Dhaka Univ. J. Sci. 74(1): 10-16, 2026 (January)
Prof. M. Muhibur Rahman Memorial Issue

ISSN 1022 - 2502, E-ISSN 2408- 8528
DOI: https://doi.org/10.3329/dujs.v74i1.84122

Bangla TTS Performance Evaluation: A Benchmark Study on Synthesized Speech Quality and Intelligibility

Mehadi Hasan, Dipto Shaha, and Md. Rezaul Karim*

Department of Computer Science and Engineering, Faculty of Engineering and Technology
University of Dhaka, Dhaka, Bangladesh

(Received : 2 September 2025; Accepted : 1 January 2026)

Abstract

Bangla Text-to-Speech (TTS) systems have seen significant advancements in recent years, yet comprehensive benchmarking of their performance remains limited. This study establishes a robust evaluation framework to compare different Bangla TTS models, including Tacotron2[1], FastSpeech2[2], VITS[3], and Grad-TTS[4]. The benchmarking approach integrates both objective and subjective assessment methodologies. Objective evaluation employs signal processing metrics such as Mel Cepstral Distortion (MCD), Mel-Spectrogram Mean Squared Error (Mel-MSE), Phoneme Error Rate (PER), Word Error Rate (WER), Signal-to-Noise Ratio (SNR), and Real-Time Factor (RTF). Subjective evaluation involves human perceptual tests such as Mean Opinion Score (MOS) test with native Bangla speakers rating speech quality and intelligibility. The study's experimental setup ensures a fair comparison by utilizing a standardized dataset, uniform computational conditions, and diverse sentence structures. Results demonstrate the relative strengths and weaknesses of various models, highlighting the need for improved phonetic accuracy and naturalness in Bangla TTS synthesis. This research provides critical insights for advancing Bangla TTS systems and aligning them with state-of-the-art English TTS models.

Keywords: Text-to-Speech (TTS), speech synthesis, benchmarking, objective evaluation metrics, subjective evaluation metrics.

I. Introduction

The emergence of Text-to-Speech (TTS) technology has transformed human- computer interaction by enabling machines to generate intelligible and natural-sounding speech from written text. Such systems are particularly instrumental in enhancing accessibility for individuals with visual impairments, reading difficulties, or low literacy, while also serving broader functions in education, customer service, language learning, and digital accessibility. Early TTS systems, based on template-driven approaches, often produce monotonous and robotic outputs. With the advancement of deep learning, particularly sequence-to-sequence and non-autoregressive neural architectures, contemporary TTS systems have significantly enhanced their ability to reproduce human-like intonation, expressiveness, and speech clarity.

Despite considerable progress in speech syn-thesis for languages such as English, Bangla TTS development remains constrained by linguistic intricacies, limited annotated datasets, and the absence of standardized benchmarking protocols. Given that over 26% of Bangladesh's population is illiterate according to the Household Income and Expenditure Survey 2022 a robust Bangla TTS system has the potential to bridge significant accessibility gaps. Additionally, visually impaired individuals, estimated in the millions globally by the World Health Organization, stand to benefit immensely from well-designed Bangla TTS systems that convert digital text into accessible audio content. Such systems also support language acquisition and literacy improvement in rural and underserved communities, where access to quality education remains limited.

To fully leverage the advantages of TTS technology, rigorous benchmarking is essential for evaluating the naturalness, intelligibility, and computational efficiency of synthesized speech. While objective metrics such as Mel Cepstral Distortion (MCD), Word Error Rate (WER), Phoneme Error Rate (PER), and Real-Time Factor (RTF) are crucial, subjective human evaluations using Mean Opinion Score (MOS) offer insights into the perceptual quality of speech output. The scarcity of comprehensive benchmarking frameworks for Bangla has led to inconsistent evaluation practices across studies, hampering the reliability and comparability of different models.

In addressing these challenges, this work proposes a complete benchmarking framework for Bangla TTS systems. The contributions include the development of a standardized evaluation system combining objective and subjective metrics, the curation of a refined dataset featuring high-quality

recorded audio, and the implementation of a web platform with public API access. Through comparative analysis of several prominent TTS architectures including Tacotron2[1], FastSpeech2[2], VITS[3], and Grad-TTS[4], the study examines trade-offs in quality, inference speed, and usability.

