STFT-GradTTS: A Robust, Diffusion-based Speech Synthesis System with iSTFT decoder for Bangla

Abstract

Text-to-speech (TTS) synthesis is a critical area in speech and language processing, with extensive applications in assistive technologies, virtual assistants, and automated content generation. Despite Bangla being the seventh most spoken language globally, high-quality TTS datasets remain scarce, limiting advancements in Bangla speech synthesis. To bridge this gap, we introduce a meticulously curated single-speaker Bangla audio dataset comprising over 20 hours of clean speech. Our dataset ensures diverse linguistic coverage, incorporating compound letters, long vowels, Sanskrit words, and varied sentence structures while maintaining phonetic balance.

In addition to the dataset, we propose *STFT-GradTTS*, a novel diffusion-based TTS model featuring a Multi-Stream iSTFT decoder and a Stochastic Duration Predictor (SDP). The iSTFT-based decoder synthesizes high-fidelity waveforms by decomposing signals into sub-bands using learnable synthesis filters, enhancing spectral clarity. The SDP models phoneme duration distributions, capturing the natural variability in speech pacing. Together, these innovations address key limitations of GradTTS, particularly in duration accuracy and audio naturalness.

We validate our approach through blind subjective evaluations using Mean Opinion Score (MOS) assessments and objective evaluations using Mel Cepstral Distortion (MCD) and ViSQOL scores, demonstrating that *STFT-GradTTS* significantly outperforms the baseline GradTTS in speech quality, naturalness, and duration modeling. Our work not only establishes a new benchmark for Bangla Text-to-Speech (TTS) systems but also provides a solid foundation for future research in low-resource language speech synthesis.

Keywords: Text-to-Speech, Low-Resource Language, Bangla Speech Dataset, Diffusion-Based TTS

1 Introduction

Text-to-Speech (TTS) systems synthesize natural-sounding speech from textual input using deep neural networks, facilitating applications such as audiobook narration, voice anonymization, public service accessibility, and language preservation [1–4]. These systems typically consist of three main stages: text preprocessing, intermediate representation generation (e.g., mel-spectrograms), and waveform synthesis.

Despite significant advancements in neural TTS, most models are predominantly trained on English due to the availability of large-scale datasets such as LJSpeech [5] and Multilingual LibriSpeech [6]. Extending these models to other languages requires language-specific datasets or adaptation techniques such as zero/few-shot learning [3, 4]. However, such approaches face severe limitations in low-resource languages like Bangla due to the scarcity of high-quality speech corpora [7–9]. As a result, existing Bangla TTS models often rely on linguistic approximations to mitigate data limitations, leading to unnatural speech synthesis and difficulties in handling complex phonetic structures.

Bangla presents unique challenges for TTS synthesis, including compound letters (Juktoborno), long vowels, Sanskrit-origin words, and phoneme duration variations. Zero-shot and few-shot TTS models trained primarily on high-resource languages struggle to capture these linguistic complexities, resulting in speech that is inconsistent or unintelligible. Addressing these limitations requires a large-scale, high-quality Bangla speech dataset that not only enhances model training but also establishes standardized evaluation metrics.

To improve Bangla TTS synthesis, we propose **Stft-GradTTS**, a diffusion-based TTS model featuring a **stochastic duration predictor (SDP)** and a **multi-stream inverse short-time Fourier transform (iSTFT) decoder**. The SDP models phoneme duration distributions, improving speech pacing and prosody, while the iSTFT decoder enhances spectral clarity by decomposing waveforms into sub-bands using learnable synthesis filters. Together, these components address the limitations of GradTTS [2], particularly in duration accuracy and speech naturalness.

Furthermore, we introduce a high-quality **20-hour single-speaker Bangla speech dataset**, curated according to Coqui TTS guidelines[1]. This dataset ensures diverse linguistic coverage, balancing phoneme distributions, grammatical structures, and sentence complexity. Compared to prior datasets such as Google Bangla TTS [10], which contains only three hours of speech, our dataset provides a significantly larger and more consistent corpus for training high-quality Bangla TTS models.

Beyond dataset size, we introduce a novel **evaluation metric** that assesses phoneme coverage, compound letter representation, and linguistic diversity, setting a benchmark for future Bangla TTS research. Our dataset serves as a foundational resource for training diffusion-based TTS models and advancing Bangla speech synthesis.

1.1 Contributions

This paper makes the following key contributions:

- We propose **Stft-GradTTS**, a diffusion-based TTS model integrating a stochastic duration predictor and a multi-stream iSTFT decoder, improving Bangla speech synthesis quality and naturalness.
- We curate a **20-hour high-quality Bangla TTS dataset**, significantly improving upon existing datasets and addressing the scarcity of Bangla speech resources.
- We introduce a **novel evaluation metric** for Bangla TTS datasets, ensuring robust phoneme coverage, linguistic diversity, and standardized benchmarking.

1.2 Paper Organization

The remainder of this paper is structured as follows: Section 2 reviews related works in TTS synthesis, diffusion models, and Bangla TTS. Section 3 details the dataset creation process and evaluation criteria. Section 4 describes our proposed diffusion-based TTS model. Section 5 presents experimental results and analysis, followed by conclusions and future directions in Section 6.

2 Related Works

Recent advancements in text-to-speech (TTS) synthesis have been driven by deep learning architectures, each improving synthesis quality and efficiency. This section reviews key models categorized by their underlying architectures, highlighting their primary components and modifications. Additionally, we discuss Bangla TTS models and their unique challenges, emphasizing the need for robust datasets and novel modeling techniques.

2.1 Neural TTS Models

One of the foundational neural TTS models, Tacotron [11], introduced an end-to-end approach for generating speech from text, utilizing a CBHG (Convolutional Block and Highway Networks with Gated Recurrent Units) encoder and a decoder producing spectrograms, later converted to waveforms using the Griffin-Lim algorithm [12]. Tacotron2 [13] refined this by generating mel-spectrograms and employing a WaveNet vocoder [14] for higher-fidelity synthesis. However, Tacotron-based models suffer from sequential dependencies, leading to slow inference.

Convolutional architectures were explored to address these limitations. DeepVoice [15] introduced a modular TTS pipeline with phoneme conversion, segmentation, duration prediction, and waveform synthesis. DeepVoice3 [16] replaced feed-forward components with CNN-based encoders and convolutional attention, improving parallelization but struggling with long-range dependencies.

