delta^2(c, c')}{\sum_i n_i(n_i - 1)} \quad (1)$$

(Krippendorff, 2011)

where:

- i indexes the items being rated,
- $N_{i,cc'}$ is the number of times categories c and c' were used by different raters for item i,
- n_i is the number of raters who rated item i.

Expected Disagreement (D_e)

Expected disagreement is calculated based on the marginal probabilities of each category across the entire dataset:

$$D_e = \sum_c \sum_{c'} p_c p_{c'} \delta^2(c, c') \quad (2)$$

(Krippendorff, 2011)

where p_c and $p_{c'}$ are the proportions of ratings assigned to categories c and c', respectively.

The interpretation of Krippendorff's Alpha values follows established guidelines indicating the strength of inter-rater agreement. An Alpha value of 0.80 or higher signifies almost perfect reliability, suggesting that the ratings can be considered highly reliable across raters. Values ranging from 0.667 to 0.80 allow for tentative conclusions, implying moderate reliability where findings may be cautiously interpreted but require further validation. When Alpha falls below 0.667, reliability is considered low to draw trustworthy conclusions (Krippendorff, 2004).

According to Table 5,based on the computed Krippendorff's Alpha values for each perceptual aspect, we observe varying levels of inter-rater agreement. For expressiveness, intelligibility, naturalness, and overall quality, the agreement among raters can be considered moderate agreement. However, all of these values remain below the commonly accepted threshold of 0.667, indicating that it may be not reliable to draw any trustworthy conclusions. In the case of low listening fatigue, the alpha value is notably low, reflecting very slight agreement, which suggests substantial variability in rater judgments for this aspect. Alpha values below 0.667 indicate that the reliability of the ratings is limited, and caution should be exercised when interpreting results based on such data (Krippendorff, 2004).

The low Krippendorff's Alpha values below 0.667 across perceptual aspects indicate limited reliability in the ratings, likely due to the subjective nature of evaluating synthetic voices. Factors such as varying exposure to synthetic speech, differences in linguistic expertise (linguists vs. non-linguists), and diverse geographic backgrounds can influence how raters perceive voice qualities. Additionally, individual differences in familiarity with other languages and personal sensitivity to voice characteristics make judgments more human-dependent and less reliable. This is especially true for low listening fatigue, where the very low value reflects the highly personal and variable experience of fatigue, influenced by individual auditory tolerance and prior exposure to synthetic voice and podcasts, leading to greater inconsistency in ratings compared to other perceptual aspects.

5. Linguistic challenges in Tamil TTS

Tamil, with its rich linguistic qualities, presents significant challenges for Text-to-Speech (TTS) systems due to its phonology and diglossic nature (Catherine and Abirami, 2023). While these features enhance the language's cultural richness, they introduce complexities for computational modeling in TTS systems (Madhavaraj et al., 2022).

5.1. Dialectal variations

Tamil features extensive dialectal diversity influenced by regional and social factors. These dialects vary in pronunciation, vocabulary, and syntax, requiring TTS systems to model a broad spectrum of linguistic nuances. Studies in computational phonology highlight the importance of large-scale speech datasets that reflect this variability across regions and social groups (Renganathan, 2008).

Dialectal variation affects how TTS systems synthesize regionally appropriate pronunciations. Phonetic features like vowel harmony and regional sound shifts can significantly influence the perceived naturalness of generated speech, making robust modeling essential (Ramakrishnan et al., 2007).

5.2. Diglossia

Tamil's diglossic nature, characterized by distinct formal and colloquial varieties, introduces another layer of complexity. TTS systems must be capable of generating both formal and colloquial speech, which often requires separate training datasets. This challenge is compounded by the lack of standardized orthography in colloquial Tamil, where spelling and usage vary across regions and social groups (Catherine and Abirami, 2023).

For example, the formal sentence அவன் போவான் *avan povaan* "He will go" may appear colloquially as அவன் போஆன் *avan poan*, requiring TTS systems to learn flexible phoneme–grapheme mappings to ensure natural synthesis.

