Sylhet Engineering College

(Shahjalal University of Science and Technology)

Department of Computer Science and Engineering

CSE800

Bengali Voice Cloning: A Neural Network for Zero-shot Synthesis

Durjoy Chandra Paul Mahmudul Hasan Badhan Rubel Ahmod

2019331514 2019331565 2019331545

Department of Computer Science and Engineering

Supervisor

Mohammad Shahidur Rahman, PhD
Professor
Department of Computer Science and Engineering (CSE)

Shahjalal University of Science and Technology (SUST)

22 July 2025

Recommendation Letter from Thesis Supervisor

The thesis titled "**Bengali Voice Cloning: A Neural Network Approach for Zero-Shot Synthesis,"** submitted by Durjoy Chandra Paul, Mahmudul Hasan Badhan, and Rubel Ahmod, is being presented under my supervision.

After thoroughly reviewing the thesis, I confirm my approval for its submission for examination.

Mohammad Shahidur Rahman, PhD
Professor, Department of Computer Science and Engineering (CSE), SUST
Date: 22 July, 2025

Certificate of Acceptance

The thesis is titled **"Bengali Voice Cloning: A Neural Network Approach for Zero-Shot Synthesis"** submitted by **Durjoy Chandra Paul**, **Mahmudul Hasan Badhan** and **Rubel Ahmod**; Student ID. **2019331514, 2019331565,** and **201933145**; Session **2019-20**, to the Department of Computer Science and Engineering, Sylhet Engineering College, has been accepted as satisfactory in partial fulfilment of the requirement for the Degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

Internal

Nayan Kumar Nath

Lecturer

Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet

Internal

Md Lysuzzaman

Lecturer

Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet

Internal

Md. Rasel Ahmed

Assistant Professor

Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet

Internal

Md. Nagrul Islam

Assistant Professor

Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet

Chairman

Md. Abu Naser Mojumder

Head

Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet

Member (External)

Dr. Mohammad Shahidur Rahman

Professor

Department of Computer Science and Engineering
Shahjalal University of Science and Technology

Abstract

Voice cloning technology has wide-ranging applications, such as assisting individuals who have lost their ability to speak, enabling realistic movie dubbing, and facilitating multilingual speech translation while preserving the speaker's identity. However, high-quality Bengali voice cloning remains underexplored due to limited linguistic resources and training data. In this work, we present a comprehensive study of zero-shot Bengali voice cloning using both single-model and ensemble-based approaches, capable of generating natural-sounding cloned audio from just a few seconds of speech—even for Bengali speakers not seen during training.

For model development, we used two publicly available Bengali speech corpora: SUBAK.KO for training the multi-speaker model, and the Development of Annotated Bangla Speech Corpora for training the single-speaker TTS model.

We evaluate two core techniques: (1) the Speaker Encoder (SE) model, based on YourTTS, fine-tuned on a 59-speaker Bengali audio data totaling over 226 hours. It achieved a Mean Opinion Score (MOS) of 3.8 and a cosine similarity of 83**%**, closely approaching the performance of state-of-the-art models like VALL-E and XTTS, which were trained on significantly larger and more diverse corpora. (2) The Speaker Converter (SC) model, trained on a single-speaker dataset with 4 hours of audio, yielded a MOS of 4.0 and 82.64% cosine similarity. To further enhance performance, we introduce an ensemble framework combining the strengths of both models. In our earlier setup with 12 speakers, the ensemble model reached a cosine similarity of 78.34%. In the current expanded setup with 59 speakers, our ensemble models (FreeVC and OpenVoice) achieved cosine similarities of 80.83% and 86.11%, respectively, and a MOS of 4.4, reflecting a significant quality improvement.

Despite using comparatively low-resource training data, our adaptation strategy enables high-fidelity voice cloning, demonstrating that increasing speaker diversity and training iterations can lead to quality rivaling large-scale systems. This work offers a scalable and accessible foundation for Bengali voice cloning, with promising implications for personalized speech synthesis, assistive technologies, and natural human-computer interaction in the Bengali-speaking community.

Acknowledgments

We would like to sincerely thank our respected teachers in the Department of Computer Science and Engineering (CSE) for their constant support throughout this project. Special thanks go to our supervisor, Dr. Mohammad Shahidur Rahman, Professor at the Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet. We are very grateful for his guidance and for providing us with all the resources we needed to complete our research.

We are also inspired by the brave freedom fighters of Bangladesh, whose sacrifices encourage us to help advance technology in our language. Finally, we thank the Department of Computer Science and Engineering for their help, as well as everyone else who contributed to this project.

Table of Contents

List Of Figures

List Of Tables

Chapter 1

1 Introduction

Voice cloning, a fascinating part of speech synthesis, has gained attention for its innovative applications. The main goal is to create realistic voices using only a small sample of audio. Traditional concatenative methods were replaced in 2016 by deep learning models, which produce more natural-sounding speech [1]. Since then, researchers have worked to improve these models, making speech sound even more realistic and allowing end-to-end training.

Over the years, voice cloning technology has improved significantly. Earlier models were extremely slow, taking much longer than real-time to generate speech, even on powerful GPUs [2]. However, by 2018, real-time speech synthesis became possible on mobile devices [3]. Some studies have even shown that AI-generated speech is nearly indistinguishable from human voices [4], with some researchers claiming it may have already surpassed natural human speech [5].

Voice cloning is widely used in areas such as animated movies, deep fake videos, advertising, and multilingual speech translation while maintaining the speaker's identity. It also improves human-computer interactions by enabling personalized voice assistants. Additionally, it offers a voice to nonverbal individuals, giving them a way to communicate [6]. To make these applications effective, it is crucial to control speech characteristics like tone, speed, emphasis, and emotions.

Traditionally, voice cloning has been defined as recreating a person's voice using recorded audio. This is usually done by training a multi-speaker text-to-speech (TTS) model, which extracts speaker-specific traits from voice samples. However, while this method captures a speaker's identity well, it struggles to provide finer control over tone, speaking speed, and emotional expression. Researchers have worked on training deep learning models to learn different speech styles, but these models are limited by the voices in their training datasets and cannot easily generate voices they haven't encountered before [7].

One major challenge in voice cloning is data efficiency. High-quality speech datasets are rare, and to produce realistic voices with clear pronunciation and natural intonation, massive amounts of training data are required. Models like Tacotron need hundreds of hours of recorded speech to generate high-quality output [8]. Researchers have explored frameworks for voice cloning and voice conversion. Voice cloning replicates a speaker's voice for text-to-speech generation, while voice conversion changes an existing speech recording to sound like another person's voice.

In our research, we have trained a voice-cloning model that can fine-tune speech characteristics with high precision. We use a specialized model called VITS, which enhances voice synthesis by capturing the subtle details of human speech. Our experiments show that this approach

significantly improves expressiveness and can mimic voices it has never heard before. Essentially, this technology gives AI a more natural and human-like voice, making speech synthesis more lifelike than ever before.

1.1 Motivation

Technology is constantly evolving, bringing new advancements and increasing the demand for voice cloning systems in various fields such as education, entertainment, customer service, and accessibility tools. However, compared to languages like English, high-quality text-to-speech (TTS) models for Bengali are still very limited. This has created a growing need for a Bengali voice-cloning model that can produce natural and expressive speech.

Voice cloning is especially helpful for visually impaired and nonverbal individuals, allowing them to communicate more easily through synthesized speech. A Bengali voice cloning system can benefit Bengali speakers in many ways. It can create personalized learning experiences for students, generate realistic virtual characters for games, animated films, and virtual reality, and improve interactive voice assistants. This technology also enhances user-friendly digital experiences by making interactions more natural and engaging.

1.2 Objective

The main goal of this project is to create a Bengali zero-shot voice cloning model that allows computers to speak naturally in Bengali, making it easier for Bengali speakers to interact with technology. While there are many voice cloning models for English, there are very few for Bengali. This project aims to bridge that gap by developing a model that can accurately read Bengali text and mimic the voices of speakers it has never heard before. This will help create a more inclusive and expressive digital communication experience for Bengali speakers.

Overall, a Bengali voice cloning model can greatly support the Bengali-speaking community, making digital communication more accessible and personalized.

Chapter 2

2 Background Study

Voice cloning is a technology that enables the generation of synthetic speech using a small audio sample from a new speaker. It has gained significant attention due to its potential applications in areas such as virtual assistants, audiobooks, and personalized speech synthesis. There are two primary voice cloning approaches: speaker encoder and speaker adaptation methods. The speaker adaptation approach involves fine-tuning a pre-trained multi-speaker text-to-speech (TTS) model using recordings of a new speaker's voice. However, when only a small amount of data is available, the model is prone to overfitting, meaning it may learn specific details too well and struggle to generalize across different sentences, speech styles, or contexts.