[1] https://github.com/coqui-ai/TTS/wiki/What-makes-a-good-TTS-dataset

The adoption of Transformer architectures further optimized TTS models. TransformerTTS [17] replaced RNN-based components with self-attention, enhancing scalability. FastSpeech [18] removed explicit attention mechanisms using a length regulator, improving inference speed. FastSpeech2 [19] added variance predictors for pitch, energy, and duration, refining speech prosody. Attentron [20] extended this by integrating few-shot speaker adaptation, though transformer-based models sometimes struggle with prosody control.

2.2 Diffusion-Based Models

Diffusion-based models have emerged as a promising alternative, leveraging stochastic processes for high-quality speech synthesis. GradTTS [2] replaces the flow-based decoder of GlowTTS [21] with a stochastic differential equation (SDE)-based diffusion model, improving naturalness and inference efficiency. GuidedTTS [22] employs an unconditional diffusion model guided by a WaveNet-based phoneme classifier [14], ensuring accurate phonetic representation. E3 TTS [23] eliminates intermediate representations like spectrograms, directly modeling waveforms using a U-Net-based diffusion model [24] and BERT-based text encoding [25], enabling zero-shot speaker adaptation. Despite their superior speech quality, diffusion-based models require significant computational resources.

2.3 Neural Audio Codecs for TTS

Neural audio codecs have been explored for compressing speech into compact representations while maintaining high fidelity. SoundStream [26] and Encodec [27] use vector-quantized variational autoencoders (VQ-VAEs) to generate token-based audio representations.

NaturalSpeech2 [4] extends NaturalSpeech by integrating Encodec for audio compression and using a diffusion-based decoder instead of a VAE. Additionally, it introduces attention-based pitch and duration prediction, enabling zero-shot synthesis while preserving prosodic features. Vall-E [3] employs a neural audio codec in an encoder-decoder framework, treating TTS as a language modeling problem. It synthesizes high-quality speech from unseen speakers using only a three-second prompt, maintaining speaker identity, emotion, and acoustic environment. While neural audio codec-based models achieve state-of-the-art performance, they require extensive data, making adaptation to low-resource languages challenging.

2.4 Existing Bangla TTS Systems

Bangla TTS research remains limited, with most efforts focusing on text normalization and rule-based phoneme mapping. Tanzir et al. [9] proposed a text normalization scheme for Bangla, integrated into an end-to-end system using Merlin [28] and the WORLD vocoder [29]. Their approach included:

1. Replacing long vowels with short vowels
2. Decomposing vowel diphthongs into separate vowels
3. Decomposing consonant conjuncts

While normalization enhances text processing, it may alter phonological characteristics, reducing naturalness. Additionally, the feed-forward neural network in Merlin struggles to model complex linguistic structures.

Khandaker et al. [30] proposed romanization of Bangla text using phoneme rules to simplify synthesis, but this approach strips essential linguistic features, resulting in robotic, unnatural speech. Gutkin et al. [8] addressed the low-resource challenge by bootstrapping text normalization from Hindi and crowdsourcing speech data. Their LSTM and Hidden Markov Model-based [31] system, however, failed to model Bangla phonological features effectively, leading to unnatural outputs.

Despite advancements in TTS synthesis, existing models for Bangla remain underdeveloped and face key challenges. While state-of-the-art English TTS models leverage diffusion-based approaches and neural audio codecs, their reliance on large datasets makes adaptation difficult for low-resource languages like Bangla. Existing Bangla TTS systems focus on text normalization and phoneme mapping, often compromising linguistic richness and naturalness. These limitations highlight the need for a robust Bangla TTS system capable of modeling phonological nuances while leveraging modern deep learning techniques for high-quality and efficient synthesis.

3 Dataset

To develop an effective Text-to-Speech (TTS) model, the dataset must meet specific criteria to ensure high-quality training and output. A well-constructed TTS dataset is recommended[2] to exhibit the following characteristics:

- Free from mistakes[3] and noise.
- Uniform tone and pitch across all audio clips.
- Coverage of all phonemes in the target language.
- Natural-sounding speech.
- Recorded in a lossless format, such as WAV.
- Gaussian-like distribution in both clip and text lengths.

3.1 Existing Datasets and Limitations

The quality of the dataset is fundamental to the success of TTS research. This section explores existing datasets and their potential limitations, particularly in the context of Bangla TTS.

3.1.1 Existing Datasets in English and Other Languages

Several high-quality datasets have significantly contributed to TTS research, particularly in English. Notable examples include:

- **LJSpeech**[5]: A public-domain dataset containing 13,100 short audio clips derived from seven non-fiction books.
- **LibriSpeech**[32]: A large-scale dataset featuring 1,000 hours of speech sampled at 16 kHz.
- **VCTK Corpus**[33]: A dataset with recordings from 110 English speakers, each reading around 400 sentences from newspapers, the Rainbow Passage, and an elicitation paragraph. It covers various accents, facilitating diverse regional speech modeling.
- **TEDx Spanish Dataset**[34]: A Spanish-language dataset containing 24 hours of TED Talks, aiding Spanish TTS research.

These datasets provide natural and diverse speech samples, sourced from books, news articles, and other materials. Each dataset contains at least 20 hours of clean, noise-free audio, essential for training robust TTS models.

3.1.2 Limitations of Existing Bangla Datasets

Currently, high-quality Bangla TTS data is limited, with Google's dataset[10] being the primary source. This dataset consists of recordings from native Bangladeshi and Indian Bangla speakers, sampled at 48 kHz in a 16-bit, mono, lossless WAV format.

Despite its quality, the dataset has several drawbacks:

- It is not entirely error-free, as it was collected from volunteers, potentially introducing variations and distortions.
- The dataset size is insufficient, failing to meet the minimum requirement of 20 hours for robust model training.
- Some audio clips contain NULL segments, where a 6-second WAV file may have only 4 seconds of speech, with silence at the beginning and end.
- Background noise is present in several clips, reducing overall dataset quality.
- It includes multiple speakers and two dialects (Bangladeshi and Kolkata Bangla), which may introduce inconsistencies in speech synthesis.

To address these challenges, we propose creating a high-quality Bangla TTS dataset:

- Record audio in a controlled environment with a single speaker to ensure consistency.
- Implement a robust data-cleaning pipeline to remove noise and silence from audio clips.
- Ensure adequate dataset size (minimum 20 hours) for effective TTS model training.

This approach will help overcome the current limitations and facilitate the development of a robust Bangla TTS system.

[2] https://github.com/coqui-ai/TTS/wiki/What-makes-a-good-TTS-dataset

[3] Mistakes include errors, inaccuracies, or unintended deviations from the intended pronunciation. Speech irregularities such as stutters, hesitations, or mispronunciations are also considered mistakes.