5.3. Phonetic complexity

Tamil's phonetic system includes a one-to-many grapheme-to-phoneme mapping, meaning the same character can be pronounced differently depending on context. Stops and fricatives can change articulation within words (Ronanki et al., 2016).

Tamil also features five basic vowels with length distinctions, which are phonemic and affect meaning. Proper modeling of coarticulatory effects and vowel harmony is essential for producing natural and intelligible speech (Madhavaraj et al., 2022).

5.4. Unique phonemes

Tamil contains phonemes not common in other languages, such as retroflex consonants and the distinctive letter ழ *zha* ɻ . Synthesizing these accurately requires high-quality acoustic modeling. HMM-based approaches have shown success in improving the intelligibility of such sounds (Srinivasan et al., 2010).

5.5. Limited digital resources

Tamil lags behind rich resourced languages like English in TTS research primarily due to the scarcity of large, annotated linguistic corpora. This limits the training and evaluation of advanced models. Recent work on subword grammar modeling emphasizes the importance of rich, representative datasets for effective synthesis (Madhavaraj et al., 2022).

5.6. Prosody and intonation

Prosodic features such as vowel/consonant length, stress, and intonation play a crucial role in meaning. For instance, the length difference between கல் *kal* "stone" and கால் *kaal* "leg" is meaningful. Similarly, intonation distinguishes between declarative and interrogative statements, as in அவன் வந்தான் *avan vandhaan* "He came" vs. அவன் வந்தானா? *avan vandhaanaa?* "Did he come?" (Geetha and Raja, 2007).

TTS systems must model these subtle prosodic patterns to produce expressive, natural-sounding speech. This includes handling rhythmic patterns and tonal shifts, which are crucial to semantic interpretation (Catherine and Abirami, 2023).

5.7. Linguistics related limitations of current Tamil TTS systems

Overall, the development of Tamil Text-to-Speech (TTS) systems encounters numerous challenges stemming from the linguistic complexities of Tamil. Dialectal variations, influenced by regional and social factors, create significant variability in pronunciation and vocabulary, complicating the design of universal TTS systems. Tamil's diglossic nature, characterized by distinct formal and colloquial forms, necessitates separate datasets and processing techniques, further adding to the complexity. Phonetic challenges, such as the precise modeling of vowel length distinctions, retroflex consonants, and co-articulatory effects, hinder the production of natural-sounding speech. Limited annotated datasets exacerbate these issues, thereby restricting the training of high-quality models. Furthermore, current systems struggle with prosodic modeling, making it difficult to model natural rhythm, stress, and intonation. The lack of standardized linguistic subjective evaluation methods designed such as proper segmental intelligibility test, sentence level intelligibility test specific to Tamil TTS further complicates the comparative assessment of system performance. Without proper evaluation mechanisms, it becomes challenging to identify and address the underlying linguistic issues, thereby impeding the development of more effective and linguistically sound Tamil TTS systems.

6. Conclusion

This paper presents a survey of Text-to-Speech Systems for the Tamil Language and evaluation of widely used Text-to-Speech Systems for the Tamil Language. The survey covers development methodologies, system performance, and linguistic challenges. Traditional approaches such as concatenative and statistical parametric synthesis have been largely replaced by neural models like Tacotron2 and FastSpeech2, which offer improved naturalness and scalability.

A critical evaluation of existing Tamil TTS systems revealed that while some systems like MMS and Google Cloud TTS deliver strong performance, others still lag in intelligibility, naturalness, and expressiveness. The study employed both objective (Word Error Rate) and subjective (Comparative Mean Opinion Score) evaluation metrics to benchmark performance across dimensions such as intelligibility, naturalness, expressiveness, listening fatigue, and overall quality. Results indicate substantial room for improvement, especially in handling expressive and emotionally nuanced speech.

Additionally, the paper highlighted core linguistic challenges including Tamil's diglossia, dialectal diversity, and phonetic complexity, which are often underrepresented in existing TTS models and datasets. These challenges, coupled with the lack of dialect-sensitive evaluation frameworks and diversity in training data, hinder the development of inclusive and robust Tamil TTS systems.

