PRECISION-DRIVEN LOW-RESOURCE SPEECH SYNTHESIS FOR BANGLA TEXT-TO-SPEECH SYSTEM

Tabassum Sadia Shahjahan[1], **Md. Ismail Hossain**[1], **Kazi Rafat**[1], **Md. Ruhul Amin**[2], **Fuad Rahman**[3], and **Nabeel Mohammed**[1]

[1]*Apurba-NSU R&D Lab, Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh*
[2]*Fordham University, NY, USA*
[3]*Apurba Technologies, Sunnyvale, CA 94085, USA*
{*tabassum.sadia, ismail.hossain2018,nabeel.mohammed,shafin.rahman*}*@.northsouth.edu*
mamin17@fordham.edu,fuad@apurbatech.com

ABSTRACT

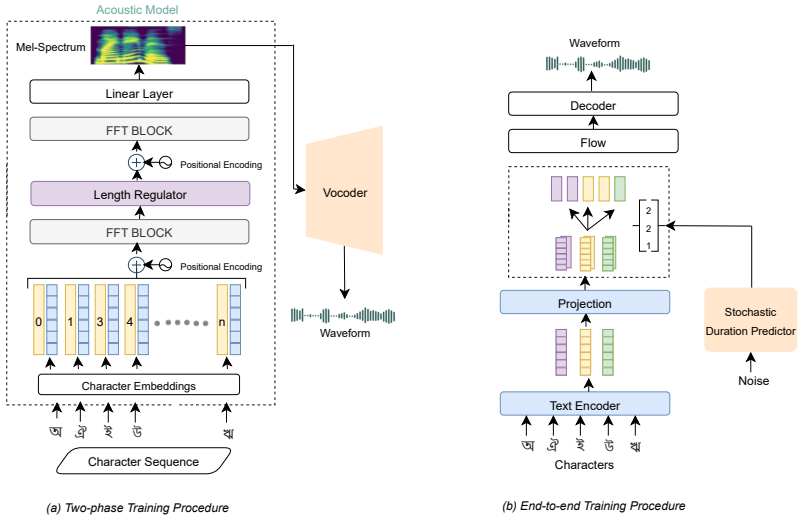
Recent developments in deep learning and artificial intelligence have facilitated widespread commercial adoption of text-to-speech models that can produce intelligible and natural-sounding speech. Although numerous synthetic models are widely available for languages such as English, Chinese, etc., extremely low-resourced languages like Bangla continue to pose a formidable challenge for synthesizing speech data. In this paper, we adopt a single-stage and a two-stage training approach, followed by quantization techniques, to generate high-quality speech from Bangla dataset. Our experimental results show that the proposed models achieve both intelligibility and naturalness with reduced inference time even under extremely low settings. We are the first to provide a robust Bangla Text-To-speech system usable for both academic and commercial applications.

1 INTRODUCTION

The advent of technology has significantly increased the academic and commercial use of text-to-speech models driven by the success of state-of-the-art neural network architectures, such as Tacotron (Shen et al., 2018), Parallel Tacotron2 (Elias et al., 2021), Fast Speech (Ren et al., 2019), Char2wav (Sotelo et al., 2017), Deep Voice 3 (Ping et al., 2018), AlignTTS (Zeng et al., 2020) and Transformer TTS (Li et al., 2019). These simpler seq2seq-based approaches have outperformed the traditional concatenative (Hunt & Black, 1996) and statistical parametric (Tan et al., 2021) speech synthesis processes. The concatenative approach synthesis pre-recorded human phonemes or syllables to generate natura-sounding waveforms (Oloko-oba et al., 2016),whereas the statistical parametric methods involves producing waveforms from similar speech fragments through parametric extraction(Zen et al., 2009).Conventional neural speech models consist of three components: a text analyzer to extract linguistic features from textual inputs, an acoustic model for obtaining acoustic features from the audio files, and a vocoder to generate the speech waveforms. However, the task of speech synthesis for languages with limited data resources still needs to be solved.

There are about 6000 languages spoken worldwide, with Bangla being the 4th most widely spoken language in the world (Xu et al., 2020). Bangla is a linguistically intricate language with idiosyncratic dialects, complex lexicons, gemination, diphthongs, and triphthongs (Showrav, 2022; Rakib et al., 2023). Despite being a widely spoken language, the need for annotated datasets, grapheme-to-phoneme conversion, and high-quality voice recordings make the training process computationally expensive and time-consuming.