3.2 Design and Implementation of the Bangla TTS Dataset

Following the initial survey, we determined the necessity for a superior and larger dataset[4]. This section details the methodology behind the dataset creation.

3.2.1 Data Collection Environment

To ensure high-quality recordings, we designated a dedicated laboratory within our department for data collection. A studio-style recording area was constructed using cardboard and wood, with acoustic foam strategically placed to absorb sound and minimize echoes. Table 1 provides details about the recording equipment.

Table 1 Audio Equipment Used

Equipment Name	**Device Name**
Microphone	Neumann TLM 103 Large-Diaphragm Condenser Microphone
Headphone	Audio Technica ATH-M30x Professional Studio Monitor
DAC (Digital Audio Converter)	Focusrite Scarlett 2i2 (3rd Gen) USB Audio Interface
Reflector	Studio Mic Sound Absorbing Foam Reflector

3.2.2 Raw Data Collection

A female voice artist was engaged for recording, using Audacity as the recording software. The reading material was sourced from books, newspapers, and online articles, with individual tracks ranging from 5 to 15 minutes in duration.

To ensure naturalness, clips shorter than 10-15 seconds were avoided. Longer clips allowed the artist's voice to retain natural intonations while preventing fatigue. Ultimately, the final dataset consists of audio segments lasting no more than 12 seconds.

Aiming for 20 hours of processed audio, we collected 27.5 hours of raw data, accounting for a 25% loss in conversion. Metrics were developed to ensure linguistic diversity, including:

- Coverage of all phonemes using a 50,000-word Tadbhava dictionary.
- Balance of active and passive sentences across different grammatical persons.
- Inclusion of Bangla compound letters (306 Juktoborno).
- Content diversity spanning multiple genres to prevent topic bias.
- Ensuring colloquial language to match natural Bangladeshi speech patterns.

3.2.3 Preprocessing Raw Data

The preprocessing of raw data represents the most challenging and labor-intensive phase of dataset creation. This phase is divided into three phases:

1. **Data Cleaning:** The primary objective of this stage is to eliminate noise from the raw audio data. Static or mechanical noise was removed using iZotope RX10, an advanced signal processing software that identifies and isolates unwanted noise components while preserving the original audio quality. For breathing noise and other human-made disturbances, a manual approach was adopted, involving careful listening and manual editing.

   After the cleaning process, the dataset was reduced from 27.5 hours to 25.5 hours, resulting in a data loss of approximately 7.27%.
2. **Clipping and Text Alignment:** In this stage, cleaned audio tracks are segmented into shorter clips of up to 12 seconds. Each audio clip is paired with a corresponding text file, ensuring consistent naming conventions across both file types. Audacity software was used for manual segmentation.

   The clipping strategy varied depending on sentence structure. While individual clips were ideally mapped to single sentences, longer sentences were divided into multiple clips, and shorter sentences were grouped together when necessary.

   The text corresponding to each clip is extracted manually or using OCR tools such as Google Lens for printed material. The final

[4] Our dataset exceeds 20 hours in volume, ensuring comprehensive coverage. It encompasses a diverse array of materials, including books from various genres, articles, newspapers, and more. Notably, for simplicity, our database features a single speaker and a single dialect, providing distinct advantages for model training.

text dataset is encoded in UTF-8 and stored in structured files.

During clipping, additional cleaning is performed to remove residual background noise. Each clip is validated several times using high-performance headphones. The clipping phase reduced the total dataset duration from 25.5 hours to 21.21 hours, representing a data loss of approximately 16.82%. The cumulative data loss from raw to clipped data amounted to 22.8%.

The major sources of data loss include the removal of unnecessary silence, repetitions, and unintelligible segments.

3. **Data Validation**

After generating approximately 16,000 clips (21.21 hours), a thorough validation was conducted to ensure:

- Text and audio clips were correctly paired.
- Noise-free and high-quality audio files.
- Clear pronunciation and absence of speech irregularities.

The rechecking process resulted in an additional 2.21% data loss, reducing the final dataset to 20.74 hours. The overall data loss from the raw dataset to the final processed dataset amounted to 24.58%.

The entire dataset preprocessing is summarized in Figure 1.

3.3 Data Statistics and Metrics

This section provides a comprehensive statistical analysis to evaluate the distribution of information within our dataset, identify trends, and ensure the robustness and reliability of the data for research purposes. Furthermore, we present key metrics that serve as benchmarks for assessing the accuracy, completeness, and relevance of our dataset.

3.3.1 Statistics of Audio Data

Our dataset is composed of audio clips with durations from 15 to 12 seconds. Figure 2 presents a bar chart depicting the distribution of clip durations.

The mean duration of the audio clips is computed to be 4.98 seconds, with a median duration of 4.49 seconds, indicating a right-skewed distribution. Figure 3 displays the distributions of average pitch and amplitude separately.

Speakers inherently exhibit variations in pitch and amplitude, influenced by factors such as context and emotional state. A diverse range of pitch and amplitude values is essential in a Text-to-Speech (TTS) dataset to enable the model to capture and synthesize natural speech patterns effectively.

3.3.2 Statistics of Text Data

The dataset statistics include various metrics such as total word count, unique word count, average word count per clip, total number of clips, and average clip length. The dataset comprises a total word count of 165,486, with 26,448 unique words.

Bangla script includes 306 compound letters (zukto borno), many of which are not commonly used in contemporary Bangla. However, it is important to incorporate the compound letters relevant to daily communication. Our analysis indicates that 230 unique compound letters are present in the dataset. Table 2 provides examples of some common compound letters.

The dataset contains a total of 17,351 sentences, though approximately 300 sentences involve instances where multiple sentences correspond to a single clip. Table 3 provides a statistical summary of the dataset.

3.3.3 Genre Distribution

A diverse dataset encompassing various genres is crucial for training a robust TTS model in Bangla. The dataset incorporates genres such as news articles, autobiographies, religious discussions, economic literature, novels, and war-related content. Figure 4 illustrates the distribution of different genres within the dataset.

3.3.4 Sentence-Type Distribution

Incorporating a variety of sentence structures—simple, compound, and complex—is essential for building a natural and diverse voice synthesis model. The dataset contains 8,978 simple sentences, 3,031 compound sentences, and 2,980 complex sentences. Figure 5 provides a visual representation of the different distributions within the dataset.

