EasyChair Preprint

№ 15582

Leveraging Transfer Learning for Voice Cloning in Bengali Language

T Taruneshwaran, Sanjay Chidambaram, Parthvi Manoj,
S Sakthi Swaroopan and G Jyothish Lal

December 16, 2024

Leveraging Transfer Learning for Voice Cloning in Bengali Language

Taruneshwaran T, Sanjay Chidambaram, Parthvi Manoj, Sakthi Swaroopan S, Jyothish Lal G

Amrita School of Artificial Intelligence, Coimbatore,
Amrita Vishwa Vidyapeetham, India.

Contributing authors: cb.en.u4aie21170@cb.students.amrita.edu;
cb.en.u4aie21160@cb.students.amrita.edu;
cb.en.u4aie21143@cb.students.amrita.edu;
cb.en.u4aie21159@cb.students.amrita.edu;
g_jyothishlal@cb.amrita.edu;

Abstract

Voice cloning was an incredible innovation in the field of AI, which replaced human-machine interaction. Unlike other conventional voice synthesis methods, voice cloning requires the least amount of data in order to re-create the voice of a speaker and can offer personalized options for communication. But the creation of such small and powerful models working with scarce voice samples remains a big challenge. This is even challenging, especially for low-resource languages like Bengali, due to the scarcity of data itself apart from the intricacy of regional accents. Our study looks into voice cloning for Bengali using a transfer learning technique from speaker verification models. In this study we have adapted the model for Bengali using Mozilla Common Voice Bengali dataset with the SV2TTS framework. This dataset contains voices ranging in a wide variety of accents and dialects. Retraining the encoder, synthesizer, and vocoder components to capture the unique phonetic features specific to Bengali allows our approach to generate realistic, high-quality voice replications. It is evident from the results, as obtained by evaluation using the Mean Opinion Score method, that the cloned voices turn out very natural and similar in likeness to the speaker. These findings demonstrate prowess for under-resourced languages and extend into customized communications, voice acting, and speech-based assistive tools. This research is focused on the development of methods and models for Bengali speech processing to tackle challenges associated with low-resource language processing; further advances in Bengali speech technologies stand on such bedrock.

Keywords: AI-driven, Bengali, Speech synthesis, Text-to-speech, Voice cloning, Voice replication.

1 Introduction

The wide usage of virtual assistants like Alexa, Siri, and Google Home shows the extent of how artificial intelligence has been integrated into people's lives. But most of these AI-driven technologies have been designed for languages with rich linguistic resources. This leaves a lot of other languages, which have fewer computational tools, at a disadvantage and without the datasets needed to compete on equal footing. Bengali is one such language that is the second most spoken after Hindi in India. It also lacks proper representation in mainstream AI systems. The present study attempts to fill this gap by studying voice cloning technology designed for Bengali-speaking people. Within the many methods of generating speech, each holding unique characteristics of a human voice, lies voice cloning, which can bridge the gaps among languages. While large amounts of training data are required in traditional speech synthesis systems, the goal of voice cloning is to create high-quality, personalized voices with minimal data. In doing so, it opens up new horizons for poorly represented languages such as.

While Bengali is the seventh most spoken language in the world, it is virtually absent from the applications of modern AI. Our work goes beyond addressing this language barrier since it tries to understand the intrinsic linguistic and cultural traits which Bengalis possess and how they can strive toward transferring such nuances into advanced voice cloning systems and assistants. The process of developing voice cloning for underrepresented languages like Bengali faces many challenges for many reasons. Such a task, therefore, innately challenges innovative approaches, from the conditions of limited availability of good quality and annotated speech data to specific prosody and phonetic features typical of this language.

