Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 47 No. 05
May 2026

Next Wave of Neural TTS: A review of Efficiency, Zero-Shot adaptation, and Expressiveness

Yuvraj Sinha[1], Dr Sandeep Kumar[2]

[1, 2] Department of Computer Science & Engineering, Sharda University, Greater Noida, Uttar Pradesh, India

Email Id: [1] yuvrajsinha150@gmail.com, [2] Sandeep.csengg@gmail.com

Abstract

Neural Text-to-Speech (TTS) synthesis has become remarkably natural, making the research frontier transition to specialized and real-world use. A decade of (2025) recent contributions are intersumed in this review to define key trends and serious gaps in research. We examine developments in three main dimensions: (1) Efficiency and Accessibility, (2) Data Efficiency and Adaptation, (3) Expressiveness and Robustness, in the context of emotion classification, linguistic sensitivity in low-resource languages, and security watermarking. Our synthesis indicates that there is a gap in research: individual models are doing great in a specific area (e.g., efficiency or zero-shot), but there are no unified frameworks, which are efficient (on device), data-scarce (zero-shot), and expressive models (prosody/emotion controlled).

Keywords: Neural Text-to-Speech, Data-Efficient Learning, Expressive Speech Synthesis, Zero Shot Adaptation, Speech Model Efficiency.

1. Introduction

In the today's smart and digital world, which is moving rapidly for daily hard and dynamic digital productions, we need a really quick and efficient relation or interaction between computer and humans. We have many technologies like NLP, Deep learning, Machine learning, Cloud computing etc. and in recent years of research, we came across many models of NLP (Natural Language Processing), in which one the technologies are Text to Speech (TTS systems). The significant advancements in text to speech technology has made the artificial voices that sounds more and more realistic now a days. India, a country which has a very diverse environment and it it boasting diverse block of languages and takes pride of its heritage. The most existing and used TTS model is developed for English and in India there is 22 languages and 1600 dialects which creates a really big problem. Formidable challenges like language barriers, mutual communication, hindering development, and constraining effective communication and impeding progress.

By taking 1000's of steps or walking across this nation will tell you that there is a really big differences between languages not just because of their pride in speaking other languages but also the lack of awareness to the new variables of different language. Hence we can also see that everyone in India is not familiar with one common language. We can find many people across the who interact via Indian languages

The 2025 literature, represented by the 10 reference papers, highlights several key challenges:

- Efficiency: How can high-fidelity models run on low-power, local devices (e.g., for accessibility) or in real-time streaming applications?
- Data Scarcity: How can models synthesize new voices with minimal (low-resource) or zero (zero-shot) audio samples from the target speaker?.
- Control and Expressiveness: How can synthesis be controlled to convey specific emotions, prosodies, or linguistic nuances for non-English languages?
- Robustness and Accessibility: How can TTS be secured against misuse (e.g., deepfakes) or adapted for new modalities, such as sign language?

This paper is a systematic review of these overlapping streams of research. We integrate the methods used by the reference materials, categorize them into themes and define a critical research gap that arises out of the synthesis. The traditional Text-to-speech models tend to encounter a difficulty when tuning the text and

speech. This minor miss cooperation gives rise to information gap, which the current methods tried to cover through gathering more inputs, but which usually results in inconsistency in data and causes more complications. Text to speech (TTS) systems plays and important role in human-computer synergy. Text to Speech (TTS) system remodels written text to articulated words using Computer systems. This enables machines to speak text uniformly. The materialization and success of speech potrayal models, have galvanized a shift in shift zero shot text to speech research towards crumbling the synthesis process. Zero shot TTS model is a structure that arranges speech in a target voice via small audio sample, even if the vice is new to the system or system has encountered the voice for the first time. Text to Speech Systems are not just helpful or actively used for this single domain, this technology is also used in image processing. Text recognition is a major area of the study within a domain of image processing and then after text recognition to speech conversion via text to speech. This is used in personal assistant which is used to classify or extract text out of an image

2. The Push for Efficiency

A primary driver of current research is the need to move TTS systems from high-latency cloud servers to local, low-power devices. This is critical for accessibility applications, privacy, and real time conversational AI.