Addressing these gaps will require the creation of diverse, annotated datasets, improved evaluation protocols, and greater emphasis on modeling regional and stylistic variations. This survey thus provides a foundation for future research aimed at building expressive, intelligible, and dialect-aware Tamil TTS systems.

Future directions for Tamil Text-to-Speech (TTS) systems include the development of large-scale, diverse datasets, which are paramount to addressing Tamil's phonetic and dialectal diversity. Enhancing multilingual integration will further improve TTS systems by enabling seamless transitions in code-switched speech. Research on advanced prosodic modeling is essential for improving the naturalness and emotional expressiveness of synthesized speech, making interactions more engaging and human-like. And, the integration of Automatic Speech Recognition (ASR) and TTS technologies can pave the way for automatic dubbing solutions, improving educational and entertainment content accessibility in multilingual contexts.

CRediT authorship contribution statement

Ahrane Mahaganapathy: Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kengatharaiyer Sarveswaran:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT and Paperpal in order to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ahrane Mahaganapathy reports financial support was provided by German Academic Exchange Service. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research study was carried out under the DigSAL project, supported by the German Academic Exchange Service (DAAD) and funded by the Federal Ministry for Economic Cooperation and Development (BMZ), Germany through SDG Partnerships (2024–2027).

References

Ahmad, H.A., Rashid, T.A., 2024. Planning the development of text-to-speech synthesis models and datasets with dynamic deep learning. J. King Saud Univ. - Comput. Inf. Sci. 36 (7), 102131. http://dx.doi.org/10.1016/j.jksuci.2024.102131.

AI, L., 2024. LOVO AI tamil text-to-speech. URL https://www.lovo.ai/ (Accessed 12 November 2024).

Amazon, 2024. Amazon Polly. URL https://aws.amazon.com/polly/ (Accessed 15 November 2024).

Annamalai, E., 1979. Linguistic Complexity of Tamil Phonology. Madras University Press.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G., 2020. Common voice: A massively-multilingual speech corpus. URL https://arxiv.org/abs/1912.06670.

Artstein, R., Poesio, M., 2008. Inter-coder agreement for computational linguistics. Comput. Linguist. 34 (4), 555–596. http://dx.doi.org/10.1162/coli.07-034-R2.

Arulprakash, A., Synthiya, M., Vijila, T., Rajabhusanam, C., 2023. Tamil speech synthesizer app for android: Text processing module enhancement. Indian J. Sci. Technol. 16 (7), 485–491. http://dx.doi.org/10.17485/IJST/v16i7.2165.

Balasubramanian, T., Asher, R., 1984. Intervocalic plosives in Tamil. In: Topics in linguistic phonetics in honour of ET Uldall: Occasional Papers in Linguistics and Language Learning, vol.9.

Baljekar, P., Black, A.W., 2016. Utterance selection techniques for TTS systems using found speech. In: 9th ISCA Speech Synthesis Workshop. pp. 184–189, URL https://www.cs.cmu.edu/~awb/papers/ssw9_baljekar.pdf.

Balyan, A., Agrawal, S.S., Dev, A., 2013. Speech synthesis: a review. Int. J. Eng. Res. Technol. (IJERT) 2 (6), 57–75. http://dx.doi.org/10.17577/IJERTV2IS60087.

Behdad, M., Gharavian, D., 2023. Improving CycleGAN-VC2 voice conversion by learning MCD-based evaluation and optimization. In: 2023 31st International Conference on Electrical Engineering. ICEE, pp. 971–976. http://dx.doi.org/10.1109/ICEE59167.2023.10334883.

Bhashini, 2024a. Bhashini IITM indic-TTS. URL https://bhashini.gov.in/ulca/search-model/6576a29700d64169e2f8f441/model (Accessed 7 October 2024).

Bhashini, 2024b. Bhashini IITM Indic-TTS. URL https://bhashini.gov.in/ulca/search-model/6576a29700d64169e2f8f441/model (Accessed 15 November 2024).

Bhavi, V., Kumar, G., 2015. Text processing for developing unrestricted tamil text to speech synthesis system. Indian J. Sci. Technol. 8, http://dx.doi.org/10.17485/ijst/2015/v8i29/72294.

Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. URL https://arxiv.org/abs/cmp-lg/9602004.