In the case of low-resource languages, cross-lingual transfer learning is widely used. However, such methods can significantly differ in prosodic features and input space mismatches, resulting in a lack of spontaneous intonation and semantics in the synthesized speech. Compared to ASR, only a few researches have been conducted on Bangla text-to-speech (Showrav, 2022). To address the



(a) Two-phase Training Procedure *(b) End-to-end Training Procedure*

Figure 1: Schematic representation of the proposed model architectures with two-stage and single-stage training pipelines. (a) The two-stage architecture features an acoustic model with a Feed-Forward Transformer to generate mel-spectrum from Bangla text, followed by a vocoder to produce raw waveforms. (b) The end-to-end model consists of an encoder, decoder, and conditional prior with a flow-based stochastic duration predictor.

issue of high-quality, efficient speech synthesis for morphologically rich languages (Hossain et al., 2022) with limited data resources, we present two approaches for a robust text-to-speech model with sparse datasets. Our significant contributions are as follows: (1) We introduce a Feed-Forward two-stage approach and an adversarial parallel end-to-end method for natural-sounding speech in Bangla, (2) We implement Quantization-Aware training (QAT) (explained in Section 3.3) followed by neural framework compression to enhance the model performance, (3) Our low-resourced model outperforms the state-of-the art base model in terms of efficiency and naturalness.

2 Related Works

Extensive studies have been conducted in the domain of speech synthesis, yet, the field of Bangla Text-To-Speech (TTS) still needs to be explored with a limited number of resources. In 2007, Alam et al. (2011) pioneered the first Bangla Text To Speech model Katha, a unit-selection-based TTS system using diphone concatenation within the Festival Framework. The heavy lexical system lacked natural prosody with slow inference and high runtime, which was minimized by the Subhachan TTS Software Naser et al. (2010), which employed a similar methodology. One of the major challenges in a morphologically rich language like BanglaHossain et al. (2022) is synthesizing speech with the consecutive flow of diphthongs Seddiqui et al. (2002) based on the syntactic structure of a sentence and tokenization of consonants with innate vowels Rashid et al. (2010). Gutkin et al. (2016)& Ahmed & Islam (2019) address this problem by introducing an LSTM-RNN and Hidden Markov Model-based acoustic and syllable-based approaches, which work satisfactorily in several situations.

Regarding performance for low-resourced languages, SPSS systems have shown better results than conventional concatenative methods Fan et al. (2014). Inspired by this methodology, Bhattacharjee et al. (2021) proposes an autoregressive Bangla synthesizer without needing any grapheme-phoneme (G2P) conversion. While these models can successfully convert a Bangla input text to speech, lower inference time and natural-sounding audio are required for a robust and commercial-ready TTS system.

The resource constraints create major hurdles in real-world applications with low-resourced models. In this paper, we introduce an end-to-end Bangla TTS system trained on both a small, unlabeled dataset and an annotated, moderately sized dataset using a monotonic alignment mechanism and probabilistic modeling.

3 METHODOLOGY

Figure 1 illustrates the two different approaches for the training pipeline we used in our work. For the first series of experiments, we used AlignTTS(Zeng et al., 2020), a feed-forward network that generates a mel-spectrogram from a sequence of characters for the acoustic features, and HiFi-GAN Kong et al. (2020) as the vocoder. To minimize the Real Time Factor (RTF) value, we have replaced the heavy-weighted HiFi-GAN with a universal vocoder. We used an end-to-end model called VITS for the next series, which predicts waveforms from an input text sequence. The modules for datasets, pre-processing pipeline, and model architectures for each case are described in Sections 3.1, 3.2 and 3.3 respectively.

3.1 DATASET

We train the models on the open-source CRBLP dataset Alam et al. (2010), published by BRAC University and features a single male speaker. The phonetically imbalanced dataset contains around 9000 audio files with disruptive noises. Compared to large-scale datasets, the CRBLP corpus contained raw and unprocessed files with bilingual composition and was channeled at 44000 Hz stereo. As a result, after the pre-processing step, we were left with only 6000 audio files for training, and the data was split into 5400 in the train set and 600 in the test set.