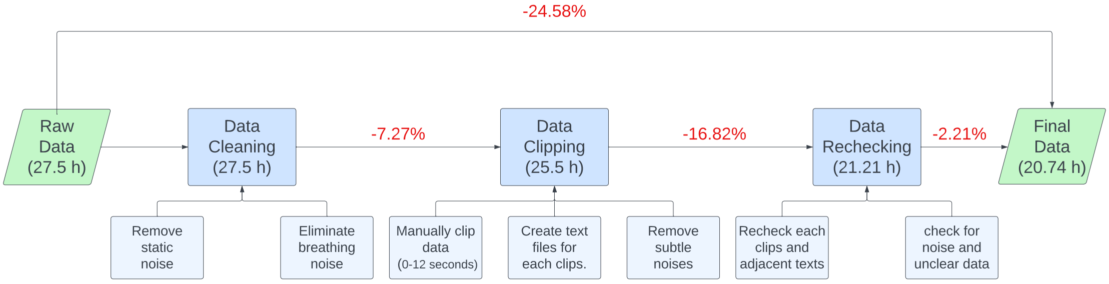
-24.58%
Raw Data (27.5 h)
Data Cleaning (27.5 h)
-7.27%
Data Clipping (25.5 h)
-16.82%
Data Rechecking (21.21 h)
-2.21%
Final Data (20.74 h)
Remove static noise
Eliminate breathing noise
Manually clip data (0-12 seconds)
Create text files for each clips.
Remove subtle noises
Recheck each clips and adjacent texts
check for noise and unclear data

Fig. 1 Summary of data preprocessing and loss

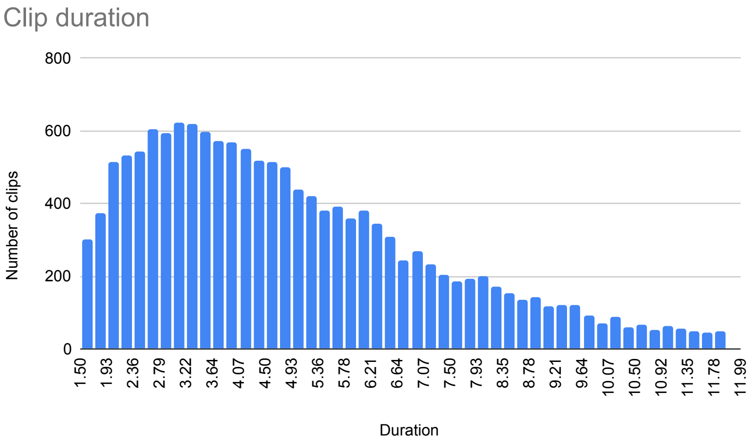

Fig. 2 Histogram of clip durations

This analysis highlights the linguistic diversity present in our dataset, which ensures that the TTS model can effectively synthesize speech across a broad range of textual styles and linguistic structures. The inclusion of diverse genres, sentence types, and linguistic elements enhances the capability of the model to produce fluent and contextually appropriate Bangla speech synthesis.

4 Proposed TTS Model

Our evaluation of GradTTS when trained on our Bangla dataset revealed several critical limitations: poor audio clarity, inadequate reproduction of certain phonemes and words, and inconsistent speech pacing. To address these issues, we developed STFT-GradTTS with two key enhancements: (1) Multi-Stream iSTFT blocks applied to the magnitude and phase variables for improved sub-band signal generation, significantly enhancing audio clarity; and (2) a Stochastic Duration Predictor replacing the deterministic approach of GradTTS to better capture the natural variability in speech timing. The complete architecture of our proposed model is illustrated in Figure Figure 7.

4.1 Multi-Stream iSTFT

Our extensive experimentation with GradTTS revealed persistent challenges in accurately reproducing the nuanced phonetics of Bangla speech. The model particularly struggled with certain compound characters, diphthongs, and phonemic combinations unique to Bangla, resulting in unclear articulation and reduced intelligibility. To overcome these limitations, we integrated Multi-Stream Inverse Short-Time Fourier Transform (iSTFT) blocks into our architecture—a design choice inspired by the work of Kim et al. [35], who demonstrated significant improvements in

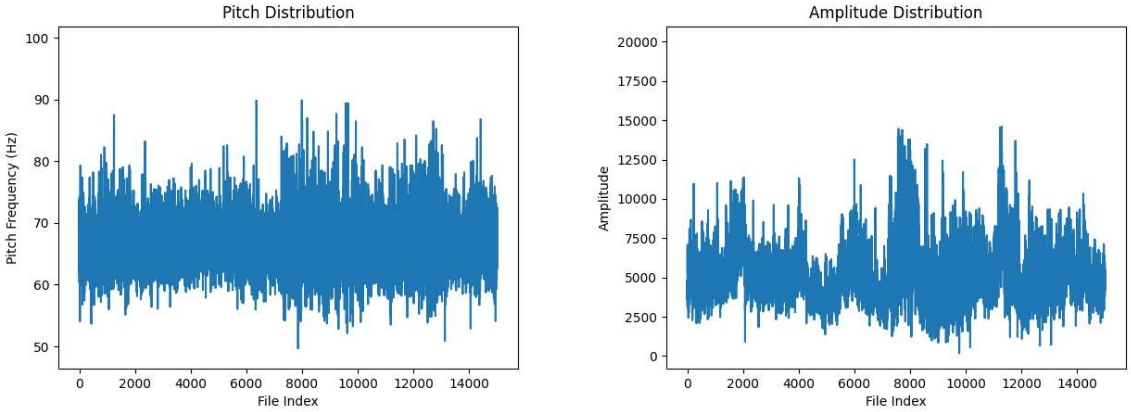

Fig. 3 Distributions of average pitch and amplitude

Table 2 Examples of Bangla compound letters

Compound letter	Decomposition	IPA	Example	IPA	Meaning
ক্ক	ক + ক	k + k	ধাক্কা	d̪ʰɐkkɐ	Push
খ্র	খ + র	Kʰ + r	খ্রিস্টান	kʰɹiʃtan	Christian
গ্ধ	গ + ধ	g+ dʱ	মুগ্ধ	mugd̪ʱo	Fascinated
ঘ্র	ঘ + র	gʱ+ r	ঘ্রাণ	gʱrɐn	Scent
ঙ্খ	ঙ + খ	ŋ + kʰ	শঙ্খ	ʃɔŋkʰo	Conch-shell
ত্ত্ব	ত + ত + ব	t + t + b	সংখ্যাতত্ত্ব	ʃɔŋkʰkʰɐt̪ɔt̪t̪o	Numerology
ঢ্য	ঢ + য	ɖʱ + dʒ	বর্ণাঢ্য	bɔrnɐddʱo	Colourful
ত্র্য	ত + র + য	t + r + dʒ	বৈচিত্র্য	boi̯cɪt̪ro	Diversity
প্ত	প + ত	p + t	সপ্তাহ	ʃɔpt̪ɐho	Week
ম্ম	ম + ম	m + m	সম্মান	ʃommɐn	Respect

Table 3 Statistical Summary of the Dataset

Category	**Value**
Number of clips	14,989
Mean clip duration (seconds)	4.98
Total audio duration (hours)	20.74
Total word count	165,486
Average word count per clip	11.04
Unique word count	26,448
Unique compound letters	230
Total number of sentences	17,051
Interrogative sentences	1,275
Exclamatory sentences	380

speech quality through the substitution of traditional neural network layers with specialized iSTFT blocks.