However, the path toward high-performing Bangla TTS systems is fraught with challenges, including phonetic complexity, prosody modeling, real-time inference requirements, and generalization across diverse recording

*Author for correspondence. e-mail: rkarim@du.ac.bd

conditions. Issues such as ethical concerns in voice cloning, lack of dataset uniformity, and scalability in low-resource environments also persist. Despite these limitations, the proposed work aims to bridge the gap between Bangla and globally benchmarked TTS technologies. It provides a foundation for systematic evaluation and improvement of Bangla speech synthesis systems, advancing accessibility and digital inclusion across linguistic, social, and technological dimensions.

II. Background and Related Works

Text-to-Speech (TTS) systems convert written text into audible speech using neural models. The modern TTS pipeline consists of a text analysis module, an acoustic model that generates mel-spectrograms, and a vocoder that converts spectrograms into audio. Initially, feedforward neural networks were used, but due to their limitations in modeling temporal dependencies, recurrent models such as RNNs and LSTMs became popular. Subsequently, Transformers introduced attention[16] mechanisms to better handle long-term dependencies in sequences, significantly improving TTS performance.

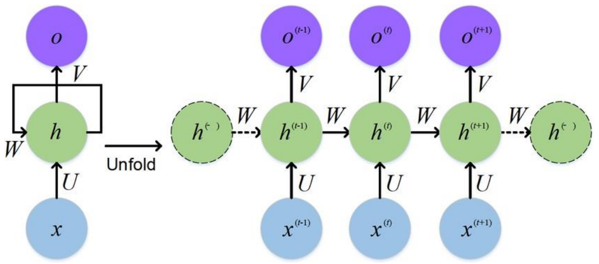

Fig.1. Structure of a Recurrent Neural Network

Basically, the recurrent neural network (RNN) shown in Figure 1 introduces memory by retaining hidden states across time steps. The hidden state at time t is computed based on the input at time t and the hidden state from the previous time step:

$h^{(t)} = f\,(Ux^{(t)} + Wh^{(t-1)})$ The forward pass can be expanded as:

$$a^{(t)} = b + Wh^{(t-1)} + Ux^{(t)}$$

$$h^{(t)} = \tanh(a^{(t)})$$

$$o^{(t)} = c + V\,h^{(t)}$$

$$y^{(t)} = \mathrm{softmax}(o^{(t)})$$

Here, $a^{(t)}$ and $h^{(t)}$ denote activations and hidden states, $o^{(t)}$ is the output before activation, and $y^{(t)}$ represents the final prediction at each time step.

Transformer-based and diffusion-based models have become foundational in modern TTS systems, enabling high-quality, efficient synthesis. Recent benchmarking platforms such as TTS Arena by Hugging Face[12], Picovoice's Latency Benchmark[13], OpenVoice Benchmark [14], and the TTS Benchmark 2025[15] study have evaluated commercial and open-source systems like ElevenLabs, Smallest.ai, and OpenVoice, analyzing trade-offs in expressiveness, intelligibility, and real-time performance. These efforts utilize both objective metrics (MCD, SNR, RTF) and subjective metrics (MOS) to measure synthesis quality. Our own benchmarking on Tacotron2[1], FastSpeech2[2], VITS[3], and Grad-TTS[4] follows a similar evaluation strategy, providing comparative analysis of autoregressive, non-autoregressive, and diffusion-based models to assess the balance between quality and latency.

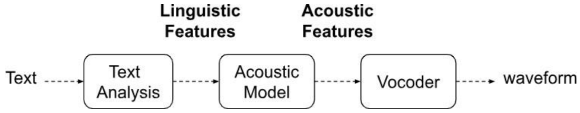

Fig. 2. Workflow of a TTS Model

Significant progress has been made in non- native and low-resource TTS, particularly through transfer learning and fine-grained speaker adaptation. Studies employing t-SNE on d-vectors have demonstrated strong speaker identity preservation across models like Grad-TTS[4] and FastSpeech2[2], especially in data-scarce settings. Research on Kurdish, Kazakh[20] and Bangla TTS has shown that using native phoneme corpora and refining datasets greatly enhances synthesis quality. Our current work builds on these findings by applying Tacotron2[1], FastSpeech2[2], VITS[3], and Grad-TTS[4] to Bangla TTS using a cleaned and augmented dataset, while incorporating both subjective (MOS) and objective (MCD, RTF) metrics to ensure a robust evaluation. The integration of deep learning-based vocoders like HiFi-GAN[17] further boosts audio fidelity, making our approach suitable for real-time and scalable deployment in low- resource language contexts.