To address this issue, a more advanced technique involves the use of a separate speaker-encoding network. This network captures speaker-dependent information and allows for zero-shot voice cloning, meaning it can synthesize speech in a new voice without requiring transcriptions or extensive retraining. By operating directly on waveform-based features, the speaker-encoding network extracts unique characteristics of the speaker's voice and applies them to speech synthesis. This approach enhances the flexibility and scalability of voice cloning, making it possible to generate realistic speech for multiple speakers with minimal training data.

Despite advancements in multi-speaker voice cloning, several challenges remain, particularly in terms of language support. Many existing models are designed primarily for widely spoken languages such as English, leaving limited resources for languages like Bangla. The lack of diverse linguistic datasets and specialized models for Bangla makes achieving high-quality voice synthesis in this language difficult. As research in voice cloning continues to evolve, efforts are being made to improve language inclusivity and enhance the performance of speech synthesis models across a wider range of languages.

2.1 Literature review

Our primary goal was to develop a zero-shot voice cloning model capable of cloning a target speaker's voice in Bangala using a short audio sample during inference. The model aims to achieve cloning results comparable to state-of-the-art models.

Text-to-Speech: Text-to-speech (TTS) technology converts written text into utterance through a series of steps. First, it takes text input and processes it through normalization and linguistic

analysis to break it down into phonemes and prosodic features. Next, an encoder transforms the processed text into a latent representation, which is then used to generate a Mel-spectrogram via a Variational Autoencoder (VAE). Subsequently, a Generative Adversarial Network (GAN) enhances the quality and naturalness of the Mel-spectrogram. Finally, the HiFi-GAN [9] vocoder converts the Mel-spectrogram into a high-quality speech waveform.

VITS (Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech) [10] exemplifies this process. It's an end-to-end TTS system that directly generates expressive speech waveforms from text using a combination of GlowTTS[11] encoder and Hifi-GAN vocoder architectures. VITS leverages techniques like GANs, VAE, and normalizing flows, to learn alignment between text and audio through Mel-Spectrogram Audio Similarity (MAS) without requiring external alignment annotations.

Zero-shot generative modeling: Zero-shot generative modeling refers to a model's capacity to produce high-quality personalized speech using only a short reference audio of an unseen speaker as an acoustic prompt. The paper [12] introduced an approach called the Speaker Encoder Approach. This process generally utilizes an encoder which produces a speaker's latent vector for each frame of the spectrogram by processing the mel-spectrogram of a speaker’s utterances. The resulting embedding vectors, Multi-Scale Speaker (MSS) vectors, intend to outperform state-of-the-art speaker embeddings regarding speech naturalness and speaker similarity. These MSS vectors capture the speaker's distinct vocal characteristics, allowing text-to-speech (TTS) models like XTTS [13] or VALL-E [14] to produce synthesized speech in the target speaker's voice. The TTS model can produce speech with pitch, tone, and timbre similar to the reference speaker by conditioning the synthesis process on the extracted speaker embeddings. This creates personalized speech synthesis without the need for extensive speaker-specific training data [15].

OpenVoice is an adaptable instant voice cloning method that requires only a brief audio clip from the reference speaker to produce an unseen speaker's high-quality, personalized speech in multiple languages [16]. It allows granular control over vocal characteristics, such as emotion, accent, rhythm, pauses, and intonation, while preserving the reference speaker's tone quality. This is achieved by utilizing a base speaker text-to-speech (TTS) model to control style parameters and languages, along with a tone color converter to integrate the reference speaker's tone quality into the synthesized voice.

Unlike previous methods, OpenVoice does not rely on an extensive massive-speaker multi-lingual (MSML) dataset to clone voices in new languages. Instead, this model achieves zero-shot cross-lingual voice cloning, which enables voice replication in languages not included in the training dataset. OpenVoice exemplifies the SC approach.

Chapter 3

3 Dataset Preparation

Data Source: For our model training, we used two publicly available datasets, SUBAK.KO [17] and the development of an annotated Bangla speech corpora dataset [18] to train the multi-speaker TTS model.

Both datasets contain recordings of native Bengali speakers reading text fluently. The recordings were collected in a controlled environment to eliminate background noise, ensuring clear and high-quality speech. The pronunciation in these recordings is accurate, making them suitable for training a voice cloning model.

We used over 226 hours of short audio clips, each lasting between 10 and 15 seconds. After gathering the data, we carefully labeled and categorized it to prepare it for training. These high-quality recordings played a crucial role in improving the performance of our Bengali voice cloning model.

Raw Dataset: From the SUBAK.KO dataset, we selected 59 unique speakers, with all audio recordings sampled at 16 kHz. Additionally, from the Development of Annotated Bangla Speech Corpora dataset, we utilized recordings from one speaker, comprising approximately 4 hours of data. The processed is similar to the VCTK-old [19] format.

Table 1: Training Audio Duration per Speaker (SUBAK.KO Dataset)

Speaker	Total Duration(hours)
s1	13:33:00
s2	12:05:24
s3	10:48:36
s4	9:23:24
s5	9:21:00
s6	8:43:12
s7	8:37:48
s8	7:11:24
s9	6:57:00
s10	6:56:24
s11	6:45:36
s12	6:42:00

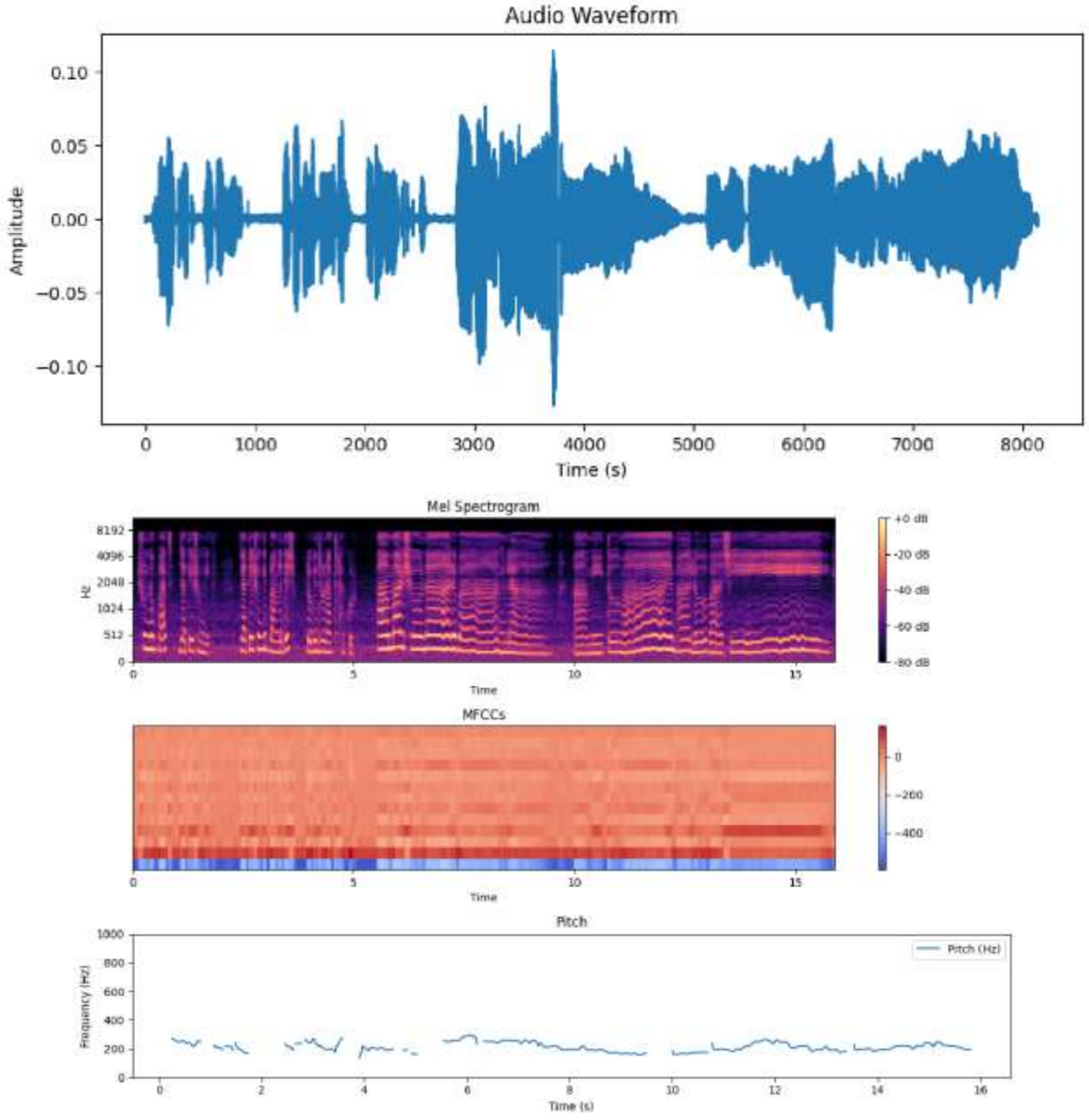
s13	6:36:36
s14	6:33:00
s15	6:33:00
s16	6:00:00
s17	5:19:12
s18	5:01:48
s19	5:01:48
s20	4:43:48
s21	4:10:47
s22	3:55:12
s23	3:48:00
s24	3:43:48
s25	3:19:48
s26	3:18:00
s27	3:12:00
s28	3:07:12
s29	3:06:00
s30	2:54:36
s31	2:54:00
s32	2:39:36
s33	2:33:00
s34	2:21:36
s35	2:00:35
s36	1:57:36
s37	1:54:00
s38	1:43:48
s39	1:43:12
s40	1:42:00
s41	1:28:48
s42	1:20:24
s43	1:20:24