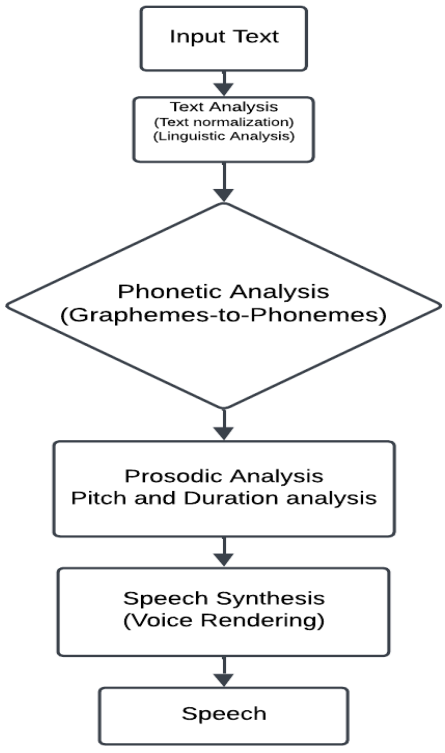

Fig. 1. Text To Speech Workflow

Compact Models for Accessibility

Traditional accessibility solutions involved a trade-off: low-latency but unnatural device-based synthesis, or high-quality but slow cloud-based neural TTS. "Compact Neural TTS Voices for Accessibility" addresses this by developing a high- quality compact neural TTS system. The goal is to achieve extremely low latency (on the order of 15ms) and a low disk footprint, making high-quality neural voices viable for pre-installation on handheld, low-power devices.

Lightweight Architectures

This compactness can only be achieved through some basic architectural innovation. LEF-TTS (Lightweight and Efficient End-to-End Text-to- Speech) suggests the methods to decrease the complexity of the calculations and the number of parameters. Key innovations include:

- Fast Linear Attention: Replacing standard stacked Transformers with Fast Linear Attention with a Single Head.
- Conv Wave-Net: Improving the network architecture to further decrease model parameters
- Multi-Stream Generator: Accelerating inference using a multi-stream inverse short- time Fourier Transform generator.

Low-Latency Streaming

In speech generating applications (conversational AI and simmers) overall speech generation time is of less concern than time-to-first-sound (low latency). This is rigged by Kolmogorov-Arnold Networks (KANs) in "Stream-TTS". KANs are awarded as a more efficient alternative to the conventional Multi-Layer Perceptrons (MLPs), which add-on smaller computational graphs and reduced parameters. This architecture has been designed in low latency simmers speech applications.

Synthesis: The trend is obvious: adequacy is being sought out through constructive (LEF-TTS), computational (Stream-TTS) and escalation (Compact TTS) approaches to effect local deployment.

3. The Challenge of Data: Low-Resources and Zero-Shot Adaptation

The second major research flag is data efficiency. Training huge TTS models requires big, high quality,

mono-speaker datasets, which are not inexpensive and impractical for exemplification. Low-resource and zero-shot erudition aim to solve this

Low-Resource Speaker Corpora

"CMDF-TTS" clearly targets scenarios to a "limited target speaker core". It proposes a method which integrates a confining module and a multi-storied prosody modeling framework to adequately utilize the limited data.

Speaker-Adaptive Prosody

One of the problems of adoption is not only the imitation of the accent of the speaker (vowel coloring) but also an excessively large part of his or her prosody (rhythm and accent). The idea of solicitation of the prosody is presented in "Stable-TTS". It is a framework that uses a small group of best-quality prior samples of the pre-training set to control the consistency of prosody, but achieves this goal by simultaneously modeling the accent of the target speaker.

Zero-Shot High Fidelity

Zero-Shot TTS hopes to recreate a voice with the help of only one audio sample. Evidential-TTS is aimed at high fidelity on this task. It points out the key weakness: It is overconfident, which is normally the case with Iterative Parallel Decoding (IPD). In order to address this, Evidential-TTS incorporates Evidential Deep Learning (EDL) to tune model uncertainty. This ambiguity provides a more promising course of path of instability, to more advanced naturalness.