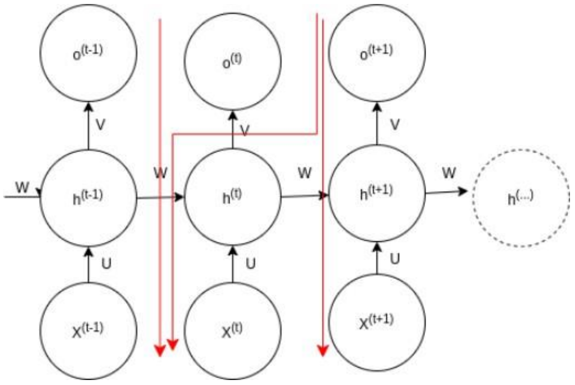

Fig. 3. Gradient Flow in Sequential Models

In Bangla TTS development, most prior efforts focused on rule-based phoneme conversion, basic LSTM architecture, and crowd-sourced datasets. A recent state-of-the-art approach introduced a diffusion-based Bangla TTS system combining Grad-TTS[4] and HiFi-GAN[17] with a Stochastic Duration Predictor, supported by an 18.43-hour single-speaker dataset. This work addressed duration modeling and expressive speech generation, offering a foundation for standardized Bangla TTS benchmarking.

Despite growing interest, the lack of robust benchmarking frameworks and high-quality annotated datasets continues to hinder fair evaluation in Bangla TTS research. This study aims to address these gaps by introducing an integrated benchmarking framework for Bangla TTS systems using refined datasets, objective signal processing metrics, and subjective perceptual evaluations.

III. TTS Dataset Evaluation and Contribution

High-quality datasets are foundational for training robust Text-to-Speech (TTS) systems. An ideal TTS dataset should exhibit the following characteristics: balanced clip and text lengths, noise-free and well-aligned recordings, consistent tone and pitch, comprehensive phonetic coverage, fluent and natural narration, uniform recording environments, and sufficient data volume. Additionally, diversity in speaker demographics (age, gender, accent) and proper segmentation of pauses contribute to better prosody modeling and broader generalization. Datasets should be provided in high- fidelity formats such as 16-bit mono WAV at 48kHz to preserve audio integrity.

Several English datasets have established benchmarks for TTS research. LJSpeech[5] offers over 13,000 clips from a single speaker, while LibriSpeech[6] and TED-LIUM 3[7] provide large-scale multi-speaker corpora derived from audiobooks and TED Talks, respectively. VCTK Corpus[8] includes speech from 110 speakers with varied accents, and Common Voice[9] offers extensive speaker diversity but suffers from inconsistent recording quality.

For Bangla, available resources are limited. The Google Bangla dataset[21] offers high- quality speech samples from native speakers, suitable for recognition and synthesis tasks. A more notable contribution is the 18.43- hour single-speaker dataset developed by Intesum and Masud[10], designed specifically for Bangla TTS training. Originally 27.5 hours, it was filtered to remove noise from page turns, breath sounds, and environmental interferences. Enhanced practices such as reading from digital screens and using pop filters or cardioid microphones were recommended to improve audio quality. Post-processing tools like Audacity and iZotope RX7 were used to refine recordings.

Despite progress, existing Bangla datasets face several limitations: limited speaker and dialect diversity, insufficient phonetic coverage, inconsistent audio quality, lack of contextual variation, and restricted public availability. These gaps hinder comprehensive benchmarking and reduce the scalability and adaptability of Bangla TTS models. Addressing these challenges through enriched, publicly accessible datasets is essential for advancing speech synthesis in underrepresented languages.