s44	1:17:24
s45	1:16:12
s46	1:14:24
s47	1:11:24
s48	1:04:48
s49	0:58:48
s50	0:55:12
s51	0:41:24
s52	0:41:24
s53	0:41:24
s54	0:36:36
s55	0:36:36
s56	0:36:36
s57	0:32:24
s58	0:22:12
s59	0:21:00
ALL	226:16:12

Table 2: Training Audio Duration of Single Speaker (Development of Annotated Bangla Speech Corpora dataset)

Speaker	Total Duration(hours)
s1	4:02:00

Sample Audio: A visual representation of a sample audio file is given below for a better understanding of the distribution and characteristics of the dataset:



VCTK Dataset: We pre-processed a custom Bengali audio dataset in the VCTK dataset [19] format.

Audio Files

We convert our audio file into FLAC format. The sample rate of those audio files is 16 kHz, and those are situated inside the speaker sub-directories. For the VCTK format, we had to maintain the naming convention as [base]_mic. files

```
Dataset_Directory/
|-- wav48_silence_trimmed/
|   |-- Speaker_Name_Subdirectory/
|   |   |-- [base]_mic1.flac
|   |   |-- ...
|   |-- Another_Speaker_Subdirectory/
|       |-- [base]_mic1.flac
|       |-- ...
|-- ...
```

Transcripts

The transcripts in our dataset are in txt format. It is structured in the same way as those audio files is. They are placed in the speakers sub-directory under the txt parent directory. The naming convention is [base].txt

```
Dataset_Directory/
|-- txt/
|   |-- Speaker_Name_Subdirectory/
|   |   |-- [base].txt
|   |   |-- ...
|   |-- Another_Speaker_Subdirectory/
|       |-- [base].txt
|       |-- ...
|-- ...
```

Chapter 4

4 Experiments

4.1 Model Selection

Our primary goal was to develop a zero-shot neural voice-cloning model for Bengali. To achieve this, we aimed to identify the best multi-speaker model available, along with an effective speaker encoder model.

We began by exploring top open-source voice cloning tools across different languages. Our testing included various zero-shot voice cloning models, such as YourTTS, VALL-E, GlowTTS, XTTS, and VITS. For these initial experiments, we used English text to evaluate the models' performance. From our tests, we observed that VALL-E and Bark produced excellent results for English voice cloning. However, none of these models had built-in support for Bengali.

To develop a Bengali neural voice-cloning model, we conducted multiple experiments to find the most effective approach. Since Bengali has unique phonemes and accents, we needed a model that could handle them naturally. We focused on multilingual open-source models, as they offered better flexibility for incorporating Bengali speech synthesis. Among the models we examined, VITS and GlowTTS stood out due to their multilingual capabilities and straightforward configurations. These characteristics made them the most suitable options for the next phase of our research.

4.2 Transition Method for Adopting the Bengali Frontier

The multilingual models we selected could not understand Bengali, which created a major challenge in developing a Bengali voice cloning model. To solve this problem, we needed to modify the encoder of our model.

In the VITS model, the encoder functions like a text-to-speech system. The model originally used GlowTTS as its encoder, which converts Bengali text into a form that the machine can process. However, GlowTTS has a built-in encoder and vocoder that cannot be modified or replaced with an external vocoder.

On the other hand, the VITS (YourTTS) model offers a more flexible architecture, allowing us to add a custom encoder and vocoder. This modification enabled us to handle Bengali characters

properly and improve speech synthesis for Bengali, making the model more effective for voice cloning.

We integrated the bn-phonemizer from the Coqui-AI/TTS repository into our model. We wrapped our synthesizer with this phonemizer and created a custom synthesizer capable of accurately processing unique Bengali characters. This change significantly improved the model's inference quality, making the speech output sound more natural and clearer. Additionally, we incorporated the HiFi-GAN vocoder, which further enhanced the efficiency and overall quality of speech synthesis.

At this stage, our model was finally able to read and understand Bengali properly. While both VITS and GlowTTS supported multiple languages, we found that GlowTTS struggled with producing natural-sounding Bengali speech. It did not perform as well as VITS in terms of fluency and pronunciation.

After applying our custom configuration and synthesizer to the VITS architecture, we achieved more realistic and natural Bengali speech output. As a result, we decided to stop using GlowTTS for further experiments. With these improvements, our model can now accurately read and generate Bengali speech, making it sound more authentic and lifelike.

4.3 Exploration of Voice Conversion Models

We also explored several voice conversions models for generating Bengali speech, including OpenVoice and FreeVC. These models proved beneficial by significantly reducing training time, as they are language-independent and do not require fine-tuning or pretraining on large datasets. While these models are designed for speech-to-speech conversion, our goal was to generate speech in the voice of a target speaker from text input.

To achieve this, we utilized a text-to-speech (TTS) model trained on a single speaker. The audio produced by this model contained the contextual information necessary for generating speech in the target speaker's voice. This approach laid the foundation for developing hybrid models that combine the strengths of TTS and VC systems, improving both performance and output quality.

4.4 Initial Breakthrough and Limitations

In our early work, we trained the Speaker Encoder (SE) using data from 12 speakers, which laid the foundation for the first-ever Bengali neural voice cloning system. This work was a significant contribution to the field and was published in ICCIT (IEEE) [20] . However, one major limitation of this early approach was its lack of robustness when the reference speaker's voice was highly different from those seen during training. The model's performance dropped when faced with greater variability in unseen voices.

To address this limitation, we expanded the training dataset significantly. We selected voice data from 59 diverse speakers to better cover a wide range of acoustic and stylistic variations. This improvement allowed our model to generalize more effectively to unseen speakers, improving robustness and voice similarity across different reference inputs.

4.5 Developing a Multi-Speaker Voice Cloning Model

To create a successful Bengali voice cloning model, we needed a multi-speaker generative model that could support the Bengali language.

First, we adjusted the configuration of the single speaker model to ensure it could understand Bengali characters. Then, we added a phonemizer in the inference phase to improve speech synthesis.

For voice cloning, we used a speaker encoder approach. This method allows the model to create a fixed-size speaker embedding vector from a short reference speech sample provided by the speaker. To do this, we introduced a concept called d_vector, which is a type of deep speaker embedding. This is a neural network that learns the unique features of a speaker's voice from the training data, helping the model recognize and replicate different voices.

For training, we used our Bengali audio dataset, which we had prepared in VCTK format during the data processing phase. The dataset contained recordings from 59 native Bengali speakers. We trained the model using audio from 59 speakers, while we chose 2 speakers from the internet, used for testing.

4.6 Training a TTS model

For training our TTS model, we used approximately 4 hours of annotated Bengali speech data from a single speaker, sourced from the Development of Annotated Bangla Speech Corpora dataset. To convert text into speech, we employed the VITS model in a single-speaker configuration.

4.7 Model Evaluation

After training, we tested the model in two ways:

1. Training Speaker Test: We provided the model with a reference voice from one of the speakers used during training. The model successfully mimics the speaker's voice, producing accurate results.
2. Unseen Speaker Test: We tested the model with a reference voice from a new speaker who was not included in the training data. Even with just a few seconds of reference audio, the model was able to clone the new speaker's voice accurately.

Both tests were conducted using Bengali text and speech, and the model was able to generate fluent Bengali speech that matched the speaker's voice.

By carefully following these steps, we successfully created a zero-shot voice cloning model capable of cloning the voices of both known and unknown speakers in Bengali.