In our STFT-GradTTS model, we strategically position these iSTFT blocks immediately following the diffusion decoder (see Figure 7). This placement allows the blocks to refine the latent representations produced by the diffusion process before final waveform generation, substantially enhancing spectral clarity and phonetic precision. The Multi-Stream iSTFT block architecture is illustrated in detail in Figure Figure 6.

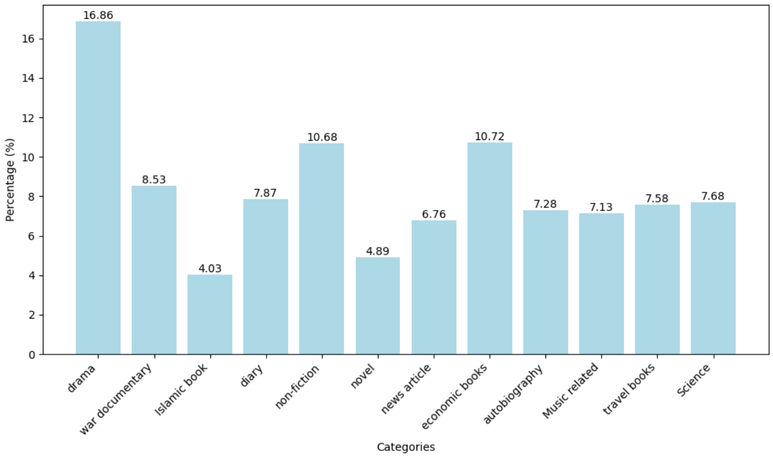

Fig. 4 Distribution of Genres in the Dataset

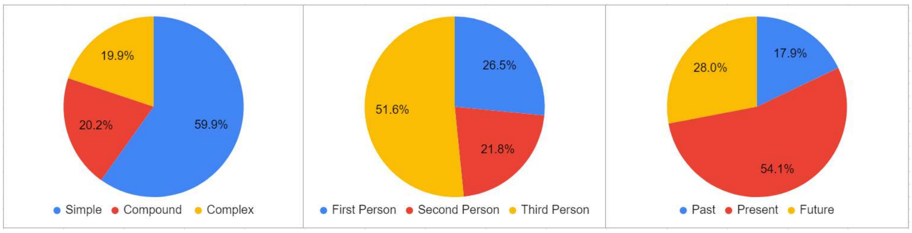

Fig. 5 Distribution of Sentence Types

The processing pipeline within our Multi-Stream iSTFT implementation consists of six carefully designed stages:

- **Input Processing**: The diffusion decoder outputs a latent representation z that encodes acoustic features derived from the input text. This representation serves as the foundation for the iSTFT block's processing chain. While the diffusion decoder captures the broad spectral characteristics, the iSTFT block refines these representations to improve phonetic accuracy.
- **Feature Extraction via Conv1D**: The latent representation undergoes processing through a series of 1D convolutional layers with varying kernel sizes. These layers implement multi-receptive field fusion—a technique that enables the simultaneous capture of both fine-grained local phonetic features (through smaller kernels) and broader contextual information (through larger kernels). This multi-scale approach is particularly beneficial for Bangla, where phonetic realization often depends on surrounding context.
- **Sub-Pixel Convolution for Upsampling**: To increase temporal resolution, we employ sub-pixel convolutional neural network layers that perform efficient upsampling. Unlike traditional upsampling methods that can introduce artifacts, our approach utilizes zero-padding-based techniques implemented through trainable CNN layers without bias parameters. This design choice ensures smoother transitions between adjacent time frames and preserves the fine spectral details crucial for accurate phoneme reproduction.
- **Multi-Receptive Field Fusion**: This stage enhances feature representation by integrating information across multiple temporal scales. The network applies convolutions with different dilation rates to capture dependencies at various temporal resolutions simultaneously. For Bangla speech, this is particularly important as certain phonetic features (such as nasalization and aspiration) manifest across different time scales.

- **Non-Linear Transformation**: LeakyReLU activation functions introduce controlled nonlinearity after each convolutional layer. With a small negative slope, these activations allow for the propagation of negative values through the network while maintaining differentiability. This property is essential for preserving the phase information in speech signals, which greatly contributes to naturalness and clarity.
- **Waveform Synthesis**: In the final stage, the processed features are transformed into the time-frequency domain and subsequently decomposed into four distinct sub-band waveforms using the iSTFT operation. These subbands effectively capture different frequency regions of speech, which are then integrated into a full-band waveform through a pseudo-QMF (Quadrature Mirror Filter) bank. This multi-band approach significantly improves the model's ability to reproduce the complex harmonic structure of Bangla phonemes.

The iSTFT block architecture, as depicted in Figure 6, builds upon the HiFi-GAN vocoder framework [36] but introduces critical modifications tailored for Bangla speech synthesis. It employs sub-pixel convolutional neural network layers for upsampling, facilitated by zero-padding-based techniques and trainable CNN layers without bias parameters. The block generates four distinct stream waveforms, each capturing different spectral components of speech, which are then synthesized in a data-driven manner by trainable CNN layers.

The advantage of this approach for Bangla TTS lies in its ability to model the complex spectral relationships inherent in the language's phonetic inventory. By decomposing the waveform generation process into multiple frequency bands, our model achieves superior reproduction of distinctive features such as aspiration contrasts, nasalization, and the nuanced vowel system characteristic of Bangla. This results in significantly improved intelligibility and naturalness compared to the baseline GradTTS model, particularly for phonetically complex utterances.

4.2 Pseudo-Quadrature Mirror Filter (Pseudo-QMF) Bank

Our model employs a pseudo-quadrature mirror filter (pseudo-QMF) bank to integrate the upsampled sub-band signals into full-band waveforms. Unlike strict QMF banks requiring perfect reconstruction conditions, pseudo-QMF banks approximate these conditions to achieve high-quality reconstruction while maintaining computational efficiency. This approach ensures that the synthesized speech waveforms maintain naturalness and minimize aliasing or distortion artifacts.