The dataset from Mushahid Intesum and Abdullah Ibne Masud's project titled *A Ro- bust Text-to-Speech System in Bangla with Stochastic Duration Predictor*[10] was analyzed for quality. We observed that while it was a valuable resource, several clips contained noise and unclear pronunciation. To facilitate refinement, we categorized the data into three quality levels as shown in Table 1.

Table 1. Dataset Categorization Based on Audio Quality

Category	Description
C1: Excellent	Clean audio with clear pronunciation and no noise.
C2: Moderate	Noise-free but includes unclear or weak pronunciation.
C3: Very Poor	Contains background noise and unclear articulation.

We manually refined the samples from Category-2 and Category-3 using high-quality recording tools and noise reduction techniques. The resulting dataset features clear pronunciation and minimal background noise, leading to improved model performance in our experiments.

IV. Proposed Methodologies

This section outlines our approaches for dataset creation, refinement, and speech synthesis.

Data Collection

To develop a Bangla TTS system, we first construct a rich text corpus that reflects the language's linguistic and phonetic diversity. Texts are collected from various domains to ensure full phoneme and intonation cover- age. Fluent native Bangla speakers with neutral accents are selected. Recordings are made in a professional studio with high-quality equipment. Scripts contain correct pronunciation and are read with multiple takes to ensure clarity and accuracy.

Data Post-processing

To eliminate background noise, we use several software tools and models:

Audacity: Uses Fourier analysis to identify and reduce background noise. Applies FFT-based gain control while preserving speech quality.

Krisp: Machine learning-based tool that re- moves background noise in real-time. Works with 800+ apps and processes audio locally to protect privacy.

iZotope RX7: Includes features like de- reverb, de-rustle, breath control, dialogue isolate, and de-bleed to enhance clarity.

Accusonus ERA Bundle: Removes ambient noise (e.g., fans, AC), adjusts gain inconsistencies, and softens harsh consonants using tools like reverb remover, voice leveler, and de-esser.

Additional Step: A two-stage CNN-based U-Net model by Moliner et al.[11] can be fine-tuned to eliminate residual noise such as hiss, clicks, and thumps, improving audio quality.

Methodology for Benchmarking Bangla TTS Models

We benchmark Bangla TTS models using two complementary approaches: objective metrics based on signal processing and subjective human evaluation via Mean Opinion Scores (MOS). This evaluates intelligibility, naturalness, and computational efficiency. Models are tested on the same diverse Bangla dataset with native speakers providing subjective feedback under consistent computational conditions.

Mathematical Comparison (Objective Metrics)

Objective metrics quantify acoustic, phonetic, and computational aspects of synthesized speech compared to natural speech.

Spectral Distance Measures: Assess spectral similarity between synthesized and reference speech; lower values imply better quality.

Mel Cepstral Distortion (MCD): Mel Cepstral Distortion is an objective metric for evaluating the difference between two sequences of mel-cepstral coefficients, typically from a reference and a synthesized signal. It quantifies the spectral distance between the original and synthesized speech signals in the mel-cepstral domain.

$$MCD = \frac{10}{\log(10)}\sqrt{2\sum_{d=1}^{D}(c_d^{ref} - c_d^{syn})^2} \qquad (1)$$

Lower MCD indicates higher similarity between the reference and synthesized audio, implying better synthesis quality.

Mel-Spectrogram Mean Squared Error (Mel-MSE): Mel-Spectrogram Mean Squared Error is an objective metric used to evaluate the average squared difference between the mel-spectrograms of reference and synthesized speech signals. It measures the discrepancy in energy distribution over time and frequency in the log mel-spectrogram representation.

$$Mel - MSE = \frac{1}{N}\sum_{i=1}^{N}(S_i^{ref} - S_i^{syn})^2 \qquad (2)$$

Mel-MSE is better when its value is lower, indicating a closer match to the reference spectrogram.

Phoneme Error Rate (PER): Phoneme Error Rate is an objective measure of the phonetic accuracy of synthesized speech, based on the comparison of predicted and reference phoneme sequences. It calculates the proportion of phoneme errors, including substitutions, deletions, and insertions, relative to the total number of reference phonemes.

$$PER = \frac{S+D+I}{N} \qquad (3)$$

Lower PER reflects better phonetic accuracy and intelligibility.