Castro, S., 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

Catherine, S.P.A., Abirami, S., 2023. Development of Phonetic Accuracy and Speech Intelligibility in Tamil Speaking Typically Developing Children. Int. J. Fundam. Multidiscip. Research (IJFMR) 5 (1), http://dx.doi.org/10.36948/ijfmr.2023.v05i01.1618.

Cherry, C., 1957. On Human Communication; a Review, a Survey, and a Criticism. The Technology Press of MIT.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46. http://dx.doi.org/10.1177/001316446002000104.

Dubverse, 2024. Dubverse tamil text-to-speech. URL https://dubverse.ai/text-to-speech/tamil/ (Accessed 13 October 2024).

ElevenLabs, 2024. ElevenLabs Tamil text-to-speech. URL https://elevenlabs.io/languages/tamil (Accessed 5 October 2024).

eSpeak, 2024. eSpeak Tamil Text-to-Speech. URL https://espeak.sourceforge.net/ (Accessed 12 November 2024).

Ezhuna, 2024. Ezhuna media articles. URL https://www.ezhunaonline.com/ (Accessed 17 September 2024).

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. 76 (5), 378–382. http://dx.doi.org/10.1037/h0031619.

Flite, 2024. Flite text-to-speech. URL https://github.com/festvox/flite (Accessed 15 November 2024).

Geetha, V., Raja, R.L., 2007. Mozhiarivial. Annamalai University Publication Division.

Gobinathan, A., Hashim, N.M.Z., Omar, A., 2024. TTS-UTeM: Text to speech transformation application. APS 44. http://dx.doi.org/10.5281/zenodo.11342295.

Google, 2024a. Google cloud text-to-speech. URL https://cloud.google.com/text-to-speech (Accessed 5 October 2024).

Google, 2024b. Google cloud text-to-speech. URL https://pypi.org/project/google-cloud-texttospeech/ (Accessed 11 October 2024).

Google, 2024c. gTTS: Google text-to-speech. URL https://pypi.org/project/gTTS/ (Accessed 29 September 2024).

Gwet, K.L., 2014. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC.

Hart, G.L., 2000. Statement on the status of tamil as a classical language. Univ. Calif. Berkeley Dep. South Asian Studies– Tamil URL https://sangamtamilliterrature.wordpress.com.

Henriksson, E., 2023. Identification and classification of TTS intelligibility errors using ASR : A method for automatic evaluation of speech intelligibility. (Master's thesis). KTH, School of Electrical Engineering and Computer Science (EECS), p. 60.

Heselwood, B., 2013. Phonetic Transcription in Theory and Practice. Edinburgh University Press, http://dx.doi.org/10.3366/edinburgh/9780748640737.001.0001.

Hoy, M.B., 2018. Alexa, Siri, Cortana, and more: An introduction to voice assistants. Med. Ref. Serv. Q. 37 (1), 81–88. http://dx.doi.org/10.1080/02763869.2018.1404391, PMID: 29327988.

Hu, Y., Chen, C., Wang, S., Chng, E.S., Zhang, C., 2024. Robust zero-shot text-to-speech synthesis with reverse inference optimization. URL https://arxiv.org/abs/2407.02243.

IIT Madras, 2024. IIT Madras Text-to-Speech Demo. URL https://asr.iitm.ac.in/demo/tts (Accessed 5 September 2024).

Inam, I.A., Azeta, A.A., Daramola, O., 2017. Comparative analysis and review of interactive voice response systems. In: 2017 Conference on Information Communication Technology and Society. ICTAS, pp. 1–6. http://dx.doi.org/10.1109/ICTAS.2017.7920660.

IndicTTS, 2024. IndicTTS - Tamil. URL https://www.iitm.ac.in/donlab/indictts (Accessed 8 October 2024).

ITU-T, 1996. Methods for subjective determination of transmission quality. Recommendation P.800, International Telecommunication Union, URL https://www.itu.int/rec/T-REC-P.800-199608-I/en (Accessed: 27 August 2024).