4.3 Multi-Resolution STFT Loss

To further enhance synthesized speech quality, we implement a multi-resolution STFT loss during training. This loss function evaluates the discrepancy between predicted and ground truth signals in the frequency domain across multiple resolutions, comprising:

- **Spectral Convergence Loss**: Measures differences in overall spectral structure between predicted and ground truth signals.
- **Log STFT Magnitude Loss**: Quantifies differences in log-scale magnitudes of STFT spectra, preserving fine-grained spectral details.

This loss is computed using multiple STFT configurations with varying FFT sizes, hop sizes, and window lengths, enabling the model to capture both coarse and fine spectral details. During training, this multi-resolution STFT loss is applied using the aforementioned pseudo-QMF-based filter, significantly improving the naturalness and clarity of the synthesized speech.

4.4 Stochastic Duration Predictor

To address the pacing issues in the generated audio, we implement a Stochastic Duration Predictor (SDP) inspired by VITS [1]. Rather than using deterministic methods, the SDP models the distribution of phoneme durations, capturing the natural variability in speech tempo—an essential characteristic as human speakers naturally vary their articulation timing for the same sentence depending on context and prosody.

The duration of a text segment is calculated by summing alignment scores across each column of the alignment matrix, where columns represent

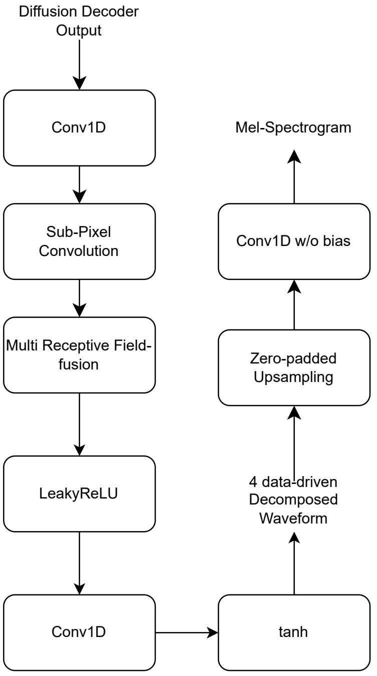

Fig. 6 Multi-Stream iSTFT Block

the alignment between phonemes and temporal segments. The SDP employs a flow-based generative model trained via maximum likelihood estimation to model a distribution of durations rather than discrete values.

To handle the discrete nature of phoneme durations, the model incorporates variational dequantization and variational data augmentation, transforming the discrete duration sequence into a continuous distribution through invertible mappings. This approach enables the model to effectively capture the inherent variability in speech timing.

Mathematically, for a token duration d computed from alignment matrix A, the model introduces auxiliary variables u and v, where $u \in [0, 1)$ ensures $d - u$ is positive, and v is concatenated with d to form a higher-dimensional latent representation. The objective is to maximize the variational lower bound of the log-likelihood of d conditioned on the input text c_{text}:

$$\log p_\theta(d|c_{text}) \geq \mathbb{E}_{q_\phi(u,v|d,c_{text})}\left[\log \frac{p_\theta(d-u,v|c_{text})}{q_\phi(u,v|d,c_{text})}\right], \tag{1}$$

where $p_\theta(d-u, v|c_{text})$ is the prior distribution, and $q_\phi(u, v|d, c_{text})$ is the posterior. The training loss L_{dur} is the negative variational lower bound:

$$L_{dur} = -\mathbb{E}_{q_\phi(u,v|d,c_{text})}\left[\log \frac{p_\theta(d-u,v|c_{text})}{q_\phi(u,v|d,c_{text})}\right] + \mathbb{E}_{q_\phi}[\log(q_\phi(c_{text}))]. \tag{2}$$

Through this stochastic approach to duration prediction, our model achieves more natural pacing and rhythm in the synthesized speech, addressing a key limitation of the original GradTTS architecture when applied to Bangla.

5 Experimental Evaluation

The code was tested on a system with the configurations listed in Table 4.

Table 4 Hardware configuration of testbench

Hardware	Details
GPU Memory	32 GB
GPU	NVIDIA RTX 3060 12GB
Memory	512 GB

5.1 Hyperparameters

The model underwent 100 epochs of training with a batch size of 16, and each training run typically lasted approximately 6 hours from initiation to completion. Due to the model's substantial size, complexity, and prolonged training durations, we were unable to conduct hyperparameter tuning. Therefore, the hyperparameter values provided were recommended by the authors of GradTTS [2], VITS [1] and [35].

5.2 Experimental Results

We present both subjective and objective evaluations of our TTS model, including Mean Opinion

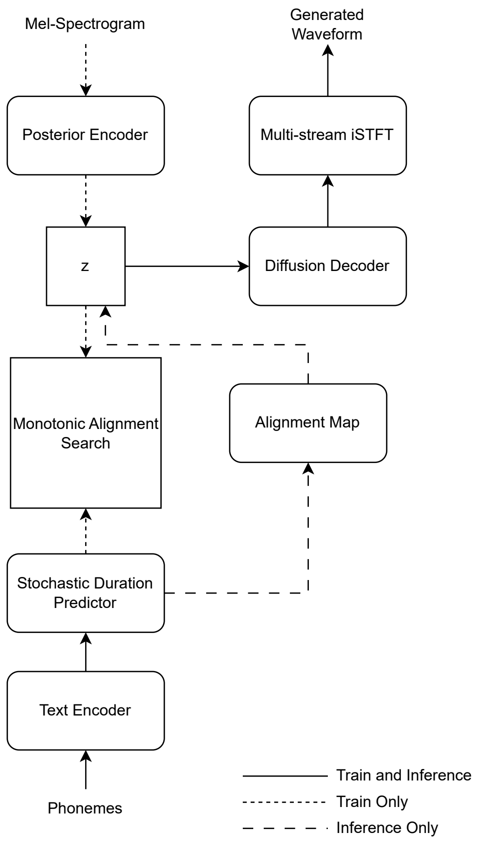

Fig. 7 Proposed model

Score (MOS), Real-Time Factor (RTF), Mel Cepstral Distortion (MCD), and Virtual Speech Quality Objective Listener (ViSQOL) [37] metrics.