Word Error Rate (WER): Word Error Rate is a metric for evaluating the intelligibility of synthesized speech by comparing the word sequence of the output against the reference transcription. It measures the word-level alignment errors, including substitutions, deletions, and insertions.

$$WER = \frac{S+D+I}{N} \qquad (4)$$

Lower WER means clearer and more accurate speech synthesis.

Signal-to-Noise Ratio (SNR): Signal-to-Noise Ratio is a measure of the relative strength of the desired signal to the background noise present in the synthesized audio. It assesses how clean or noisy the synthesized signal is by comparing the power of the clean signal to the noise (error).

$$SNR = 10log_{10}\frac{\sum_i s_i^2}{\sum_i(s_i - \hat{s}_i)^2} \qquad (5)$$

Higher SNR indicates cleaner output, with less distortion or noise.

Real Time Factor (RTF): Real-Time Factor (RTF) is a metric that evaluates the computational efficiency of a speech synthesis system by comparing the time taken to generate audio with the actual duration of the audio. It measures how fast the model can synthesize speech relative to the length of the output audio.

$$RTF = \frac{Synthesis\ Time}{Audio\ Time} \qquad (6)$$

An RTF close to 1 indicates real-time syn- thesis capability. Lower than 1 is faster than real-time; higher than 1 is slower.

Human Evaluation (Subjective Metrics)

Objective metrics lack full insight into perceived naturalness and clarity, so human evaluation is crucial. We have considered a form with eight audio and different synthesized audio of four different models and took feedback of human how the quality was.

Mean Opinion Score (MOS) Listeners rate samples on a 1–5 scale according to Table 2.

Table 2. Mean Opinion Score (MOS) rating scale.

Score	Description
5	Excellent, indistinguishable from human speech
4	Good, minor flaws but natural
3	Fair, some unnatural elements
2	Poor, robotic or difficult to understand
1	Very Bad, unintelligible or nothing is understandable

MOS is the average of native speakers' ratings over randomized samples.

V. Experimental Setup

To ensure a fair and systematic evaluation of different TTS models, we follow a well- defined experimental setup that includes dataset selection, model evaluation with human listeners, and a consistent computational environment.

Dataset Selection

A carefully curated test dataset is crucial for ensuring a diverse and representative evaluation. To achieve this, we employ a **clustering-based selection approach** combined with **TF-IDF vectorization** and **length-weighted sampling**. The steps are as follows:

1. Load text files and compute the word count for each sample.
2. Convert text to **TF-IDF feature vectors** using a **custom tokenizer** to retain Bangla words, numbers, and punctuation.
3. Apply **K-Means clustering** to group similar text samples into 50 clusters.
4. Use **length-weighted random sampling** to select one representative sample per cluster.
5. Copy the corresponding text and audio files to a new dataset folder for evaluation.

Model Evaluation and Listener Feedback

We evaluate four state-of-the-art TTS models: Tacotron2[1], FastSpeech2[2], VITS[3], and Grad-TTS[4]. For subjective evaluation, we recruit 50 native Bangla speakers to assess pronunciation accuracy, naturalness, and prosody. Their feedback ensures high linguistic reliability, given their familiarity with phonetic and prosodic nuances.

Computational Setup

All models were trained and tested under a consistent hardware configuration using NVIDIA GPUs with CUDA support, which enables efficient matrix operations and parallel processing—both critical for training deep neural TTS models. Training on CPUs was found to be inadequate due to performance bottlenecks.

The hardware used for all experiments is listed in Table 3, which fits within a single column.

Table 3. System configuration used for model training and evaluation.

Component	Specification
CPU	AMD Ryzen 5 3500X, 6-Core Processor
RAM	39 GB
GPU	NVIDIA GPU with CUDA support (unspecified model)
Storage	238.5 GB NVMe SSD

VI. Experimental Results

This section presents our experimental analysis of different Bangla TTS models based on objective and subjective evaluation metrics. The models include Tacotron2[1], Fast-Speech2[2], VITS [3], and Grad-TTS[4].