This work significantly contributes to the advancement of voice cloning for Bengali, a low-resource language. First, we adapt the SV2TTS framework, retraining its encoder, synthesizer, and vocoder components on the Mozilla Common Voice Bengali dataset to effectively capture the unique phonetic and prosodic features of Bengali speech. Second, we address the challenges of dialectal diversity and limited training data by leveraging transfer learning techniques and customized pre-processing for Bengali text. Finally, we test the system through both subjective (Mean Opinion Score) and objective (Mel-spectrogram similarity) metrics, thereby proving that our method indeed works in generating natural, speaker-specific synthesized voices. Our work contributes to low-resource language processing and sets a foundation for future advancements in inclusive AI-driven speech technologies.

Our work deploys machine learning and state-of-the-art natural language processing techniques in an effort to overcome such challenges and produce adaptable Bengali speech voice cloning models. This will inevitably pave the way for speech technologies driven by AI to go beyond mere communication and language learning to creating enabling tools for people with speech impairments. Giving a voice to Bengali speakers in the digital space, we see a way toward greater linguistic inclusion and contribute to furthering AI applications in culturally and linguistically relevant ways. This may be a first step toward broader AI inclusivity, and we

hope it will inspire new interest in further research and collaboration across the speech technology disciplines.

This paper develops voice cloning for Bengali, a low-resource language, using the SV2TTS framework and the Mozilla Common Voice Bengali dataset. Section 2 reviews related work, Section 3 describes the dataset, Section 4 outlines the methodology, and Section 5 covers the experimental setup and results. Section 6 discusses challenges, and Section 7 concludes with future work.

2 Literature Review

Speech synthesis in low-resource languages like Bengali holds immense significance related to aspects of language preservation, accessibility, and cultural inclusivity. Arik et al. [1] pioneered work on neural voice cloning using deep neural networks. Realistic voice synthesis from a few samples indeed showcased how neural architectures are adaptable across resource-scarce environments. On that basis, Dai et al. [2] addressed the sparse data problem in voice cloning and introduced ways to improve data efficiency, while calling for large-scale testing to confirm real-world practical efficacy.

Further development on the topic includes Li et al. [3], who developed Unet-tts, improving cloning for unseen speakers and style transfer, therefore achieving progress in natural-sounding synthesis. However, no application has been made to Bengali so far, which indicates that their effectiveness may need to be tried in low-resource languages. Another important contribution to neural audio synthesis, done by Kalchbrenner et al. [4], is WaveNet, a deep neural network architecture generating high-fidelity audio by predicting audio waveforms at the sample level. Although WaveNet was initially applied to other languages, its potential for Bengali synthesis remains unexplored; it also opens an avenue toward low-resource languages.

Magariños et al. [5] proposed an acoustic cloning approach, independent of the linguistic frontier and validated for English, which could be ported to Bengali. Their results point out the way synthesis methodologies may transcend language-specific features, making them even suitable for low-linguistic-resource languages. This idea aligns with the application of wav2vec2 for Bengali, as seen in recent work [6], which effectively extracts rich representations from unlabelled Bengali speech through self-supervised learning, opening the way for robust downstream tasks such as speech synthesis. Complementing these advances, a large Bengali speech recognition dataset for out-of-distribution benchmarking [7] provides diverse samples that are important to developing generalizable Bengali synthesis models. Early foundational work by Atal and Hanauer [8] in speech synthesis through linear prediction laid the groundwork for today's synthesis techniques and established methods that are still influential in speech technology today and find application in Bengali. Later, Zhang and Lin [9] extended this work with a multi-modal few-shot voice cloning approach, whereby data beyond audio can be utilized to enhance the synthesis quality of a voice—a methodology very

apt for use with the Bengali language due to its data limitations. Shen et al. [10] furthered this direction by conditioning WaveNet on mel-spectrogram predictions for natural text-to-speech synthesis, resulting in smoother and more expressive speech. Though this framework has not yet been applied to Bengali, it shows promise for adapting synthesis models across languages with complex sound patterns.