5.2.1 Subjective Evaluation

Table 5 and Figure 8 present the comparative MOS and RTF scores, respectively. We conducted an ablation study comparing three model variants: the base Grad-TTS model, Grad-TTS with Stochastic Duration Prediction (SDP), and our proposed STFT-GradTTS. For the subjective evaluation, we designed a survey containing generated audio outputs alongside ground truth samples. Participants rated each audio sample on a scale from 1 (poor) to 5 (excellent). We collected ratings for 10 different text prompts and calculated the mean score with a 95% confidence interval. For all configurations, we performed statistical analysis on the means with known standard deviations. In all cases, we established with statistical significance ($p < 0.005$) that ground truth samples had mean scores below 4.5, while synthesized audio from all models had mean scores above 3.0.

As shown in the results, Grad-TTS with SDP demonstrated a modest improvement over the baseline Grad-TTS, achieving a MOS score increase of 0.13 when trained on Bangla data. This suggests that our proposed stochastic duration prediction component enhances the naturalness and overall quality of the synthesized speech. More notably, STFT-GradTTS exhibited substantial improvement with a MOS score 0.55 higher than the baseline Grad-TTS, confirming that our approach produces significantly more intelligible audio.

Table 5 Model parameters and Mean Opinion Scores (MOS) with 95% confidence intervals

Model	Parameters	MOS
Ground Truth	-	4.56 ± 0.06
GradTTS	15,180,889	3.34 ± 0.06
GradTTS with SDP	16,225,928	3.47 ± 0.07
STFT-GradTTS	17,213,061	3.89 ± 0.05

5.2.2 Objective Evaluation

To complement the subjective assessment, we employed two objective metrics: Mel Cepstral Distortion (MCD) and ViSQOL. MCD quantifies spectral fidelity by measuring the Euclidean distance between Mel-frequency cepstral coefficients of synthesized and reference audio. Lower MCD values (in dB) indicate closer spectral alignment with the reference. High-quality TTS systems typically achieve MCD scores between 4.0–6.0 dB, with human recordings approaching zero distortion. ViSQOL predicts perceptual quality by comparing neurogram representations of synthesized and reference speech, producing a Mean Opinion Score (MOS) ranging from 1 (poor) to 5 (excellent).

Table 6 presents the MCD and ViSQOL scores for our models. We evaluated 20 synthesized samples against corresponding ground truth recordings and calculated mean scores with 95% confidence intervals. Our statistical analysis confirmed that MCD means were below 6.0 and ViSQOL means were above 3.0 with $p < 0.005$ for all models.

The MCD scores demonstrate that our proposed STFT-GradTTS model generates spectrograms with reduced distortion compared to the baseline GradTTS. The incremental improvement observed with GradTTS+SDP confirms that both the SDP and STFT components contribute to enhancing the overall model performance. The ViSQOL scores further corroborate these findings, showing that the baseline GradTTS suffers from more pronounced audio distortions during generation. The significant improvement in STFT-GradTTS (3.79 compared to 3.17 baseline) can be attributed to the introduction of the iSTFT modules. While the scores remain below 4.0, this can be attributed to variations in duration between synthesized and ground truth samples, as well as occasional prosodic inconsistencies, both of which affect the ViSQOL metric.

Table 6 Objective evaluation metrics with 95% confidence intervals

Model	MCD (dB)	ViSQOL
GradTTS	3.21 ± 1.77	3.17 ± 0.04
GradTTS with SDP	3.81 ± 1.52	3.21 ± 0.03
STFT-GradTTS	5.13 ± 1.03	3.79 ± 0.05

5.2.3 Inference Efficiency

To assess computational efficiency, we measured the Real-Time Factor (RTF) for each model variant, as illustrated in Figure 8. RTF quantifies the time required to generate one second of audio, with lower values indicating faster synthesis. Due to its increased parameter count and more complex mathematical operations, STFT-GradTTS exhibited longer synthesis times compared to the baseline. However, we observed consistent synthesis times across varying text lengths, indicating that our model's inference speed remains stable regardless of input complexity.

6 Conclusion

In this study, we curated a substantial audio dataset for Bangla, comprising over 20 hours of speech and encompassing more than 26,000 unique words, ensuring robust coverage of diverse linguistic aspects. To address the unique challenges of Bangla, such as its complex phonological structures and the absence of a dedicated grapheme-to-phoneme dictionary, we developed a Text-to-Speech (TTS) model based on diffusion models. Our model incorporates two key innovations: Multi-Stream iSTFT blocks and a Stochastic Duration Predictor (SDP). These components enable our model to outperform the baseline GradTTS in terms of audio quality, naturalness, and linguistic accuracy. The Multi-Stream iSTFT blocks play a critical role in generating high-quality waveforms by decomposing the signal into sub-band components and reconstructing them into full-band waveforms using a trainable synthesis filter. This data-driven approach ensures that the synthesized audio retains fine-grained spectral details, improving both clarity and naturalness. Additionally, the SDP models the distribution of phoneme durations rather than relying on discrete values, capturing the inherent variability in human speech. This is particularly important for Bangla, where the same sentence may be pronounced with different durations depending on context or speaker. By combining these components, our model achieves more accurate pacing and cadence, closely resembling human speech patterns.

Experimental results demonstrate that our proposed model significantly outperforms GradTTS in audio generation tasks, particularly in terms of pacing, duration accuracy, and overall naturalness. The integration of iSTFT blocks ensures high-quality waveform generation, while the SDP enables precise control over phoneme durations, addressing one of the key limitations of the baseline model. Furthermore, our model maintains computational efficiency during both training and inference, despite the added complexity of these components.

While our approach showcases promising outcomes, some limitations persist, such as occasional mispronunciations of certain words and limited improvements from the context prediction network. Nevertheless, our findings provide a solid

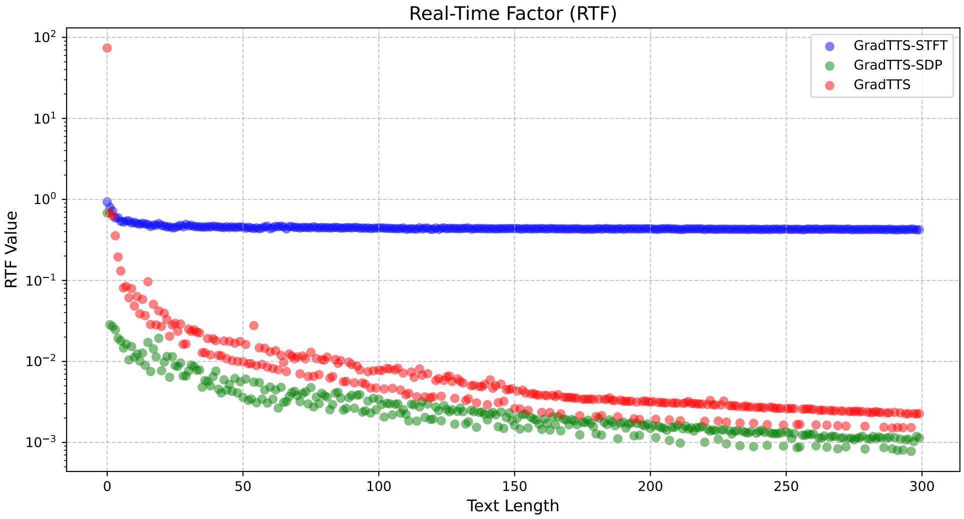

Fig. 8 Inference Speech Comparison. Text length is given in characters.

foundation and serve as a benchmark for future advancements in Bangla TTS research, opening promising avenues for further refinement and innovation in the field.