3.2 PREPROCESSING

We conducted several data preprocessing steps to ensure the quality and fidelity of the generated waveforms. Since Bangla is a low-resourced language and there are few phoneme conversion tools available, we opted for character-based training. The preprocessing steps with the CRBLP Speech Corpus consisted of the following:

- **Text Normalization**: We removed the unnecessary punctuation (such as '-','/',':', etc.) and special symbols (such as '#','@', etc.) from the sentences. To keep the sentences consistent, we standardized the abbreviations, acronyms, and numerical values for proper pronunciation. Moreover, we ran algorithms to check for any words in other languages (e.g. English) and removed them to avoid confusion during training.
- **Filtering Audio Samples**: We removed audio recordings with a duration of more than 16 seconds to enhance the performance of the proposed models.
- **Tokenization**: At the end of each sentence, we added a token "|", and each sentence was reorganized to follow the English speech dataset LjSpeech (**?**) file structure, "$FileName \| Bangla\ Text\ sentence$ ". We employed a character-based training for our chosen model as the low-resourced Bangla language still lacks en efficient grapheme-to-phome converter.
- **Feature Extraction**: We resampled all the audio files to 22050 Hz for the experiment. All WAV files were then converted from stereo to mono for better accuracy, and the extensive background noises and long silences were removed.

3.3 MODEL ARCHITECTURES AND TRAINING

For the purpose of this paper, we use COQUI TTS models to train our datasets, an open-source deep learning toolkit for SOTA text-to-speech synthesis Eren & Team (2023).

AlignTTS: The Feed-Forward Transformer entails assembling a linear layer, several FFT blocks, a length regulator, and a character embedding. Using a time predictor, AlignTTS forecasts the parallel alignment between text and mel-spectrogram. The predictor in the illustration consists of a linear network producing a scalar output, stacked FFT blocks, and a character embedding layer. The mix density network comprises many stacked linear layers, except the final layer, which is followed by dropout, ReLU activation, and layer normalization. The last linear layer represents the Mel-spectrum distribution of each character and produces mean and variance vectors for multi-dimensional Gaussian distributions (Zeng et al., 2020).

HiFi-GAN: HiFi-GAN consists of one generator and two discriminators—multi-scale and multi-period, trained adversarially with loss functions. A mel-spectrogram is fed into a fully convolu-

Table 1: MOS and RTF Of The Trained Models

	CRBLP	
Model	**MOS↑**	**RTF↓**
Baseline	4.30 ± 0.10	1.60
AlignTTS + HiFi-GAN	4.25 ± 0.12	1.85
AlignTTS + Univnet	3.78 ± 0.10	1.35
VITS	**4.43** ± 0.05	<u>1.34</u>
VITS Quantized	<u>4.40</u> ± 0.03	**0.86**

tional neural network generator via adversarial training, with two additional losses for improved stability and model performance. It up-samples the input using transposed convolutions to match the temporal resolution of the raw waveform with the length of the output sequence (Kong et al., 2020).

UnivNet: Univnet is made of a multi-resolution spectrogram discriminator (MRSD) and generator. The discriminators make use of multiple spectrograms and compute waveforms from real signals. The state-of-the-art vocoder employs multi-resolution STFT loss to train the model (Jang et al., 2021).

VITS: The model consists of variational autoencoder (VAE) with residual blocks in the posterior encoder, which are made up of skip connections and dilated convolutions with a gated activation unit. The normal posterior distribution's mean and variance are produced by the linear projection layer that sits above them. Global conditioning is used in residual blocks to accommodate speaker embedding in multi-speaker settings (Kim et al., 2021). As the acoustic model the Glow-TTS architecture and HiFi-GAN as the vocoder.

Quantization-Aware Training (QAT): To reduce computational cost and speed inference, we employ QAT(Krishnamoorthi, 2018) to the end-to-end VITS model without changing the architecture. QAT training optimizes the model through model compression and latency reduction. Initially, we train the base VITS model with full precision and store it along with the quantized weights (Tailor et al., 2020). Given a model as a differentiable function $\mathbf{F_t}(: \mathbf{W_t})$, with high-precision weights as $\mathbf{W_t}$ we try to mimic quantization during training. Weights $\mathbf{W_t}$ are quantized to low-precision weights $\mathbf{W_q}$ through an initial scale and zero for low-bit conversion. For the forward pass input $\mathbf{x_i} \in \mathbf{X}$; $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_s}\}$, with dataset size s, is convolved with quantized weights $\mathbf{W_q}$ to get outputs $\mathbf{O_q} = \mathbf{W_q} \times \mathbf{x_i}$ which is used to calculate loss against labels y known as QAT loss. Subsequently, a fake quantization is included during the forward pass, and gradients are calculated without any precision loss. The quantized values are then stored in nodes with gradients applied to them during the training phase.