Chapter 5

5 Methodology

This research aims to develop a zero-shot Bengali voice cloning model that can synthesize speech in a specific speaker's voice using only a short reference audio sample. Unlike traditional text-to-speech (TTS) models, which require large amounts of data from a speaker to generate realistic speech, zero-shot voice cloning enables the synthesis of a speaker's voice without prior training on their speech data.

To achieve this, we explored two different methodologies:

1. The Speaker Encoder (SE) Approach, which generates speech using a multi-speaker TTS model and speaker embeddings.

2. The Speaker Converter (SC) Approach, which first synthesizes speech using a single-speaker TTS model and then modifies it to match the target speaker's voice.

3. Additionally, we proposed a hybrid approach that combines both SE and SC models to improve the overall quality of Bengali voice cloning.

This section outlines the datasets used, preprocessing steps, model architectures, and evaluation metrics applied in this research.

5.1 Speaker Encoder (SE) Approach

The SE approach focuses on generating Bengali speech from text by leveraging a multi-speaker TTS model and a speaker encoder. This method allows the model to clone a speaker's voice from a short reference audio sample.

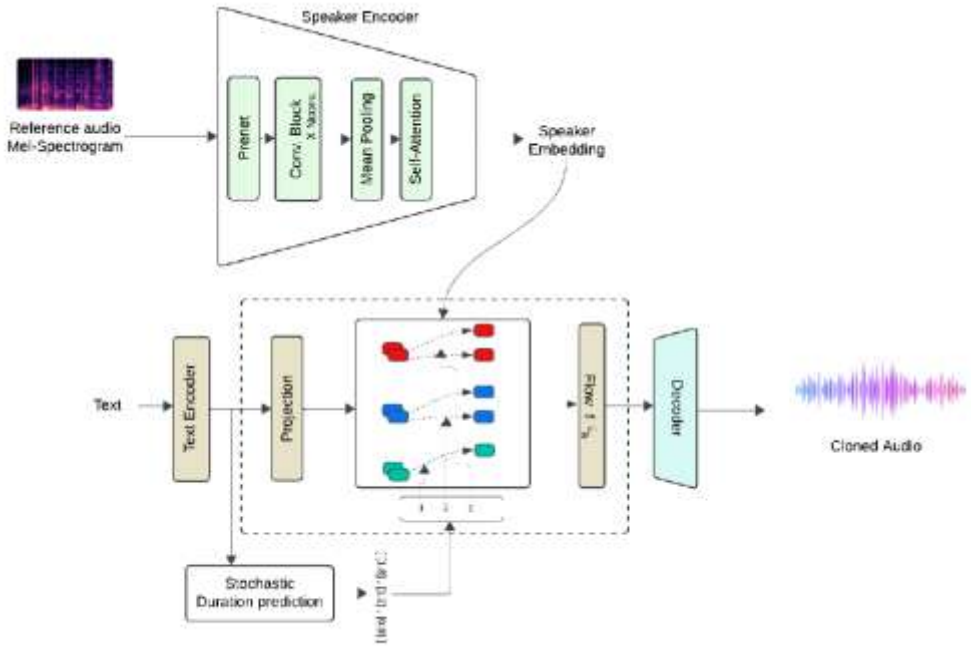


Figure 1: Architecture for SE approach

1. Extracting Speaker Embeddings

- A speaker encoder is trained to extract a speaker embedding vector from a short reference audio sample.
- This embedding captures the unique voice characteristics of the speaker, including pitch, tone, and accent.

2. Generating Speech from Text

- The speaker embedding vector is combined with text input, allowing the model to generate speech that mimics the target speaker's voice.
- A synthesizer converts the combined input into a mel-spectrogram, which visually represents the sound.

3. Converting the Mel-Spectrogram into Speech

- A HiFi-GAN vocoder is used to transform the mel-spectrogram into natural-sounding speech.

4. Loss Function

- The loss function is based on a distance metric (such as mean squared error or L2 norm) to evaluate how closely the generated speech matches the original reference speech.

The SE approach effectively enables the model to synthesize speech in an unseen speaker's voice, making it a key component of zero-shot voice cloning.

5.2 Speaker Converter (SC) Approach:

The SC approach follows a different method by converting synthesized speech into a target speaker's voice.

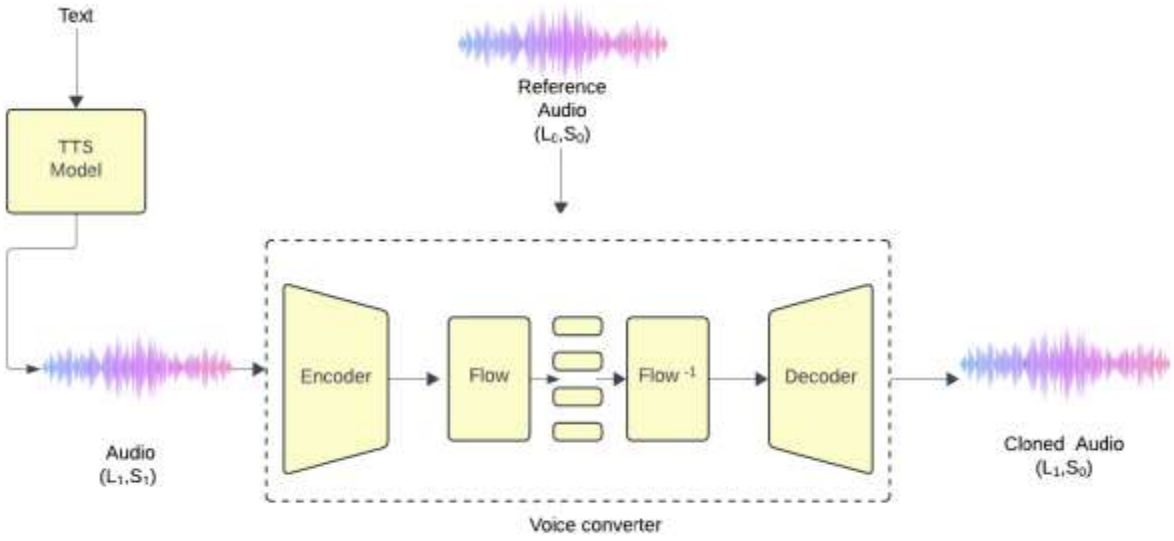


Figure 2: Architecture for SC approach

1. Generating Synthetic Speech

- Our trained single-speaker TTS model first synthesizes Bengali speech from text. However, voice conversion model depends on the base audio data. That means the voice conversion model highly depends on the first single-speaker TTS.

2. Applying Voice Conversion

- The synthesized speech or base audio is then processed through a voice conversion model, which modifies it to match a reference speaker's voice.
- The pretrained OpenVoice[21] model was used for this task. Since OpenVoice is language-independent, it could be used for Bengali voice cloning without extra fine-tuning.

- Besides the OpenVoice model, we recently tried the FreeVC [22] voice conversion model, which is an end-to-end framework of VITS for high-quality waveform reconstruction, and for the experiment, we used FreeVC from Coqui-AI [23].

This approach was highly effective in cloning voices for unseen speakers using only a short reference audio sample.

5.3 Proposed Hybrid Approach:

To enhance the quality of Bengali voice cloning, we proposed an Ensemble Model(EM) approach that integrates SE and SC models.

1. Step 1: SE Model Generates Cloned Speech

- The SE model first generates cloned speech from text and a reference speaker's voice.

2. Step 2: SC Model Refines the Output

- The SC model's voice converter then further processes the speech, enhancing clarity, fluency, and speaker similarity.

By combining both approaches, we achieved more accurate and natural Bengali speech synthesis, making the cloned voice more realistic and expressive.

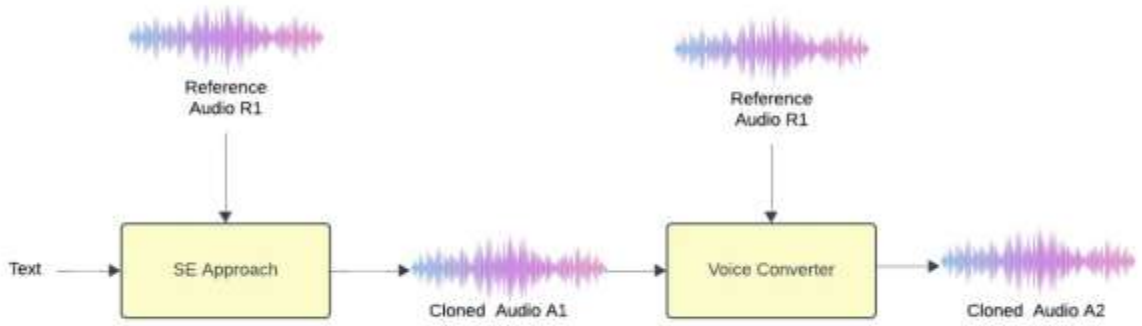


Figure 3: Proposed Architecture

5.4 Evaluation Metrics

To measure the effectiveness of our Bengali voice cloning models, we used two key evaluation metrics:

1. **Mean Opinion Score (MOS)**

   To evaluate the naturalness and intelligibility of the cloned speech, we conducted a Mean Opinion Score (MOS) test, where human listeners rated each audio sample on a scale from 1 to 5. The MOS score reflects how human-like the generated speech sounds, considering both clarity and resemblance to natural speech.To ensure fairness, all audio samples were amplitude-normalized to prevent volume differences from influencing the ratings. Each audio clip was evaluated only once by each rater.