Declarations

Competing Interests: The authors declare that they have no competing interests.
Funding: Funding information has been anonymized in accordance with the journal's double-blind review policy.
Data Availability The Bangla speech dataset used in this study is available from the corresponding author upon reasonable request.

References

[1] Kim, J., Kong, J., Son, J.: Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech (2021)

[2] Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., Kudinov, M.: Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech (2021)

[3] Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., Wei, F.: Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers (2023)

[4] Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., Qin, T., Zhao, S., Bian, J.: NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers (2023)

[5] Ito, K., Johnson, L.: The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/ (2017)

[6] Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., Collobert, R.: MLS: A Large-Scale Multilingual Dataset for Speech Research (2020)

[7] Shah, N., Tambrahalli, V., Kosgi, S., Pedanekar, N., Gandhi, V.: MParrotTTS: Multilingual Multi-speaker Text to Speech Synthesis in Low Resource Setting (2023)

[8] Gutkin, A., Ha, L., Jansche, M., Pipatsrisawat, K., Sproat, R.: TTS for low resource languages: A Bangla synthesizer. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 2005–2010. European Language Resources Association (ELRA), Portorož, Slovenia (2016). https://aclanthology.org/L16-1317

[9] Islam Pial, T., Salim Aunti, S., Ahmed, S., Heickal, H.: End-to-end speech synthesis for bangla with text normalization. End-to-End Speech Synthesis for Bangla with Text

Normalization," 2018 5th International Conference on Computational Science/ Intelligence and Applied Informatics (CSII), Yonago, Japan, 2018 **1**(1), 66–71 (2018) https://doi.org/10.1109/CSII.2018.00019

[10] Sodimana, K., Pipatsrisawat, K., Ha, L., Jansche, M., Kjartansson, O., Silva, P.D., Sarin, S.: A Step-by-Step Process for Building TTS Voices Using Open Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese (2018). http://dx.doi.org/10.21437/SLTU.2018-14

[11] 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.: Tacotron: Towards End-to-End Speech Synthesis (2017)

[12] Griffin, D., Lim, J.: Signal estimation from modified short-time fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing **32**(2), 236–243 (1984) https://doi.org/10.1109/TASSP.1984.1164317

[13] Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R.A., Agiomyrgiannakis, Y., Wu, Y.: Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions (2018)

[14] Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio (2016)

[15] Arik, S.O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., Shoeybi, M.: Deep Voice: Real-time Neural Text-to-Speech (2017)

[16] Ping, W., Peng, K., Gibiansky, A., Arik, S.O., Kannan, A., Narang, S., Raiman, J., Miller, J.: Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning (2018)

[17] Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M.: Neural speech synthesis with transformer network. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 6706–6713 (2019) https://doi.org/10.1609/aaai.v33i01.33016706

[18] Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.-Y.: FastSpeech: Fast, Robust and Controllable Text to Speech (2019)

[19] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.-Y.: FastSpeech 2: Fast and High-Quality End-to-End Text to Speech (2022)

[20] Choi, S., Han, S., Kim, D., Ha, S.: Attentron: Few-Shot Text-to-Speech Utilizing Attention-Based Variable-Length Embedding (2020)

[21] Kim, J., Kim, S., Kong, J., Yoon, S.: Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search (2020)

[22] Kim, H., Kim, S., Yoon, S.: Guided-TTS: A Diffusion Model for Text-to-Speech via Classifier Guidance (2022)

[23] Gao, Y., Morioka, N., Zhang, Y., Chen, N.: E3 TTS: Easy End-to-End Diffusion-based Text to Speech (2023)

[24] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation (2015)

[25] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019)

[26] Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., Tagliasacchi, M.: SoundStream: An End-to-End Neural Audio Codec (2021)

[27] Défossez, A., Copet, J., Synnaeve, G., Adi, Y.: High Fidelity Neural Audio Compression (2022)

[28] Zhizheng Wu, S.K. Oliver Watts: "Merlin: An

Open Source Neural Network Speech Synthesis System" (2016)

[29] MORISE, M., YOKOMORI, F., OZAWA, K.: World: A vocoder-based high-quality speech synthesis system for real-time applications. IEICE Transactions on Information and Systems **E99.D**(7), 1877–1884 (2016) https://doi.org/10.1587/transinf.2015EDP7457

[30] Ahmed, K., Mandal, P., Hossain, B.M.M.: Text to speech synthesis for bangla language. International Journal of Information Engineering and Electronic Business **11**, 1–9 (2019) https://doi.org/10.5815/ijieeb.2019.02.01

[31] Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state markov chains. The annals of mathematical statistics **37**(6), 1554–1563 (1966)

[32] Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: An ASR corpus based on public domain audio books (2015). https://doi.org/10.1109/ICASSP.2015.7178964

[33] Yamagishi, J., Veaux, C., MacDonald, K.: CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92) (2019)

[34] Hernandez-Mena, C.D.: TEDx Spanish Corpus. Audio and transcripts in Spanish taken from the TEDx Talks; shared under the CC BY-NC-ND 4.0 license. Web Download (2019)

[35] Kawamura, M., Shirahata, Y., Yamamoto, R., Tachibana, K.: Lightweight and High-Fidelity End-to-End Text-to-Speech with Multi-Band Generation and Inverse Short-Time Fourier Transform (2023). https://arxiv.org/abs/2210.15975

[36] Okamoto, T., Toda, T., Kawai, H.: Multi-stream hifi-gan with data-driven waveform decomposition. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 610–617 (2021). https://doi.org/10.1109/ASRU51503.2021.9688194

[37] Chinen, M., Lim, F.S., Skoglund, J., Gureev, N., O'Gorman, F., Hines, A.: Visqol v3: An open source production ready objective speech and audio metric. In: 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6 (2020). IEEE