Jalin, A.F., Jayakumari, J., 2017. Text to speech synthesis system for Tamil using HMM. In: 2017 IEEE International Conference on Circuits and Systems. ICCS, pp. 447–451. http://dx.doi.org/10.1109/ICCS1.2017.8326040.

Kailainathan, R., 1999. Linguistic Heritage of Sri Lanka Tamil. University of Jaffna, URL http://repo.lib.jfn.ac.lk/ujrr/handle/123456789/9882 (Accessed 2025-05-19.

Kapwing, 2024. Kapwing Tamil text-to-speech. URL https://www.kapwing.com/ (Accessed 29 November 2024).

Karthikadevi, M., Srinivasagan, K., 2014. The development of syllable based text to speech system for Tamil language. In: 2014 International Conference on Recent Trends in Information Technology. pp. 1–6. http://dx.doi.org/10.1109/ICRTIT.2014.6996126.

Keane, E., 2004. Illustrations of the IPA: Tamil. J. Int. Phon. Assoc. 34 (1), 111–116. http://dx.doi.org/10.1017/S0025100304001549.

Kim, J., Lee, K., Chung, S., Cho, J., 2024. CLaM-TTS: Improving neural codec language model for zero-shot text-to-speech. http://dx.doi.org/10.48550/arXiv.2404.02781, URL https://arxiv.org/abs/2404.02781.

Krippendorff, K., 2004. Content Analysis: An Introduction to Its Methodology, second ed. SAGE Publications, Thousand Oaks, CA, p. 413.

Krippendorff, K., 2011. Computing Krippendorff's alpha-reliability. Technical report 43, University of Pennsylvania, Annenberg School for Communication, Philadelphia, PA, URL https://repository.upenn.edu/asc_papers/43/.

Krishnamurthy, E., 1977. Automatic phonetic transcription of tamil in roman script. Proc. the Indian Acad. Sci. 86 (6), 503–512. http://dx.doi.org/10.1007/BF03046906.

Krithiga, M.V., Geetha, T.V., 2004. Improved tamil text to speech synthesis. In: Tamil Internet 2004. pp. 133–139, URL https://infitt.org/ti2004/papers/19vinukritiga.pdf.

Kumar, G.K., Praveen, S., Kumar, P., Khapra, M.M., Nandakumar, K., 2023. Towards building text-to-speech systems for the next billion users. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1–5.

Kuno, S., 1958. Phonemic structure of colloquial tamil. J. Linguist. Stud. http://dx.doi.org/10.11435/gengo1939.1958.34_41.

Ladefoged, P., Bhaskararao, P., 1983. Non-quantal aspects of consonant production: a study of retroflex consonants. J. Phon. 11 (3), 291–302. http://dx.doi.org/10.1016/S0095-4470(19)30828-9.

Łajszczak, M., Cámbara, G., Li, Y., Beyhan, F., van Korlaar, A., Yang, F., Joly, A., Martín-Cortinas, Á., Abbas, A., Michalski, A., Moinet, A., Karlapati, S., Muszyńska, E., Guo, H., Putrycz, B., Gambino, S.L., Yoo, K., Sokolova, E., Drugman, T., 2024. BASE TTS: Lessons from building a billion-parameter Text-to-Speech model on 100K hours of data. URL https://arxiv.org/abs/2402.08093.

Lăpușteanu, A., Morar, A., Moldoveanu, A., Băluțoiu, M.-A., and, F.M., 2024. A review of sonification solutions in assistive systems for visually impaired people. Disabil. Rehabil.: Assist. Technol. 19 (8), 2818–2833. http://dx.doi.org/10.1080/17483107.2024.2326590, PMID: 38469665.

Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., Williamson, M., Manohar, V., Adi, Y., Mahadeokar, J., Hsu, W.-N., 2023. Voicebox: text-guided multilingual universal speech generation at scale. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA, pp. 14005–14034.

Linguistic Data Consortium, 2023. Linguistic data consortium. URL https://www.ldc.upenn.edu/.

Listen2it, 2024. Listen2it tamil text-to-speech. URL https://www.getlisten2it.com/text-to-speech-voices/tamil (Accessed 8 October 2024).