4 EXPERIMENTS

We used a V100 GPU to train all our models from scratch. The batch size was set to 14 due to computational restrictions, and the models were trained for 405 K steps. For the AlignTTS we initiated the Adam (Kingma & Ba, 2014) optimizer with β_1=0.9 , β_2=0.998 and weight decay with 1e-06. Similarly, the VITS model was initiated with an AdamW optimizer with β_1=0.8, β_2=0.9, the stability parameter with 1e-09, and a weight decay rate of 0.01. The learning rate for each epoch was set to 0.0001 and a total of 165 Bangla characters were used for all the models.

Evaluation metric: To verify the effectiveness and intelligibility of our proposed models, we used both subjective and objective scoring systems with state-of-the-art Bangla TTS systems(Hirst et al., 1998). Speech intelligibility measures the clarity and accuracy of spoken language; high intelligibility indicates clear understanding, while low intelligibility suggests comprehension difficulties(Miller et al., 1951; Paul et al., 2020) . For the qualitative assessment of the naturalness of the models, we used a five-point Mean Opinion Score(MOS) (Vazquez-Alvarez & Huckvale, 2002). Forty randomly sampled recordings from the test set were reviewed by 15 listeners. As for the objective evaluation, the average Real Time Factor (RTF) score was measured using an Intel Core i5 CPU. The end-to-end VITS model was then converted to Onnx format for the feasibility of incorporating it into applications.

4.1 MAIN RESULTS

Table 1 shows the experimental results for the proposed models. The results show that the quantized end-to-end VITS model achieved much faster inference and MOS scores compared to the base model Srivastava et al. (2020); Mobassir & Ansary (2022) in a low-setting environment. We enhanced the performance significantly by implementing quantization-aware training and compressing the model size. The heavily parameterized AlignTTS for the acoustic model and HiFi-GAN as the vocoder takes nearly 2 seconds to generate speech on an Intel Core i5 CPU with high-quality audio. The inference time is then reduced by replacing the HiFi-GAN with UnivNet which also decreases the naturalness of the speech. This shows that using a lightweight vocoder can reduce the speed but can also heavily affect the sound quality. Conversely, the end-to-end model outperforms the other models in terms of both intelligibility and speed, and the quantized version shows a similar trend. Our results demonstrate that the proposed method successfully speeds up the inference process of the end-to-end model while being trained on a low-resource language with a small-scale dataset that makes it beneficial for industry applications.

5 CONCLUSION

In this research, we describe two-stage and one-stage Bangla TTS systems with high-speed speech synthesis capabilities for real-world on-device applications. Our suggested approach builds upon the popular end-to-end models VITS and the acoustic model AlignTTS, but uses several methods, including quantization and model compression in addition to a pre-established set of preprocessing methodologies, to accelerate inference. Unlike traditional two-stage techniques, the suggested model benefits entirely from its strong optimization process since it is optimized fully end-to-end. The experimental results demonstrated that the suggested techniques can produce speech that is as organically occurring as that produced by languages with plenty of assets, but they can also produce waveforms considerably more quickly. Further research can be performed using grapheme-to-phoneme conversion with lightweight models as it will produce better quality audio samples with less processing time .

REFERENCES

Md Kausar Ahmed and Md Monirul Islam. Syllable-based bengali text to speech system. *Australian Journal of Science and Technology*, 2019.

Firoj Alam, SM Habib, Dil Afroza Sultana, and Mumit Khan. Development of annotated bangla speech corpora. 2010.

Firoj Alam, SM Murtoza Habib, and Mumit Khan. Bangla text to speech using festival. In *Conference on human language technology for development*, pp. 154–161, 2011.

Prithwiraj Bhattacharjee, Rajan Saha Raju, Arif Ahmad, and M Shahidur Rahman. End-to-end bangla speech synthesis. In *2021 International Conference on Science & Contemporary Technologies (ICSCT)*, pp. 1–6. IEEE, 2021.

Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, RJ Skerry-Ryan, and Yonghui Wu. Parallel tacotron 2: A non-autoregressive neural tts model with differentiable duration modeling. *arXiv preprint arXiv:2103.14574*, 2021.