Objective Evaluation Summary

We evaluated 50 test samples for each model, where we have used the K-means clustering algorithm for sampling from the bigger dataset consists of 18.43 hours of audio. Table 4 shows the average performance across five metrics: Signal-to-Noise Ratio (SNR), Word Error Rate (WER), Phoneme Error Rate (PER), Mel-Spectrogram MSE, and Real- Time Factor (RTF). On the other hand, Table 5 presents the corresponding standard deviations. Note that Lower value of WER, PER, MSE are considered as better whereas higher value of SNR is considered as better result and RTF equal to 1 is considered as the best result. On the contrary, the lower standard deviation is better for each category of results.

Table 4. Average results across key metrics

Model	SNR↑	WER↓	PER↓	MSE↓	RTF
Tacotron2	-3.10	0.812	0.357	357.5	4.95
FastSpeech2	-2.97	0.835	0.371	395.7	3.67
VITS	-2.06	0.780	0.328	360.9	4.81
Grad-TTS	-1.99	0.786	0.329	368.3	4.70

Table 5. Standard deviation of metrics across test samples.

Model	SNR	WER	PER	MSE	RTF
Tacotron2	2.10	0.145	0.081	60.4	0.58
FastSpeech2	1.95	0.150	0.076	66.1	0.69
VITS	1.58	0.132	0.069	56.2	0.68
Grad-TTS	1.66	0.133	0.070	55.3	0.67

Subjective Evaluation (MOS)

A Mean Opinion Score (MOS) survey was conducted using 50 listeners. Ratings ranged from 1 (worst) to 5 (best). Table 6 presents the average scores per model.

Table 6. Average MOS Ratings based on listener survey.

Model	Avg. MOS
Tacotron2	3.45
FastSpeech2	3.50
VITS	3.20
Grad-TTS	3.30

Visual Analysis of Benchmark Metrics

To further illustrate model performance, the following figures display the benchmark metrics and their standard deviations. These visuals help highlight trade-offs between intelligibility, synthesis quality, and real-time efficiency.

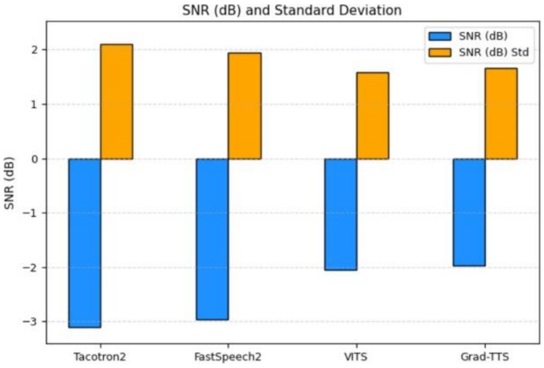

Fig. 4. SNR and standard deviation for Bangla TTS models.

SNR reflects the clarity of the generated audio. From Figure 4, we can clearly see that Grad-TTS[4] achieved the highest score at -1.99 dB, indicating it produced cleaner speech with less background noise compared to the others. In contrast, Tacotron2[1] had the lowest SNR at -3.10 dB, making it the noisiest model among the four. Although the absolute values are negative due to the metric's calculation method, a less negative value is better, favoring Grad-TTS[4] overall for this criterion.

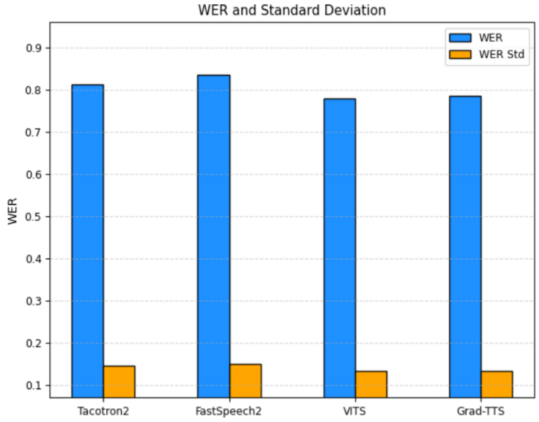

Fig. 5. WER and standard deviation for Bangla TTS models.