Madhavaraj, A., Bharathi Pilar, G., R.A., 2022. Knowledge-driven Subword Grammar Modeling for Automatic Speech Recognition in Tamil and Kannada. URL https://arxiv.org/abs/2207.13333.

Micmonster, 2024. Micmonster tamil voices. URL https://micmonster.com/ (Accessed 7 October 2024).

Ministry of Electronics and Information Technology, India, 2024. ULCA model search. URL https://bhashini.gov.in/ulca/search-model/66e41eeae2f5842563c988d8, (Accessed 15 October 2024).

Murthy, H.A., Umesh, S., 2023. FastSpeech2 model. URL https://creativecommons.org/licenses/by/4.0/, Copyright 2023, All rights reserved. Licensed under CC BY 4.0.

Narakeet, 2024. Narakeet text-to-speech. URL https://www.narakeet.com/ (Accessed 16 October 2024).

Open Speech EkStep, 2021. Text to speech model training - vakyansh. URL https://github.com/Open-Speech-EkStep/vakyansh-models#tts-models (Accessed 18 September 2024).

OpenAI, 2024a. Openai text-to-speech documentation. URL https://platform.openai.com/docs/guides/text-to-speech (Accessed: 11 October 2024).

OpenAI, 2024b. OpenAI TTS. URL https://platform.openai.com/playground/tts (Accessed: 15 November 2024).

OpenSLR, 2023. Open speech and language resources. URL https://www.openslr.org/.

Panda, S.P., Nayak, A.K., Rai, S.C., 2020. A survey on speech synthesis techniques in Indian languages. Multimedia Syst. 26 (4), 453–478. http://dx.doi.org/10.1007/s00530-020-00659-4.

Paramahansa Yogananda, 2024. Oru Yogiyin Suyasarithai - audio edition. URL https://yssofindia.org/ta/paramahansa-yogananda/paramahansa-yogananda-autobiography-of-a-yogi-mp3-edition (Accessed 18 September 2024).

Play.ht, 2024. Play.ht Tamil text-to-speech. URL https://play.ht/text-to-speech/indian-tamil/ (Accessed: 10 October 2024).

Prakash, A., Murthy, H.A., 2020. Generic indic text-to-speech synthesisers with rapid adaptation in an end-to-end framework. In: Interspeech 2020. In: interspeech_2020, ISCA, pp. 2962–2966. http://dx.doi.org/10.21437/interspeech.2020-2663.

Prakash, A., Reddy, M.R., Nagarajan, T., Murthy, H.A., 2014. An approach to building language-independent text-to-speech synthesis for Indian languages. In: 2014 Twentieth National Conference on Communications. NCC, pp. 1–5. http://dx.doi.org/10.1109/NCC.2014.6811356.

Prakash, A., Umesh, S., Murthy, H.A., 2023. Towards developing state-of-the-art TTS synthesisers for 13 Indian languages with signal processing aided alignments. In: 2023 IEEE Automatic Speech Recognition and Understanding Workshop. ASRU, pp. 1–8. http://dx.doi.org/10.1109/ASRU57964.2023.10389630.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., Auli, M., 2024. Scaling speech technology to 1,000+ languages. J. Mach. Learn. Res. 25 (97), 1–52, URL http://jmlr.org/papers/v25/23-1318.html.

Prathibha, P., Ramakrishnan, A., Muralishankar, R., 2002. Thirukkural II - A text-to-speech synthesis system. In: Tamil Internet 2002. pp. 126–133, URL https://www.infitt.org/ti2002/papers/27PRATIB.PDF.

Pubadi, D., Basandrai, A., Mashat, A., Chiurunga, Z., Gandhi, I., Ofei, W., Navaladi, L., Ramachandran, R., Ogunshile, E., 2020. A focus on codemixing and codeswitching in Tamil speech to text. In: 2020 8th International Conference in Software Engineering Research and Innovation. CONISOFT, pp. 154–165. http://dx.doi.org/10.1109/CONISOFT50191.2020.00031.

Rajendran, V., Kumar, G.B., 2019. A robust syllable centric pronunciation model for tamil text-to-speech synthesizer. IETE J. Res. 65 (5), 601–612. http://dx.doi.org/10.1080/03772063.2018.1452642.