Stable and Robust Zero-Shot Synthesis"

SSR-Speech" is also a speech sermonist, with their interest in stabilizing and robustness of the process. It uses the Transformer decoder mixed with classifier-free instruction to increase formation stability.

Synthesis: Data-efficacious TTS is no longer looking smooth in voice cloning. The tie has been on bound synthesis in both stable (SSR-Speech), high-ardor (Evidential-TTS) synthesis that will be able to reproduce not only timbre but also prosody on the nominal data (Stable-TTS, CMDF TTS)

4. Expanding The Frontier

The last research push involves the extension of the capabilities of TTS to the new areas, new languages and new obstacles, including security and poly-modality

Expressive and Emotional TTS

Standard TTS pronounces text in an informative, story-telling manner. Expressive TTS (ETTS) is set to bring speech emotion. Deep Learning assisted Bangla text normalization: This work adds an important preprocessing step in ETTS. Two main contributions that the paper can make are:

1. Bangla Text Normalization (BTN): A system to encode the non-standard form of a word (e.g., a number or date) into a standard spoken form with the help of a Temporal Convolutional Network (TCN).
2. Emotion Classification: An A Hierarchical Attention Network (HAN) model that uses normalized text to determine the desired emotion which may be inputted into an ETTS synthesizer.

This points out that expressiveness is a process that starts with profound linguistic knowledge prior to synthesis.

Low-Resource Language Difficulties

In Linguistically complex by being low-resource languages, the challenges of text normalization are increased in number. This problem space is reviewed in "A survey and evaluation of text-to speech systems to the Tamil language" article. It points out that the rich linguistic characteristics, as well as nature of Tamil, pose a great challenge to speech synthesis. The present paper supports the results of the Bangla paper, which is that strong TTS in many languages necessitates the solution of difficult problems in Natural Language Processing. It surveys the Tamil, which is a language that has a rich linguistic feature and has diagnostic nature and is in state of the art. The research has serious challenges. Its objective is to check these particular language barriers, and, with no less importance, to suggest a "perceptual evaluation framework" to measure the quality of future Tamil TTS systems correctly. In this paper, the definition of what constitutes a hard low-resource language is presented. A Language Model (LM) can be trained to learn this content (semantic tokens) and ignore the voice (speech vectors) by training on these two streams separately.

Security and Robustness

Zero-shot synthesis is a security risk: deepfake audio. This is the point that is directly focused by the proposal of a watermark Encodec in SSR-Speech. Frame-level

tidemark is embedded in the system into areas of the speech under editing, which allows identifying that the elements of audio clips were synthesized.

Cross-Modal Accessibility

At last, now, Kannada Speech to Sign gesture Language Converter is blowing the synthesis out of audio altogether. This presents a new platform that combines speech-to-text system and Indian Sign Language (ISL) origination, presumably via a 3D avatar. The application of synthesis to a visual, cross-modal dimension posits to a derogatory subset of the article of 2025 Cross-Modal Synthesis, specifically, transformation of Audio/Text into optical Sign Language.

5. Synthesis and Visualized Analysis

We can see this formed TTS conduit and relative objective of the contributions.

The flowchart diagram illustrates the pipeline of a complicated and full-length pipeline of state of-the-art Neural Text-to-Speech (TTS) systems, integrates key innovations of 2025 research panorama into one system. It is initiated by Leading Text Normalization & Emotion/dialectal Analysis, a base division that is concerned with the deep understanding of linguistics as opposed to the banal phonetization. The latter is highlighted by [2] (Bangla) and [3] (Tamil), rough text requires appropriate pre-processing solving atypical words, in dates or currencies as well as separating emotional context, a step which is particularly humiliating the task of handling the complicated phonology of low-resource languages. The second step is that the text under handling is inputted into the dialectical/metrics Encoder where cadence, rhythm, stress and accents are determined. [9]. This case starts with a statement of a phrase, the prosody prompting, and which gives the model the mandate to twin the style presentation to the reference sample, such that the product is readable and not just to be read aloud. This is fed directly to Speaker/Style alteration Module where the vocal identity is written. This literature is an important transition to the efficiency of this stage in data: Paper 1 proposes the use of adaptation with finite training corpora (low-resource), but in the meantime [5] and [8] apply the zero-shot method to produce a marked voice with the help of a short phonic prompt. The essence is observed in the Acoustic Generator/Critic whereby the linguistic features are converted into acoustic descriptions (mel-spectrograms). These stages can be defined as efficiency propulsion to support real time use and in [10] proposes Kolmogorov- Arnold Networks (KANs) to reduce the latency in order to render the system realistic in regards to live streaming. Finally, now signal is achieved in Vocoder/Post- Rarefaction stage. In [4], it is specific to ensure that vocoder is small enough to be implemented on-device, and [8] offers a security buffer, which puts apex-invisible there, which would be difficult to abuse. In the negative, the pipeline ends by following the multi-modal products, shown in [6], these systems can completely ignore the audio to drive sign dialects avatars, which brighten the synthesis range which would include the visual accessibility devices. This is fed directly to Speaker/Style alteration Module