   The evaluation was conducted by seven final-year students from Sylhet Engineering College, who were carefully trained to assess the quality of the cloned audio in comparison to its reference. During the training phase, students were familiarized with samples from existing state-of-the-art models to calibrate their understanding of speech quality and speaker similarity. Each listener evaluated only one cloned audio sample along with its corresponding reference audio for each model.

   Following the training, raters evaluated audio outputs from our models. Their task was to assess not only the naturalness of the speech but also how closely the cloned voice

matched the reference voice in tone and character. This structured and consistent evaluation approach allowed us to compute reliable MOS scores for comparing model performance.

where N is the number of listeners, and R_{tkl} is the rating given by listener k for sample l. The overall MOS is then computed as:

$$MOS_l = \frac{1}{N} \sum_{k=1}^{N} Rt_{kl}$$

$$MOS = \frac{1}{S} \sum_{l=1}^{S} MOS_l$$

where S is the total number of samples.

2. **Cosine Similarity**

This metric evaluates how closely the speaker identity is preserved by comparing the embeddings of the original and cloned voices.To evaluate how well the speaker identity is preserved in the cloned audio, we used Resemblyzer, a pre-trained speaker encoder that extracts high-dimensional speaker embeddings from audio samples. These embeddings capture various speaker-specific characteristics such as tone, pitch, speaking style. We extracted embeddings from both audio samples, then computed the cosine similarity between them. Higher similarity scores (closer to 1) indicate better identity preservation, meaning the cloned voice more accurately reflects the original speaker's characteristics.

Chapter 6

6 Results and Discussions

We developed a Bengali voice cloning model that can generate natural-sounding speech from text. Our goal was to create a system that could accurately mimic a speaker's voice while maintaining the unique characteristics of the Bengali language. To achieve this, we ensured that our model fully supports Bengali characters, capturing the language's phonetic and linguistic features.

One of the key improvements we made was adding a phonemizer to the encoder. This helped the model understand Bengali characters better, resulting in smoother and more natural text-to-speech conversion. The addition of the phonemizer improved pronunciation, making the speech sound more fluent and expressive instead of robotic.

Figure 4: Voice Cloning Process

To evaluate the performance, we tested our models with reference audio from speakers outside the training dataset to assess how well they could generalize to unseen voices. The model produced clear and fluent speech that closely matched the speaker's tone, speaking rate, and emotional nuance. It successfully captured the natural rhythm of Bengali speech, contributing to a more human-like output.

In addition to qualitative evaluation, we compared the generated audio with outputs from several existing voice cloning models, including XTTS, VALL-E, and YourTTS (our base model**)**. We categorized the models as follows:

- SE and SC are our proposed standalone models.
- YourTTS served as our baseline.
- OpenVoice and FreeVC were treated as external SC modules and used to build Ensemble models by combining them with our SE model. Specifically:
    - Ensemble (SE + OpenVoice): used OpenVoice as the SC component.

- Ensemble (SE + FreeVC**)**: used FreeVC as the SC component.

These ensemble configurations allowed us to test hybrid approaches for better speaker similarity. Across the board, our models outperformed YourTTS in terms of both subjective listening quality and cosine similarity scores. The higher cosine similarity between reference voices and generated speech in our models indicates superior speaker identity preservation.

These results highlight that our voice cloning system can produce high-fidelity Bengali speech while maintaining accurate voice characteristics, even with minimal reference data.

Table 3: MOS and Cosine Similarity Scores for Different TTS Approaches

Experimental Setup	Language	Approach	MOS	Cosine Similarity
Our Previous Works With 12 Speakers	Bengali	SE SC (OpenVoice) Ensemble Model (OpenVoice)	3.60 ± 0.11 3.80 ± 0.11 4.02 ± 0.11	73.91% 76.84% 78.34%
Our Current Works with 59 Speakers	**Bengali**	**SE** **SC (OpenVoice)** **SC (FreeVC)** **Ensemble Model (FreeVC)** **Ensemble Model (OpenVoice)**	**4.21 ± 0.11** **4.23 ± 0.11** **4.14 ± 0.11** **4.34 ± 0.11** **4.37 ± 0.11**	**84.13%** **85.20%** **85.85%** **80.83%** **86.11%**
Existing models	English	XTTS (Over 2000 speakers) [24] Your TTS (Over 2,500 speakers) [25] VALL E (Over 7000 speakers) [26]	4.11 ± 0.11 3.72 ± 0.11 4.70 ± 0.11	85.02% 75.23% 89.87%

As evident from the results, our base model YourTTS for SE, after being fine-tuned with Bengali language data, shows a significant improvement in performance. The achieved cosine similarity closely approaches those of state-of-the-art models like VALL-E and XTTS, despite those models being trained on significantly larger and more diverse speech datasets. This highlights the effectiveness of our adaptation strategy even with limited training data.

In our previous work with only 12 speakers, the Ensemble Model (OpenVoice) achieved the highest cosine similarity (78.34%) among all approaches. Building upon this, our current work with 59 speakers shows further improvements across all methods. Notably, the Ensemble Models, both with FreeVC and OpenVoice, now reach cosine similarities of 80.83% and 86.11%,

respectively. These results suggest that increasing the number of training speakers leads to more robust voice representation and higher accuracy.

We believe that with extended training time and further optimization, the ensemble models in our current setup with 59 speakers will surpass their current performance and potentially exceed the results from models trained on much larger datasets. This trend aligns with our earlier findings, where the ensemble approach consistently delivered superior cosine similarity scores.

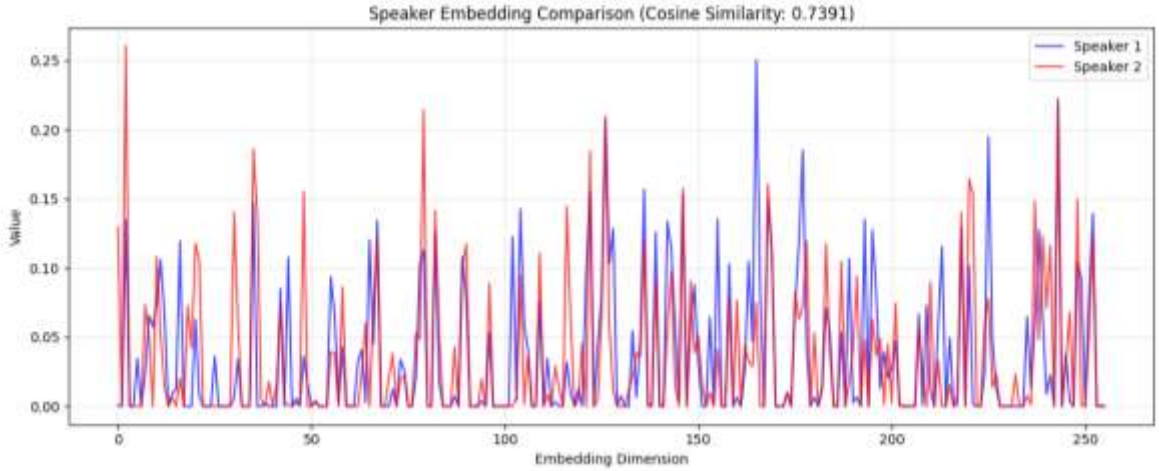


Figure 5: Speaker Embedding Comparison and Cosine Similarity of SE in 12 Speakers