Other researchers have also extended the call to focus on robust methods for voice cloning under low-resource conditions. Zhao and Chen [11] presented cloning with minimal samples by addressing the challenge of sustaining speaker individuality in languages like Bengali. Their findings signal the need for flexible methods that uphold linguistic authenticity amid data constraints. Stevens et al. [12] used electrical analogs of the vocal tract to provide insights into speech production. Primarily, these works serve to model the physical aspects of speech to provide more realistic synthesis approaches and applications indirectly helpful in Bengali synthesis. Strube [13] discussed the Kelly–Lochbaum acoustic-tube model, contributing theoretical insights toward enhancing the accuracy and quality of synthetic Bengali speech through more classic physical modelling.

The British Council [14] identifies language preservation as essential to society, promoting research to make technology accessible for underrepresented languages like Bengali. Similarly, Crystal [15] applied a speech-to-component breakdown method to offer foundational methodologies for synthesis models tailored to the phonetic features present in Bengali. Meanwhile, Heigold et al. [16] developed an end-to-end text-dependent speaker verification approach, enhancing the preservation of individual speaker characteristics in synthesis—a feature particularly relevant for Bengali speaker-specific synthesis models.

Nair et al. [17] contributed a comprehensive survey on Indian text-to-speech systems, listing strategies and general obstacles related to regional languages, insights that are applicable to Bengali. Chowdary et al. [18] established the efficiency of Transformer-based ASR models for low-resource languages in their work on Dravidian languages, offering a framework that could serve Bengali similarly. Radhakrishnan et al. [19] worked on voice cloning for Tamil, addressing issues directly relevant to Bengali, such as tonal accuracy and linguistic authenticity, providing guidance for tackling similar issues in Bengali synthesis. Reddy et al. [20] conducted a comparative study of various TTS models and vocoder combinations, identifying effective configurations for high-quality synthesized speech, which is essential for Bengali given the limitations in data. Lastly, Jia et al. [21] investigated the application of transfer learning from speaker verification to multi-speaker text-to-speech synthesis, showing the benefits of transferring knowledge across domains to improve voice cloning. This approach could be highly useful for Bengali synthesis, leveraging related-task data to enhance synthesis quality and speaker generalization.

Collectively, these studies highlight the importance of developing specialized methodologies for low-resource languages. Building on these studies, researchers can advance

inclusive technologies that empower speakers of underrepresented languages, ensuring innovations in speech synthesis reach diverse linguistic communities like Bengali.

3 Dataset

The Bengali Common Voice Corpus 17.0 is a huge publicly available dataset designed to stimulate advancement in speech technologies for the Bengali language. It was released on March 20, 2024. The data amount totals 24.66 GB. It corresponds to 1273 hours of real speech, all of which but 54 hours are validated. The corpus is under a CC0 open-source license, meaning it can be freely used for both scientific and commercial purposes. The dataset is in MP3 format serving as a highly universal one supported by the majority of digital sound data processing tools. It is a crowd source data with 22913 individual volunteers having contributed to it. It was observed that contributors either read prepared sentences or spoke extemporaneously, which was the reason for obtaining a large variety of accents, dialects, and speech styles even in Bangla-speaking population. It is this uniqueness and diversity that enhances the worth of the dataset in training speech recognition systems, TTS and other NLP tasks. The age structure of contributors has 67% of the contributors being in the 20-29 age group, 7% under 20 years old. Only 3% of contributors were within the 30 to 39 age brackets, and an even smaller proportion of those aged 40+ were represented contributing less than 1% each. Also, 22% of the total contributors chose not to share their information regarding their age. In terms of gender, 54% of the dataset is masculine and 23% feminine. Admittedly, 22% of the contributors did not answer the gender question and there is no participation of the transgender or any other alternate gender population members.