Ramakrishnan, A., Kaushik, L.N., Narayana, M.L., 2007. Natural language processing for Tamil TTS. In: Proc. 3rd Language and Technology Conference, Poznan, Poland. pp. 192–196.

Ramakrishnan, A., Vikram, L., Abhinava, S.H., 2010. Bilingual TTS for Tamil and English. In: Tamil Internet 2010. pp. 303–305. http://dx.doi.org/10.13140/RG.2.1.1156.9762.

Renganathan, V., 2008. Computational phonology and the development of text-to-speech application for Tamil. In: INFITT Conference Proceeding. pp. 194–198, URL http://www.uttamam.org/papers/14_35.pdf.

ResponsiveVoice, 2024. ResponsiveVoice text-to-speech. URL https://responsivevoice.org/text-to-speech-languages/ (Accessed: 18 November 2024).

Ronanki, S., Reddy, S., Bollepalli, B., King, S., 2016. DNN-based Speech Synthesis for Indian Languages from ASCII text. URL https://arxiv.org/abs/1608.05374.

Sailor, H.B., 2013. Objective evaluation of speech quality of text-to-speech (TTS) synthesis systems (Ph.D. thesis). Dhirubhai Ambani Institute of Information and Communication Technology.

Salesky, E., Mäder, J., Klinger, S., 2021. Assessing Evaluation Metrics for Speech-to-Speech Translation. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop. ASRU, pp. 733–740. http://dx.doi.org/10.1109/ASRU51503.2021.9688073.

Samuel Manoharan, J., 2022. A novel text-to-speech synthesis system using syllable-based HMM for Tamil language. In: Shakya, S., Du, K.-L., Haoxiang, W. (Eds.), Proceedings of Second International Conference on Sustainable Expert Systems. Springer Nature Singapore, Singapore, pp. 305–314.

Shahid, K.S., Habib, T., Mumtaz, B., Adeeba, F., ul Haq, E., 2016. Subjective testing of urdu text-to-speech (TTS) system. Lang. Technol. 65, URL https://cle.org.pk/clt16/CLTProceedings.pdf#page=73.

Sharma, A., Srivastava, A., Vashishth, A., 2014. An assistive reading system for visually impaired using OCR and TTS. Int. J. Comput. Appl. 95 (2), 13–18. http://dx.doi.org/10.5120/16566-6231.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: Uses in assessing rater reliability. Psychol. Bull. 86 (2), 420–428. http://dx.doi.org/10.1037/0033-2909.86.2.420.

Speakatoo, 2024. Speakatoo tamil text-to-speech. URL https://www.speakatoo.com/tts/view/oAMXg2Jkl25566888ccfae66288d93ce389395271Y3N2ayPBl. (Accessed 28 November 2024).

Speechify, 2024. Speechify tamil text-to-speech. URL https://app.speechify.com/ (Accessed 22 November 2024).

Srinivasan, A., Rao, K.S., Kannan, K., Narasimhan, D., 2010. Speech Recognition of the letter 'zha' in Tamil Language using HMM. CoRR abs/1001.4190 URL http://arxiv.org/abs/1001.4190.

Suseendirarajah, S., 1993. Phonology and Morphology of Jaffna Tamil. University of Jaffna Publication.

TTSFree, 2024. TTSFree Tamil text-to-speech. URL https://ttsfree.com/text-to-speech/tamil-india (Accessed: 25 November 2024).

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., 2016. WaveNet: A Generative Model for Raw Audio. URL https://arxiv.org/abs/1609.03499.

Veed.io, 2024. Veed.io tamil text-to-speech. URL https://www.veed.io/ (Accessed 10 November 2024).

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., Saurous, R.A., 2017. Tacotron: Towards End-to-End Speech Synthesis. URL https://arxiv.org/abs/1703.10135.

Zeidan, A., 2020. Languages by total number of speakers. URL https://www.britannica.com/topic/languages-by-total-number-of-speakers-2228881 (Accessed 29 December 2024).

Zvelebil, K.V., 2021. Companion Studies to the History of Tamil Literature, vol. 5. Brill.