Word Error Rate (WER) measures the accuracy of the generated words compared to the reference transcript. Figure 5 reflects that VITS [3] outperformed all other models with a WER of 0.780. This suggests it generated the most intelligible and accurate words. On the other hand, FastSpeech2[2] performed the worst with a WER of 0.835, showing more frequent word-level mistakes during synthesis.

Phoneme Error Rate (PER) evaluates pronunciation accuracy. Here, at Figure 6, VITS[3] again demonstrated the best performance with a PER of 0.328, closely followed by Grad-TTS[4] at 0.329. This reinforces VITS's[3] strength in maintaining phonetic fidelity. FastSpeech2[2] had the highest PER at 0.371, reflecting greater difficulty in accurately generating phoneme sequences.

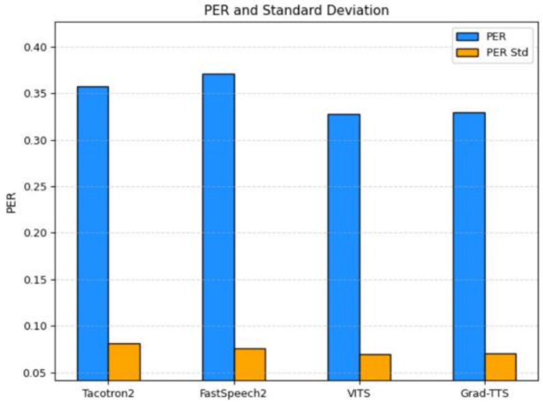

Fig. 6. PER and standard deviation for Bangla TTS models.

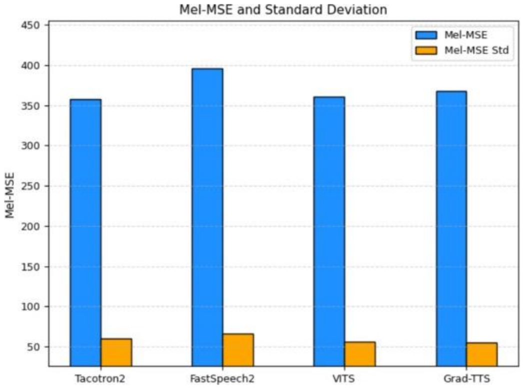

Fig. 7. Mel-MSE and standard deviation for Bangla TTS models.

The Mel-Spectrogram Mean Squared Error (Mel-MSE) indicates how closely the generated spectrograms match the ground truth. Figure 7 indicates that VITS [3] achieved the lowest MSE at 360.9, making it the most accurate in reproducing the acoustic structure of real speech. FastSpeech2[2] had the highest Mel-MSE of 395.7, suggesting poorer spectrogram generation compared to the others.

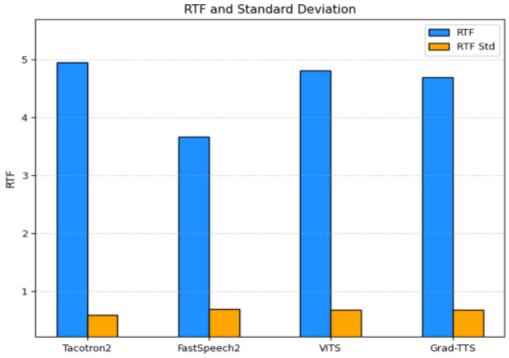

Fig. 8. RTF and standard deviation for Bangla TTS models.

The Real-Time Factor (RTF) measures the synthesis speed relative to real-time. Figure 8 portrays that FastSpeech2[2] had the lowest RTF at 3.67, making it the fastest among the tested models. While none of the models operate in real-time (RTF closer to 1), FastSpeech2's[2] performance was comparatively better. VITS[3] had the slowest synthesis speed with an RTF of 4.81, indicating it is the most computationally intensive during inference.