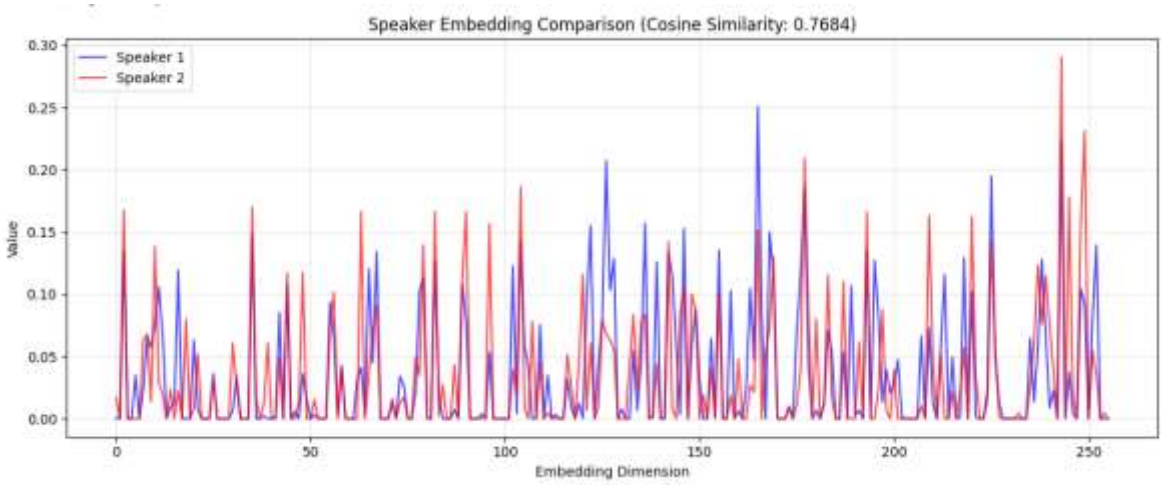


Figure 6: Speaker Embedding Comparison and Cosine Similarity of SC(OpenVoice) in 12 Speakers

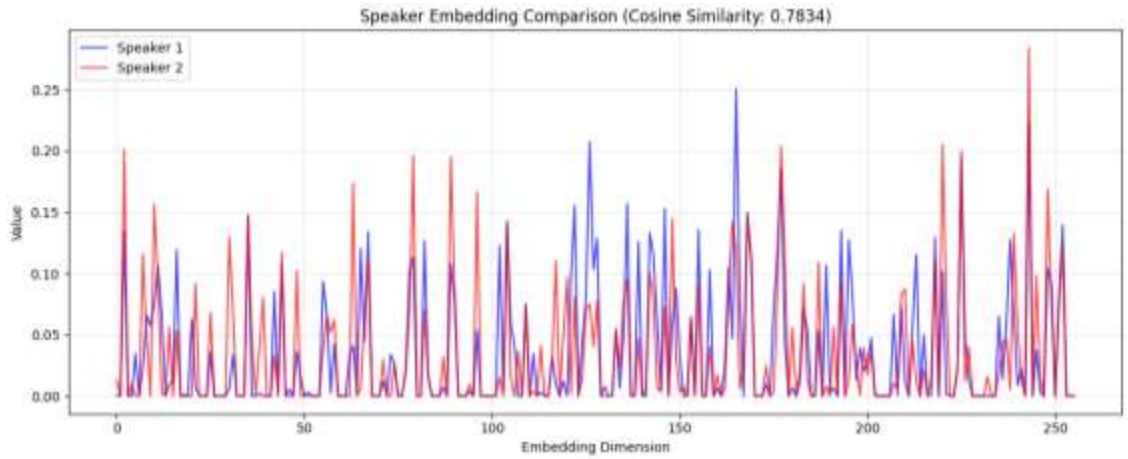


Figure 7: Speaker Embedding Comparison and Cosine Similarity of Ensemble Model (OpenVoice)

in 12 Speakers

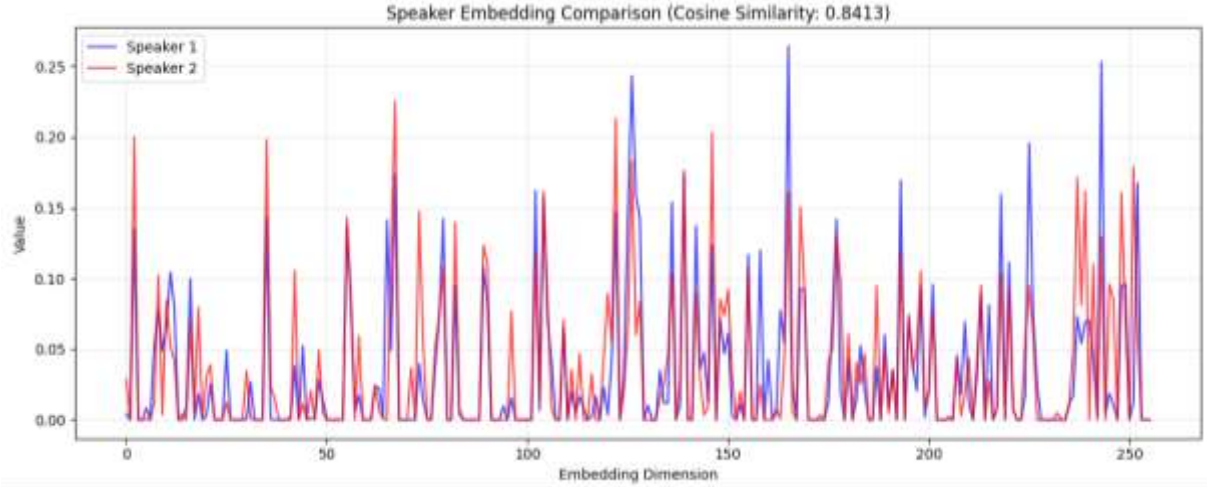


Figure 8: Speaker Embedding Comparison and Cosine Similarity of SE in 59 Speakers

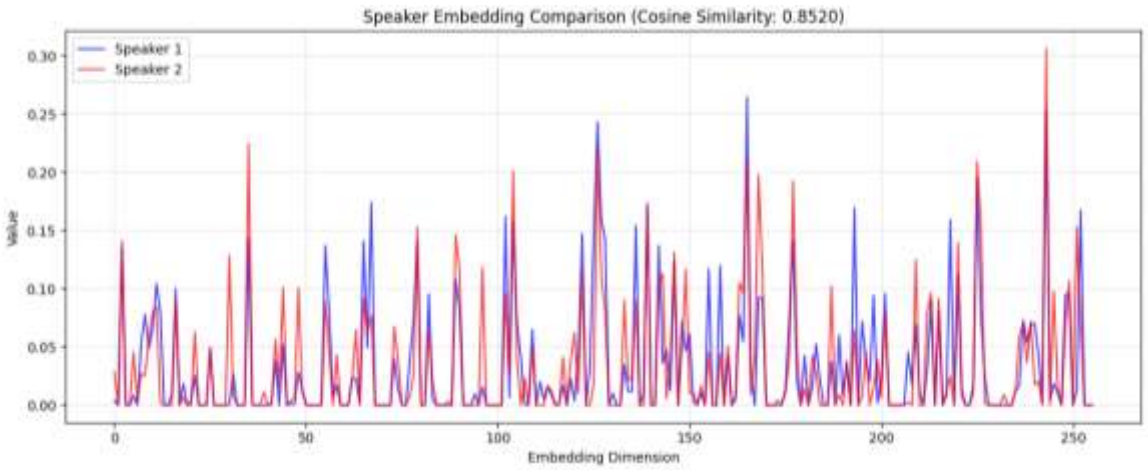


Figure 9: Speaker Embedding Comparison and Cosine Similarity of SC(Openvoice) in 59 Speakers

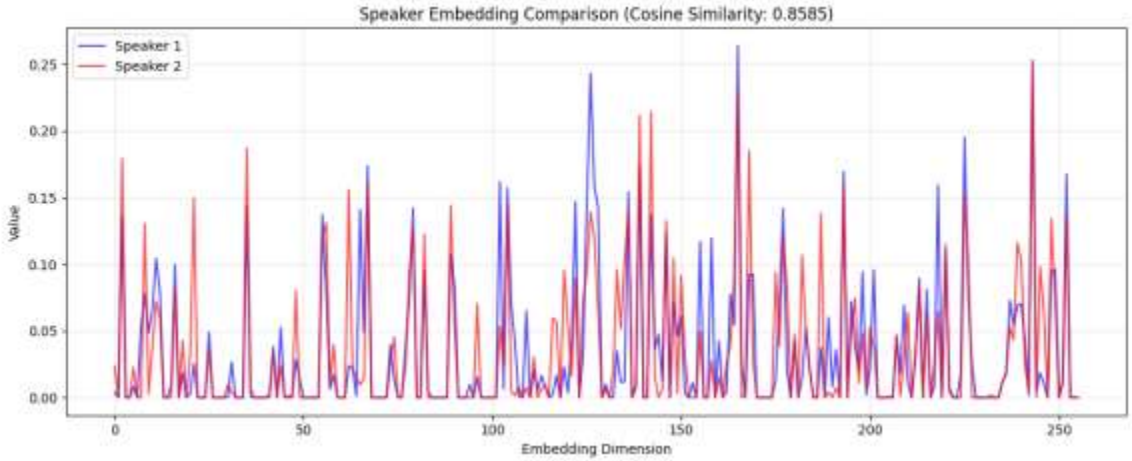


Figure 10: Speaker Embedding Comparison and Cosine Similarity of SC(FreeVC) in 59 Speakers