4 Methodology

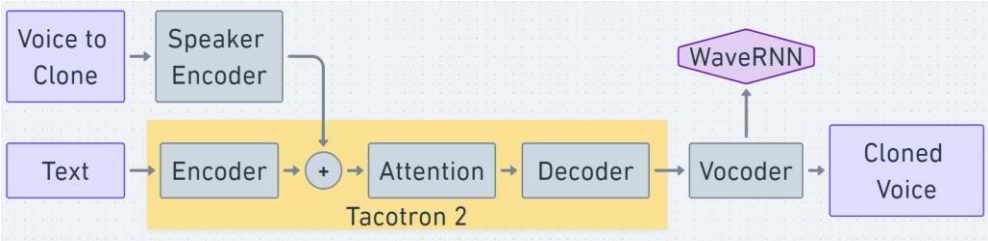


Figure 1: Flow Diagram of Voice Cloning in Methodology

In this paper, we describe the implementation of a voice cloning system through transfer learning from speaker validation (speaker verification) models as put forward by SV2TTS [20]. The method allows the synthesis of speech in a target speaker's voice with only the limited datasets for the training. Our research is about the use of this process in the Bengali language,

which is a low-resource language, and is one of the first voice cloning in Bengali that has ever been documented. Figure 1 offers a synopsis of the entire setup.

4.1 Architecture

Cloning a speaker's voice is a process which involves three main processes: extracting the speakers' vocal characteristics, generating a Mel-spectrogram of a spoken text incorporating these vocal characteristics and converting the Mel-spectrogram into an audio waveform of the voice. This is implemented in the paper through three main components: Encoder, Synthesizer, and Vocoder.

1. **Encoder:** The encoder plays a key role as the first vital part. It is designed to capture a speaker's unique voice features from a brief voice sample. This matters a lot in few voice cloning cases where there is not much speech data to work with. The encoder uses a neural network structure that was first made for speaker verification [21]. Its main job is to turn input speech into a high-dimensional embedding space. In this space, the embeddings are vectors that stand for the speaker's identity. Embeddings from the same speaker group around a center point in this space, while embeddings from different speakers sit further apart.

   To do this, we first change the input speech into log-Mel-spectrograms. These show the power spectrum of the speech over time. The neural network then processes these spectrograms to create fixed-dimensional speaker embeddings. We train the network using a contrastive loss function. This makes sure that the embeddings of speech samples from the same speaker are close in the embedding space, while keeping samples from different speakers apart. The end result is a speaker representation that we can use to condition the synthesis process on a specific speaker's voice traits. These embeddings play a key role in making sure the generated voice copies the intended speaker.

2. **Synthesizer:** After the encoder gets the speaker embeddings, the next step is to create a Mel-spectrogram based on both the text input and the speaker embeddings. The synthesizer handles this process. It produces a time-frequency picture of the spoken text that includes the speaker's unique voice features.

   We expand on Tacotron2's structure, a well-known sequence-to-sequence model that has the purpose to synthesize speech, to accomplish this task [10]. The model has three main parts: an encoder, an attention mechanism, and a decoder. The encoder turns the text input into a series of hidden states, while the attention mechanism lines up the hidden states of the input text with the time steps of the output Mel-spectrogram. In the end, the decoder creates the Mel-spectrogram one frame at a time.

   In our case, we give Tacotron2 new training on a Bengali-specific dataset, Mozilla Common Voice Bengali, to adapt it to the unique sounds and speech patterns of the Bengali

language. As the model trains, it learns to create mel-spectrograms that are both accurate and tailored to the speaker embeddings from the earlier stage. The spectrograms it produces can then mimic the target speaker's voice capturing their unique tone, accent, and way of speaking.

3. **Vocoder:** The last process of the voice cloning process converts the Mel-spectrogram back to actual speech in the form of synthesized audio. The vocoder does this to develop this time series speech wave from the Mel-spectrogram. The one we have adopted is called the WaveNet vocoder which is a neural network of its kind capable of churning out realistic sounding audio [4].