Gölge Eren and The Coqui TTS Team. Coqui tts, December 2023. URL `https://doi.org/10.5281/zenodo.10363832`.

Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K Soong. Tts synthesis with bidirectional lstm based recurrent neural networks. In *Fifteenth annual conference of the international speech communication association*, 2014.

Alexander Gutkin, Linne Ha, Martin Jansche, Knot Pipatsrisawat, and Richard Sproat. Tts for low resource languages: A bangla synthesizer. 2016.

Daniel Hirst, Albert Rilliard, and Véronique Aubergé. Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

Md Ismail Hossain, Mohammed Rakib, Sabbir Mollah, Fuad Rahman, and Nabeel Mohammed. Lila-boti: Leveraging isolated letter accumulations by ordering teacher insights for bangla handwriting recognition. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 1770–1776. IEEE, 2022.

Andrew J Hunt and Alan W Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, volume 1, pp. 373–376. IEEE, 1996.

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. *arXiv preprint arXiv:2106.07889*, 2021.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pp. 5530–5540. PMLR, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 6706–6713, 2019.

George A Miller, George A Heise, and William Lichten. The intelligibility of speech as a function of the context of the test materials. *Journal of experimental psychology*, 41(5):329, 1951.

Syed Mobassir and Md. Nazmuddoha Ansary. Comprehensive bangla tts, 2022.

Abu Naser, Devojyoti Aich, and Md Ruhul Amin. Implementation of subachan: Bengali text to speech synthesis software. In *International Conference on Electrical & Computer Engineering (ICECE 2010)*, pp. 574–577. IEEE, 2010.

Mustapha Oloko-oba, Ibiyemi& S Osagie TS, and Osagie Samuel. Text-to-speech synthesis using concatenative approach. *International Journal of Trend in Research and Development*, 3(2016): 559–462, 2016.

Dipjyoti Paul, Muhammed PV Shifas, Yannis Pantazis, and Yannis Stylianou. Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion. *arXiv preprint arXiv:2008.05809*, 2020.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: 2000-speaker neural text-to-speech. *proc. ICLR*, pp. 214–217, 2018.

Mohammed Rakib, Md Ismail Hossain, Nabeel Mohammed, and Fuad Rahman. Bangla-wave: Improving bangla automatic speech recognition utilizing n-gram language models. In *Proceedings of the 2023 12th International Conference on Software and Computer Applications*, pp. 297–301, 2023.

Muhammad Masud Rashid, Md Akter Hussain, and M Shahidur Rahman. Text normalization and diphone preparation for bangla speech synthesis. *Journal of Multimedia*, 5(6):551, 2010.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32, 2019.

Hanif Seddiqui, Muhammad Azim, Mohammad Rahman, and Muhammed Iqbal. Algorithmic approach to synthesize voice from bangla text. In *The 5th International Conference on Computer and Information Technology (ICCIT)*, 12 2002.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4779–4783. IEEE, 2018.

Tushar Talukder Showrav. An automatic speech recognition system for bengali language based on wav2vec2 and transfer learning. *arXiv preprint arXiv:2209.08119*, 2022.

Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2wav: End-to-end speech synthesis. 2017.

Nimisha Srivastava, Rudrabha Mukhopadhyay, Prajwal K R, and C. V. Jawahar. Indicspeech: Text-to-speech corpus for indian languages. In *International Conference on Language Resources and Evaluation*, 2020. URL `https://api.semanticscholar.org/CorpusID:218977387`.

Shyam A Tailor, Javier Fernandez-Marques, and Nicholas D Lane. Degree-quant: Quantization-aware training for graph neural networks. *arXiv preprint arXiv:2008.05000*, 2020.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*, 2021.

Yolanda Vazquez-Alvarez and Mark Huckvale. The reliability of the itu-p. 85 standard for the evaluation of text-to-speech systems. ISCA, 2002.

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. Lrspeech: Extremely low-resource speech synthesis and recognition. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2802–2812, 2020.

Heiga Zen, Keiichi Tokuda, and Alan W Black. Statistical parametric speech synthesis. *speech communication*, 51(11):1039–1064, 2009.

Zhen Zeng, Jianzong Wang, Ning Cheng, Tian Xia, and Jing Xiao. Aligntts: Efficient feed-forward text-to-speech system without explicit alignment. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6714–6718. IEEE, 2020.