Table I: Review table

Author(s) (Lead)	Paper Title / Topic	Primary Output	Technology / Method Used
Ye Tao, et al.	CMDF-TTS: TTS with limited target speaker corpus	A TTS method for low-resource speaker adaptation	Multi-level prosody modeling 29, corpus compression (SCAC)
Md. Rezaul Islam, et al.	Bangla Text Normalization & Emotion Classification for ETTS	A preprocessing pipeline for expressive Bangla TTS; new corpora (BTN, BNET)	Deep Learning: Temporal Convolutional Network (TCN), Hierarchical Attention Network (HAN)
A. Mahaganapathy, et al.	Survey & Evaluation of Tamil TTS	A comprehensive review and a new perceptual evaluation framework	(Survey) Reviews various methods; proposes new evaluation metrics
Kunal Jain, et al.	Compact Neural TTS for Accessibility	A high-quality, low latency (15ms) on device neural TTS system	Neural TTS model optimization for low power, low- footprint deployment.

Author(s) (Lead)	Paper Title / Topic	Primary Output	Technology / Method Used
Myeonghun Jeong, et al	Evidential-TTS: High-fidelity zero shot TTS	A zero-shot TTS model with improved naturalness.	Evidential Deep Learning (EDL) to quantify model uncertainty; Iterative Parallel Decoding (IPD)
Pushpalatha M. N, et al	Kannada Speech to Sign Language Converter	A platform for cross modal accessibility (Speech-to-ISL)	Speech-to-Text Technology, Indian Sign Language (ISL) generation (3D Avatar)
Yan Shi, et al.	LEF-TTS: Lightweight & Efficient E2E TTS	A fast, lightweight E2E TTS model.	Fast Linear Attention (replaces Transformer), Conv WaveNet, Multi Stream Generator
Helin Wang, et al.	SSR-Speech: Stable, Safe, Robust zero-shot speech editing/synthesis	A robust zero-shot TTS model with security features	Neural codec autoregressive model, Transformer decoder, "watermark Encodec"
Wooseok Han, et al.	Stable-TTS: Stable speaker- adaptive TTS via prosody prompting.	A speaker-adaptive framework that resists overfitting.	"Prosody Prompting" (using "prior samples"), Prior- Preservation Loss.
Giridhar Pamisetty, et al	Stream-TTS: Low latency TTS for streaming.	A low-latency (RTF 0.0795) multilingual streaming TTS model.	Kolmogorov-Arnold Networks (KANs) as MLP replacement, Supervised Auxiliary Learning.

6. Conclusion

This review has captured 10 2025 papers mapping the frontier in TTS studies. Our analysis of strong trends revealed three: the pressure towards on-device efficiency, the art of data-efficient adaptation, and the increase in expressive, multilingual, and secure synthesis. The gap to be addressed by the research is the absence of integration of these three areas. The future work should aim at coming up with single frameworks that are compact, adaptive and at the same time expressive. This will probably need new architectures, multi-task learning models, and new knowledge distillation strategies to reduce large, expressive, zero-shot models to lightweight and efficient architectures without compromising quality. It is the answer to this integration problem that makes Neural TTS, present, personal, and safe for purpose of communication

References

[1] Han, W., Kang, M., Kim, C., & Yang, E. (2025). *Stable- TTS: Stable Speaker-Adaptive Text- to-Speech Synthesis via Prosody Prompting*. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[2] Islam, M. R., & Rahman, M. S. (2025). Deep learning-based Bangla text normalization with emotion classification for expressive text-to-speech synthesizer. *Applied Soft Computing Journal, 184*, 112899.