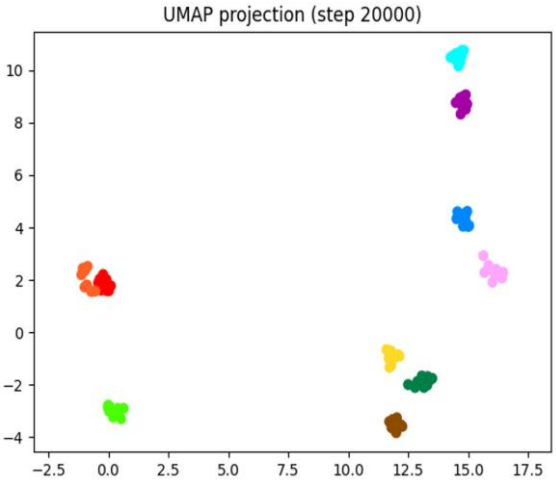
   WaveNet takes the Mel-spectrogram frame by frame and produces one audio sample after the other in a sequential manner. Unlike most other recurrent neural networks (RNNs), WaveNet uses causal convolutions. This enables it to work more by using samples it has diagnosed in the past and which are contained in its database. Therefore, one of the main advantages of this model is its ability to generate long wave forms with great timing accuracy.

   In this particular case, we employ the WaveNet vocoder that is trained in advance. In general, it is supposed to have the ability to produce clear and natural speech waveforms independent of language or speaker. The vocoder is used in conjunction with the mel-spectrogram that the synthesizer generates. This means that it is able to identify the speaker specific features that are embedded in the Mel-spectrogram. The result is more realistic, individualized speech.

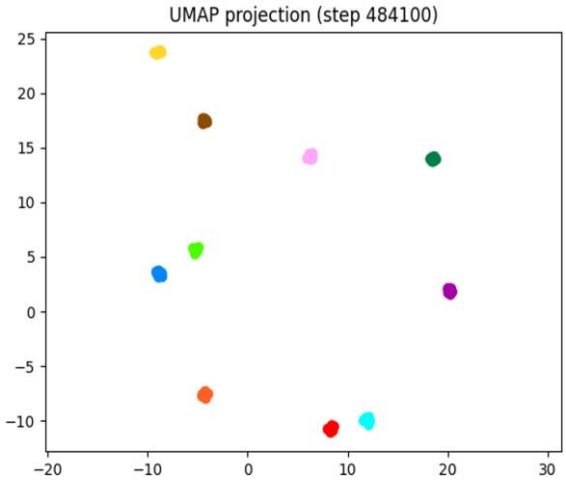
4.2 Bengali Language Adaptation

Modifying this system to mimic Bengali voices made us think about the peculiar phonemes and language characteristics of Bengali. As for the consonants, it is also necessary to understand that Bengali has its own vowels, consonants, and even different tones compared to the voice copying datasets of other languages. Therefore, we have utilized Mozilla Common Voice Bengali dataset to train the system. There are many examples of Bengali speech in this dataset. It eased the synthesizer in understanding how Bengali speakers change the intonation of words and set the pace. We also created a personalized pre-processing system for Bengali text. This system standardizes text and writes out letters of a word. This makes sure that the input text aligns with the linguistic characteristics of Bengali.

In this study, we retrained the speaker encoder synthesizer, and vocoder for Bengali instead of using pre-trained models from other languages. This thorough method was essential to capture the unique sound and language features of Bengali, which are very different from other languages. We needed to retrain the speaker encoder, vocoder, and synthesizer because of the specific details of Bengali sounds, speech melody, and rhythm. Training these parts from the ground up allowed us to clone voices more designed for the language's unique traits.



2(a) Early-stage UMAP of speaker voice clusters



2(b) Late-stage UMAP of speaker voice clusters

Figure 2: Progression of speaker voice clustering in UMAP projections at different training steps

4.3 Speaker Encoder and Vocoder Performance

Training the speaker encoder from scratch needed a big dataset so we used the Mozilla Common Voice Bengali dataset to do this. The encoder did a good job capturing things like pitch, tone, and accent that make each speaker unique. It made distinct embeddings for each speaker that worked well during the synthesis phase. Since we trained the speaker encoder on Bengali data, it was better at telling apart and representing the vocal traits of speakers. UMAP projections in Figure 2 shows the clustering of voice embeddings from different speakers while training a model for voice cloning. Each colour represents a unique speaker, while the different clusters show the way the model differentiates between their vocal characteristics. The step number reflects the progress in training. In later steps, the separation between clusters is clear and indicates the improved ability of the model to distinguish different speakers' voices.