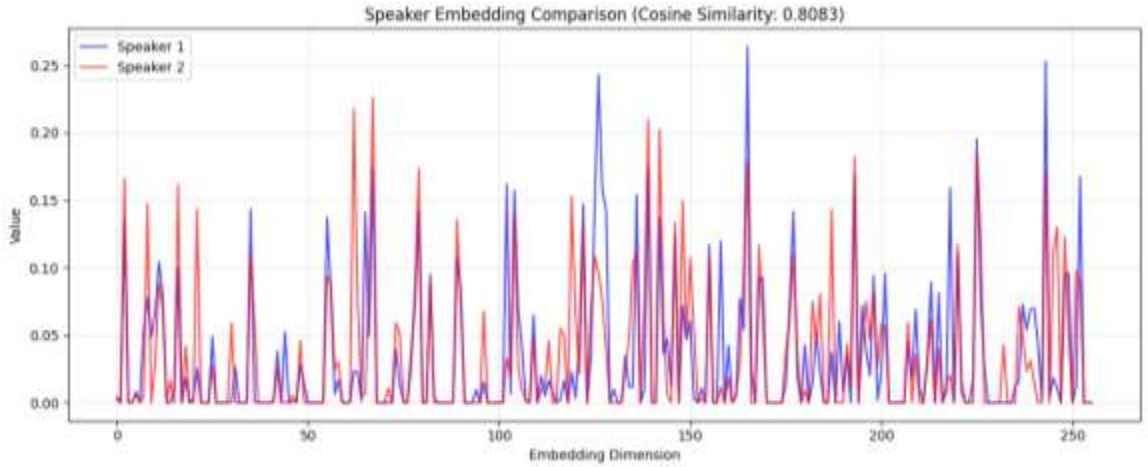


Figure 11: Speaker Embedding Comparison and Cosine Similarity of Ensemble Model (FreeVC) in 59 Speakers

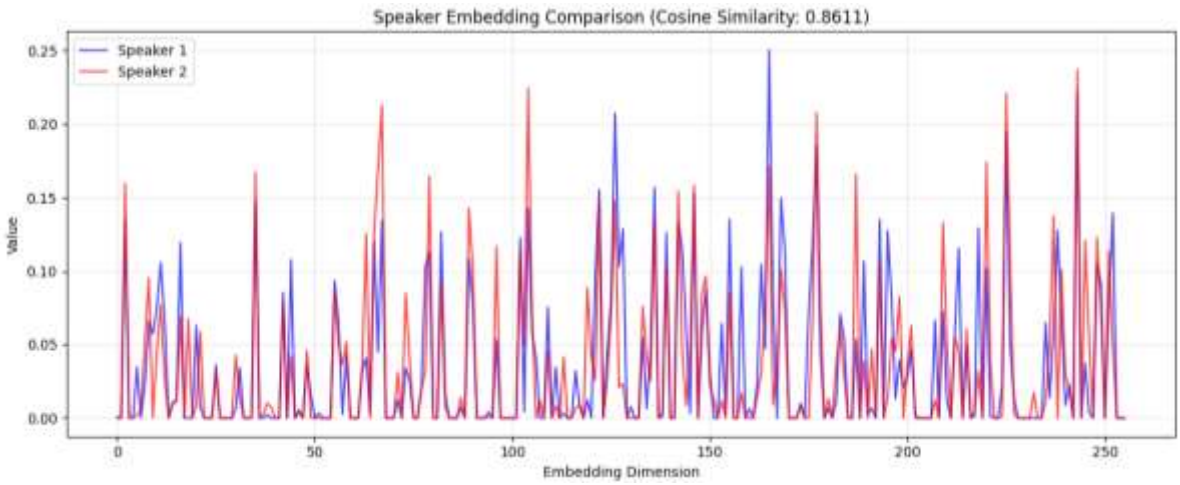


Figure 12: Speaker Embedding Comparison and Cosine Similarity of Ensemble Model (OpenVoice) in 59 Speakers