VII. Conclusion

This work marks a significant step toward advancing Bangla TTS research and supporting the broader digital transformation in Bangladesh. We developed a structured benchmarking framework that evaluates leading Bangla TTS models—Tacotron2[1], FastSpeech2[2], VITS[3], and Grad-TTS[4]—using both objective metrics (e.g., SNR, WER, PER, Mel-MSE, RTF) and subjective evaluations (MOS). A refined dataset was created to ensure reliable testing, and a web interface with an API was implemented to promote accessibility and interaction. These contributions aim to close the performance gap between Bangla and high-resource in speech synthesis, laying the groundwork for inclusive and natural human-computer interaction in the Bangla language.

Looking ahead, expanding the dataset to include more speakers and regional variations is a key priority. Future efforts will also focus on integrating perceptual and prosody-based evaluation metrics, optimizing models for real-time performance, and enhancing expressiveness in synthesized speech. Lightweight architecture suited for mobile deployment, improved prosody modeling, and multilingual transfer learning are promising directions. We also plan to develop a publicly available benchmarking platform, allowing community engagement and continuous improvement of Bangla TTS technologies.

Acknowledgments

Authors are thankful to Mushahid Intesum and Abdullah Ibne Masud for providing the dataset of 18.43 hours.

References

1. Shen J. et al., "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," in IEEE ICASSP, pp. 4779–4783, 2018.
2. Ren Y., Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fast- Speech 2: Fast and high-quality end-to-end text to speech," in Proceedings Int. Conf. Learn. Represent., 2021.
3. Kim J., S. Kim, J. Kong, and S. Yoon, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in Proceedings Int. Conf. Mach. Learn., pp. 5530–5540, 2021.
4. Popov V. et al., "Grad-TTS: A diffusion probabilistic model for text-to-speech," in Proc. Int. Conf. Mach. Learn., pp. 8599–8608, 2021.
5. Ito K. and L. Johnson, "The LJ Speech Dataset," 2017. Available: https://keithito.com/ LJ-Speech-Dataset
6. Panayotov V., G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in Proc. IEEE ICASSP, pp. 5206–5210, 2015.
7. Rousseau A., P. Deléglise, and Y. Estève, "Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks," in Proc. LREC, pp. 3935–3939, 2014.
8. Yamagishi J. et al., "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2019. Available: https: //datashare.ed.ac.uk/ handle/10283/3443
9. Ardila R. et al., "Common voice: A massively-multilingual speech corpus," in Proceedings LREC, pp. 4218–4222, 2020.
10. Intesum M., A. I. Masud, M. A. Islam, and M. R. Karim, "A robust text-to-speech system in Bangla with stochastic duration predictor," Technical Report, Independent Research Work, at University of Dhaka, 2023.

    Available: https: //github.com/mushahidintesum/ speech_synthesis_in_bangla
11. Moliner A. and J. L. Serrano, "Audio denoising using U-net convolutional networks," in Proc. IEEE ICASSP, pp. 238–242, 2019.
12. Hugging Face, "TTS Arena." Available: https://huggingface. co/spaces/fffiloni/tts-arena
13. Picovoice, "Latency Benchmark." [Online]. Available: https:// picovoice.ai/benchmark/
14. OpenVoice, "Voice Cloning Bench- mark." Available: https:// github.com/myshell-ai/OpenVoice
15. ElevenLabs, Smallest.ai, and Open- Voice, "TTS Benchmark 2025." Available: https:// ttsbenchmark.com
16. Vaswani A. et al., "Attention is all you need," in Proceedings NeurIPS, pp. 5998–6008, 2017.
17. Kong J., J. Kim, and J. Bae, "HiFi- GAN: Generative adversarial networks for efficient and high-fidelity speech synthesis," in Proc. NeurIPS, 2020.
18. Schneider S. et al., "Exploring trans- fer learning with TTS for low-resource languages," in Proceedings IEEE SLT, pp. 531–538, 2021.
19. van A. den Oord et al., "WaveNet: A generative model for raw audio," arXiv preprint, arXiv:1609.03499, 2016.
20. Kurdish K., M. Kazakh, and B. Bangla, "Improving low-resource TTS using native phoneme corpora: A comparative study on Bangla, Kurdish, and Kazakh," in Proc. Int. Conf. Speech Technol., pp. 102–107, 2021.
21. Bengali.AI and Google, "Bengali.AI Speech Dataset," 2020. Available: https://bengali.ai/ datasets/