5 Results and Discussion

We trained the vocoder on the same dataset to convert the generated Mel-spectrograms into audio waveforms. This step had an impact on fine-tuning the waveform generation for Bengali phonetic and prosodic structures. The retraining of the vocoder made sure the synthesized voice had a more natural sound and matched the intended speaker's voice. It also prevented any language artifacts that might have been present in a pre-trained, language-agnostic vocoder.

5.1 Synthesizer Training

We trained the synthesizer, which creates Mel-spectrograms from text, on the Bengali dataset for 48 hours. They used powerful computing tools, including 28-core processors, 128 GB of RAM, and a Tesla V100 GPU. This allowed the model to grasp the unique sounds of Bengali such as specific vowel and consonant noises as well as the rhythm and tone changes that make up natural speech in the language. The updated synthesizer made spectrograms that matched both the text input and the speaker patterns letting it produce clear and accurate Bengali speech.

Table 1: Hyperparameters - Synthesizer Training

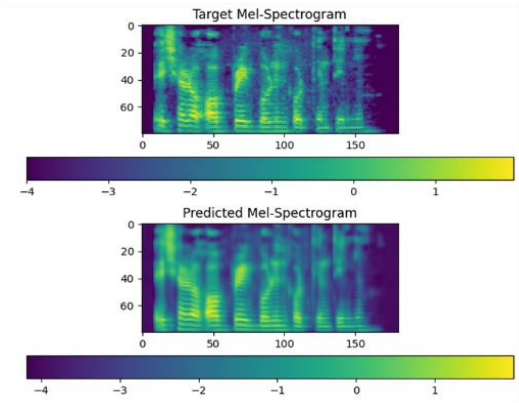
Hyperparameter	**Value**
sample rate	16000
num mels	80
tts embed dims	512
tts encoder dims	256
tts decoder dims	128
tts lstm dims	1024
tts schedule	[(2, 1e-3, 20 000, 12), (2, 5e-4, 40 000, 12), (2, 2e-4, 80 000, 12)]
tts clip grad norm	1.0
tts eval interval	500
synthesis batch size	16
speaker embedding size	256

Since each language has its own unique sound patterns and speech rhythms, it was essential to train the synthesizer. This step made sure the model could produce Bengali speech that sounded natural. The hyperparameters used in the synthesizer training are listed in Table 1.

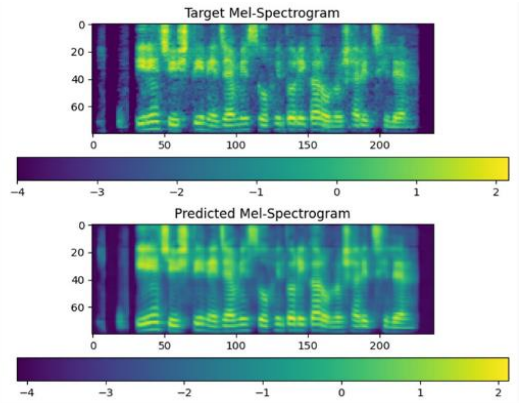
5.2 Validation Process

To assess how well the model performed, we set aside part of the dataset to validate. In Figure 3, the Mel-spectrograms are used to make an objective comparison of the original (target) utterances with the synthesized audio. Precisely, the generated Mel-spectrograms visually representation to access the quality of the produced speech. The top row in each sub-figure presents the target Mel-spectrogram, while the bottom shows the one predicted from the synthesized voice. As training progresses from 50,000 to 220,000 steps, predicted Mel-spectrograms look more and more similar to the target ones, which indicates that the model

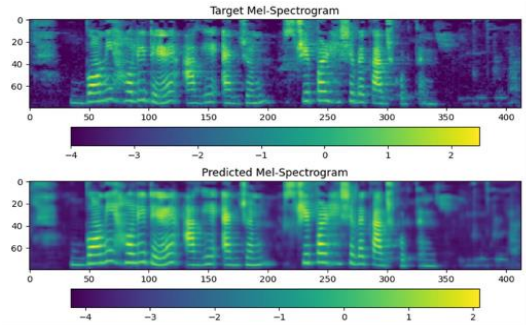
was getting better in generating not only the phonetic information but also the speaker-specific characteristics. This clearly shows the refinement of the model for Bengali.