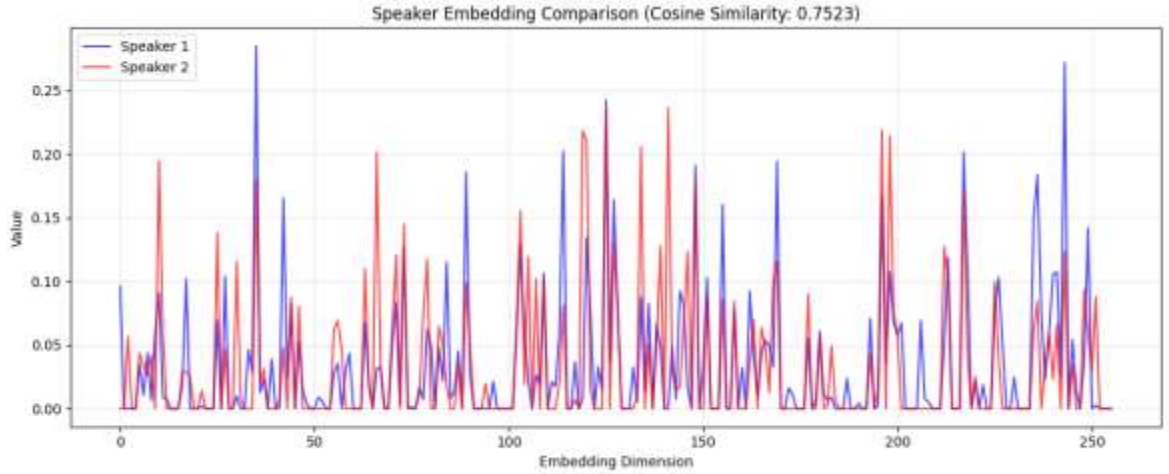


Figure 13: Speaker Embedding Comparison and Cosine Similarity of YourTTS model

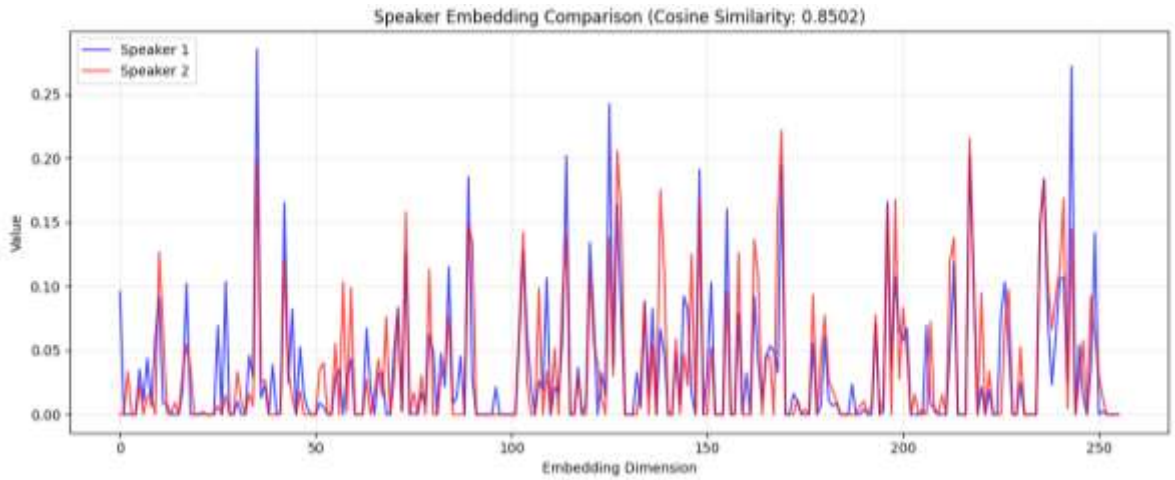


Figure 14: Speaker Embedding Comparison and Cosine Similarity of XTTS model

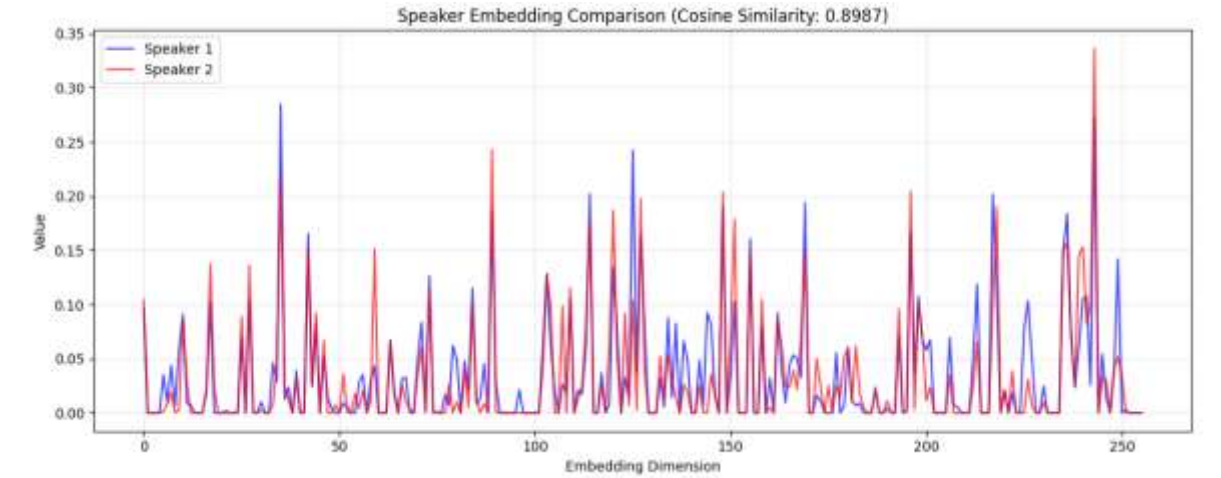


Figure 15: Speaker Embedding Comparison and Cosine Similarity of VALL E model

Chapter 7

7 Limitation

While developing our Bengali voice cloning model, we faced several challenges, which are listed below:

- Limited GPU Resources: Our available GPU power was not enough to handle large audio samples efficiently. Ideally, each audio clip should have been 6 to 10 seconds long, but our dataset contained clips around 15 seconds, which took up extra space and processing power.
- Long Training Time: Due to the limited GPU resources, we had to reduce the batch size during training, which slowed down the process and increased the overall training time.
- No Bengali Phonemizer: A phonemizer helps in converting text into accurate speech sounds. Since there was no dedicated Bengali phonemizer, we had to manually customize the character set to improve pronunciation. If we had a proper phonemizer, the model would have performed even better.
- Challenges in Model Evaluation (MOS Testing): To properly test the quality of our model, we needed a large number of people, time, and resources for Mean Opinion Score (MOS) testing. However, due to resource constraints, conducting extensive evaluations was difficult.

Despite these limitations, we managed to create a high-quality Bengali voice cloning model. With better datasets, resources, and tools, we can further improve its accuracy and performance in the future.

Chapter 8

8 Future Work

In this thesis, we have successfully developed a high-quality Bengali voice cloning model to generate Bengali speech from written text. The speech sounds natural and closely resembles the voice of the target speaker. In the future, we can further improve and expand the use of Bengali voice cloning technology.

Fine-tuning Voice Conversion Models: We will explore and fine-tune state-of-the-art voice conversion models like FreeVC and OpenVoice using custom datasets. This will enable more effective and flexible voice adaptation.

Improving Ensemble Models: We aim to develop an improved ensemble model to enhance flexibility, adaptability, and overall quality in voice cloning applications.

Eliminating Base Speaker Accent Influence: To achieve a more authentic voice cloning experience, we will work on removing the accent influence of the base speaker in our speaker conversion (SC) model, making the cloned voice fully dependent on the target audio.

Exploring Multiple Voice Conversion Models: We plan to investigate and experiment with multiple voice conversion models to identify the best approaches for high-quality and natural voice cloning.

Chapter 9

9 Conclusions

In this thesis, we proposed and developed a zero-shot Bengali voice cloning system that addresses the challenges of low-resource language support and speaker generalization in speech synthesis. Our goal was to build a flexible and high-quality Bengali voice cloning model capable of generating speaker-specific speech from just a few seconds of reference audio.

To achieve this, we designed two core models: a Speaker Encoder (SE) model trained on a large multi-speaker dataset to extract speaker embeddings, and a Speaker Converter (SC) model built using a single-speaker TTS system for high-quality Bengali speech generation. Recognizing the complementary strengths of these models, we introduced an ensemble framework that combines the SE-generated embeddings with VC models like FreeVC and OpenVoice, transforming TTS-generated intermediate speech into the target speaker's voice. This approach effectively bridges text-to-speech and voice conversion, enabling more accurate and expressive voice cloning.

We adapted existing multilingual models—specifically the VITS architecture—by integrating a Bengali phonemizer and vocoder to support the Bengali language. This allowed the system to handle Bengali characters and phonemes effectively, overcoming a major limitation in most pre-existing voice cloning frameworks.

Reference

1. [Arik et al., 2018] Arik, S., Chen, J., Peng, K., Ping, W., and Zhou, Y. (2018). Neural voice cloning with a few samples. *Advances in neural information processing systems*, 31.

2. [Awan, 2023] Awan, A. A. (2023). Bark: The ultimate audio generation model. Published on KDnuggets in Artificial Intelligence.

3. [Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *NeurIPS*.

4. [Chen et al., 2021] Chen, Q., Li, Y., Qi, Y., Zhou, J., Tan, M., and Wu, Q. (2021). V2c: Visual voice cloning. *arXiv preprint arXiv:2111.12890*.

5. [Jemine, 2019] Jemine, C. (2019). Real-time voice cloning. Master's thesis, Faculté des Sciences appliquées, Université de Liège. Defended on 26-Jun-2019/27 Jun-2019.

6. [Kalchbrenner et al., 2018] Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., van den Oord, A., Dieleman, S., and Kavukcuoglu, K. (2018). Efficient neural audio synthesis. *arXiv preprint arXiv:1802.08435*.

7. [Kibria et al., 2022] Kibria, S., Samin, A. M., Kobir, M. H., Rahman, M. S., 40 Selim, M. R., and Iqbal, M. Z. (2022). Bangladeshi bangla speech corpus for automatic speech recognition research. Speech Communication, 136:84–97

8. [Kim et al., 2020] Kim, J., Kim, S., Kong, J., and Yoon, S. (2020). Glow-tts: A generative flow for text-to-speech via monotonic alignment search. Advances in Neural Information Processing Systems, 33:8067–8077.

9. J. Kong, J. Kim, and J. Bae, 'HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis', *arXiv [cs.SD]*. 2020.

10. J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," ICML, 2021, comments: ICML 2021. [Online]. Available: https://doi.org/10.48550/ arXiv.2106.06103

11. Jaehyeon Kim. 2020. Glow-TTS. https://github.com/jaywalnut310/glowtts Accessed on May 10th, 2023.

12. S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, “Neural voice cloning with a few samples,” Advances in neural information processing systems, vol. 31, 2018.

13. CoquiAI, “Xtts taking text-to-speech to the next level,” Technical Blog, 2023, https://coqui.ai/blog/xtts.

14. C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, “Neural codec language models are zero-shot text to speech synthesizers,” arXiv preprint arXiv:2301.02111, 2023.

15. T. W. Cory, “Speaker encoding for zero-shot speech synthesis,” Master’s Thesis, Missouri State University, 2022. [Online]. Available: https://bearworks.missouristate.edu/theses/3796

16. Z. Qin, W. Zhao, X. Yu, and X. Sun, “Openvoice: Versatile instant voice cloning,” arXiv preprint arXiv:2312.01479v5, 2024. [Online]. Available:

https://arxiv.org/abs/2312.01479

17. S. Kibria, A. M. Samin, M. H. Kobir, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, ‘Bangladeshi Bangla speech corpus for automatic speech recognition research’, *Speech Communication*, vol. 136, pp. 84–97, 2022.

18. F. Alam, M. Habib, D. Sultana, and M. Khan, ‘Development of annotated Bangla speech corpora’, 09 2010.

19. C. Veaux, J. Yamagishi, K. MacDonald et al., “Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” 2016

20. A. Ahmed, P. Roy, D. Paul, and M. S. Rahman, ‘Efficient Zero-Shot Voice Cloning for Bengali Speech Synthesis’, 12 2024, pp. 1470–1474.

21. Qin, Z., Zhao, W., Yu, X., & Sun, X. (2024). OpenVoice: Versatile Instant Voice Cloning. *arXiv [Cs.SD]*. Retrieved from http://arxiv.org/abs/2312.01479

22. Li, Jingyi, Weiping Tu, and Li Xiao. "Freevc: Towards high-quality text-free one-shot voice conversion." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, 2023.

23. https://github.com/coqui-ai/TTS/tree/dev

24. E. Casanova *et al.*, ‘XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model’, *arXiv [eess.AS]*. 2024.

25. E. Casanova, J. Weber, C. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, ‘YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone’, *arXiv [cs.SD]*. 2023

26. C. Wang *et al.*, ‘Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers’, *arXiv [cs.CL]*. 2023