[3] Jain, K., Murphy, E., Gupta, D., Dyke,J., Shah, S., Tsiaras, V., Petkov, P., & Conkie, A. (2025). *Compact Neural TTS Voices for Accessibility*. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[4] Jeong, M., Kim, M., Kim, S., & Kim, N. S. (2025). *Evidential- TTS: High Fidelity Zero-Shot Text-to-Speech Using Evidential Deep Learning*. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[5] Mahaganapathy, A., & Sarveswaran, K. (2025). A survey and evaluation of text-to-speech systems for the Tamil language *Natural Language Processing Journal, 12*, 100171.

[6] Pamisetty, G., Easow, R. A., Gupta, K., & Murty, K. S. R. (2025). *Stream-TTS: A Low- Latency Text-to-Speech using Kolmogorov-Arnold Networks for Streaming Speech Applications*. In *ICASSP 2025-2025 IEEE International (ICASSP)*.

[7] Pushpalatha M. N., Parkavi A., Koley, A., & Naik, A. M. (2025). *Kannada Speech to Sign Language Converter: A Novel Platform Integrates Indian*

Sign Language and Speech-to-Text Technology. In *2025 International Conference on Microwave, Optical, and Communication Engineering (ICMOCE)*.

[8] Shi, Y., Shi, J., Chen, M., Miao, C., Fang, M., Cheng, N., Wang, S., & Xiao, J. (2025). *LEF-TTS: Lightweight and Efficient End-to-End Text-to-Speech Synthesis with Multi-Stream Generator*. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[9] Tao, Y., Liu, J., Lu, C., Liu, M., Qin, X.,Tian, Y., & Du, Y. (2025). CMDF-TTS: Text-to-speech method with limited target speaker corpus. *Neural Networks, 188*, 107432.

[10] Wang, H., Yu, M., Hai, J., Chen, C., Hu,Y., Chen, R., Dehak, N., & Yu, D. (2025). *SSR-Speech: Towards Stable, Safe and Robust Zero-shot Text-based Speech Editing and Synthesis*. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[11] P. Patil, V. Phaltankar, P. Patil, A. Bombarde, and C. Lande, "Text Echo Personalized TTS System," in 2025 Global Conference in Emerging Technology (GINOTECH), Pune, India, 2025.

[12] S. Priyadharshini and M. R. Nachiyar, "A Study on Intelligibility, Naturalness and Fluency of Text-to-Speech Systems," in 2025 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2025.

[13] C. Du et al., "VALL-T: Decoder-Only Generative Transducer for Robust and Decoding-Controllable Text-to- Speech," in ICASSP 2025, 2025.

[14] X. Zhu et al., "Vec-Tok Speech: Speech Vectorization and Tokenization for Neural Speech Generation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 33, 2025.

[15] J. Lee, N.-S. Song, and J.-H. Chang, "Vector Field Decomposition-Based Flow Matching for Zero-Shot Cross-Lingual Text-to-Speech," IEEE Signal Processing Letters, vol. 32, 2025.

[16] J. Yeom et al., "VoiceGuider: Enhancing Out-of-Domain Performance in Parameter-Efficient Speaker-Adaptive Text-to-Speech via Autoguidance," in ICASSP 2025, 2025.

[17] C. Wang et al., "DetailTTS: Learning Residual Detail Information for Zero-shot Text-to-speech," in ICASSP 2025, 2025.

[18] J. Xing et al., "DecoupledSynth: Enhancing Zero-Shot Text-to-Speech Via Factors Decoupling," in ICASSP 2025, 2025.

[19] H. Zhang et al., "A Text-to-Speech Synthesis Method Based on Speaker Transfer," in 2024 IEEE 24th International Conference on Communication Technology (ICCT), 2024.