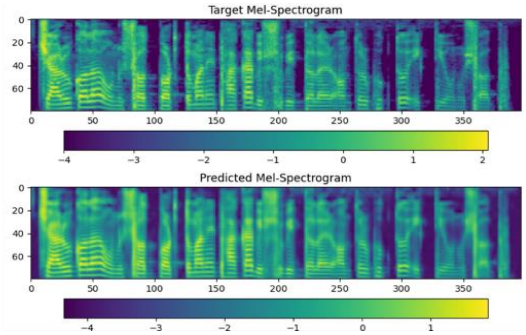
3(a) Predicted vs Target Mel-spectrogram (Step 50,000)



3(b) Predicted vs Target Mel-spectrogram (Step100,000)



3(c) Predicted vs Target Mel-spectrogram (Step 150,000)



3(d) Predicted vs Target Mel-spectrogram (Step 220,000)

Figure 3: Comparison of target and predicted Mel-spectrograms over different training steps.

The Mel-spectrograms indicate that the model is able not only to replicate the text but also to capture speaker-specific vocal features. From this comparison, it is possible to evaluate how well the model can reproduce accurate, speaker-specific speech by visually providing a clear representation of progress during training.

The average cosine similarity obtained is 0.9693 and an average MSE equals to 0.0036 as can be seen in **Error! Reference source not found.**. That is, the predicted mel spectrograms are highly similar with small error in comparison to the target mel spectrograms, therefore showing efficient model performance.

5.3 Evaluation

To figure out how natural and similar to the speaker the created audio sounded, we used the Mean Opinion Score (MOS) method, which is widely accepted as the subjective evaluation metric for accessing the perceptual quality of synthesized speech. MOS involves listeners giving their thoughts on how alike the made-up and real audio are. They rate it from 1 to 5 where 1 means “Very Different” and 5 means “Very Alike”.

To assess the quality, we asked five Bengali native speakers who knew about text to-speech tech to rate 10 samples we created. We gave them the original recordings to compare. Table 2 breaks down the MOS results. The audio we made scored between 3.8 and 4.3 on average. These numbers show that native listeners thought the fake voices sounded a lot like the real speaker. This proves our retraining method works well.

Table 2: Mean Opinion Score for Audio Samples

Audio ID Sample	**Expert Ratings**					**Mean Scores**
	Expert 1	*Expert 2*	*Expert 3*	*Expert 4*	*Expert 5*	
2145	4	3	4	5	4	4.0
8791	5	5	4	5	4	4.6
6512	3	4	4	3	4	3.6
3428	2	3	3	4	2	2.8
7982	5	4	5	5	5	4.8
5376	3	4	3	4	3	3.4
4903	4	3	4	5	4	4.0
6719	2	3	2	3	3	2.6
8546	4	5	4	5	5	4.6
1298	3	3	4	4	3	3.4

6 Conclusion

In this study, we have explored the feasibility of voice cloning for Bengali, a low resource language. This is done by implementing a speaker verification-based model using SV2TTS. The use of the Mozilla Common Voice Bengali dataset has enabled us to capture the unique phonetic characteristics of the language. Our approach has demonstrated that transfer learning when tailored with an appropriate dataset can produce realistic synthetic voices even with limited training data. By retraining key components like the encoder, synthesizer, and vocoder

specifically for Bengali, we have tried to ensured that the resulting synthetic voices not only matched the target speaker's identity but also retained their natural language features. The Mean Opinion Score (MOS) evaluation of our results reflected high levels of naturalness and similarity to the real speaker voices thus proving the effectiveness of the system.

Future work will focus on improving the robustness of the system by incorporating much larger and diverse Bengali datasets along with exploring advanced techniques like speech-to-speech conversion. This could further expand the potential applications of voice cloning in real-time communication, preservation of languages, and other domains requiring synthetic speech solutions. Our findings demonstrate that voice cloning technologies can indeed be extended to languages with lower resources, thus opening avenues for a more inclusive and accessible technology solutions.

References

[1] S. Arik et al., "Neural voice cloning with a few samples," in Adv. Neural Inf. Process. Syst., vol. 31, 2018.

[2] D. Dai et al., "Cloning one's voice using very limited data in the wild," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 8322–8326, 2022.

[3] R. Li et al., "Unet-tts: Improving unseen speaker and style transfer in oneshot voice cloning," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 8327–8331, 2022.

[4] N. Kalchbrenner et al., "Efficient neural audio synthesis," in International Conference on Machine Learning, PMLR, pp. 2410–2419, 2018.

[5] C. Magarin˜os, D. Erro, and E. R. Banga, "Language-independent acoustic cloning of HTS voices: A preliminary study," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 5615–5619, 2016.

[6] Applying wav2vec2 for Speech Recognition on Bengali Common Voices Dataset, [Online].

[7] OOD-Speech: A Large Bengali Speech Recognition Dataset for Out-of-Distribution Benchmarking, [Online].

[8] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am., vol. 50, no. 2B, pp. 637–655, 1971.

[9] H. Zhang and Y. Lin, “Improve few-shot voice cloning using multi-modal learning,” in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 8317–8321, 2022.

[10] J. Shen et al., “Natural tts synthesis by conditioning wavenet on melspectrogram predictions,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 4779–4783, 2018.

[11] L. Zhao and F. Chen, “Research on voice cloning with a few samples,” in 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA), IEEE, pp. 323–328, 2020.

[12] K. N. Stevens et al., “An electrical analog of the vocal tract,” J. Acoust. Soc. Am., vol. 25, no. 4, pp. 734–742, 1953.

[13] H. W. Strube, “The meaning of the Kelly–Lochbaum acoustic-tube model,” J. Acoust. Soc. Am., vol. 108, no. 4, pp. 1850–1855, 2000.

[14] British Council, “The English effect: The impact of English, what it’s worth to the UK and why it matters to the world,” British Council, UK, 2013.

[15] T. H. Crystal, “Speech analysis,” in The Journal of the Acoustical Society of America, R. W. Schafer and J. D. Markel, Eds. Acoustical Society of America, 1980, vol. 68, no. 6, pp. 1906–1906.

[16] Nair, Jayashree, Akhila Krishnan, and S. Vrinda. “Indian text to speech systems: A short survey.” 2022 International Conference on Connected Systems & Intelligence (CSI). IEEE, 2022.

[17] Chowdary, Divi Eswar, et al. “Transformer-Based Multilingual Automatic Speech Recognition (ASR) Model for Dravidian Languages.” Automatic Speech Recognition and Translation for Low Resource Languages (2024): 259-273.

[18] Radhakrishnan, Vishnu, et al. “Voice Cloning for Low-Resource Languages: Investigating the Prospects for Tamil.” Automatic Speech Recognition and Translation for Low Resource Languages (2024): 243-257.

[19] Reddy, Penaka Vishnu, et al. “A Comparative Study of Text-to-Speech (TTS) Models and Vocoder Combinations for High-Quality Synthesized Speech.” 2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2023.

[20] Y. Jia et al., “Transfer learning from speaker verification to multi-speaker text-to-speech synthesis,” in Adv. Neural Inf. Process. Syst., vol. 31, 2018.

[21] Wan, L., Wang, Q., Papir, A., Moreno, I.L., Generalized end-to-end loss for speaker verification, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 4879–4883, 2018.