University *of*
New Haven

Department of Computer Science and Engineering

Report of Evaluating and Enhancing Persian and Bengali Text-to-Speech (TTS) Technology

Bahareh Arghavani Nobar, Sharara Hossain
Supervised by: Dr. Vahid Behzadan

November 29, 2023

Abstract

This project is dedicated to exploring and enhancing the state of the art in Persian and Bengali Text-to-Speech (TTS) technology. Building upon a comprehensive report and leveraging a wealth of available resources, our primary goal is to evaluate and improve the existing TTS models for both Persian and Bengali languages.

Persian and Bengali are languages of significant cultural and linguistic importance, with millions of speakers in the Middle East, Central Asia, Iran, Bangladesh, India, and beyond. Despite their rich linguistic heritage, TTS technology for these languages remains underdeveloped, limiting access to voice-assisted applications, digital content, and communication tools for Persian and Bengali speakers.

To achieve this, our project utilizes the state-of-the-art VITS multi-speaker TTS model, trained on a dataset comprising audio and text from Common Voice in Persian and Openslr for Bengali. The VITS model provides exceptional voice synthesis capabilities and the ability to generate natural, expressive speech in multiple voices and styles. By incorporating this advanced model, we aim to create more accurate, natural, culturally relevant, and open-source TTS solutions for both Persian and Bengali speakers.

1 Introduction

The project aims to improve Text-to-Speech (TTS) technology for the Persian and Bengali languages, both of which carry significant cultural and linguistic value but are currently underserved in TTS technology. The venture is important because both Persian and Bengali are widely spoken across multiple regions, and improving TTS technology for these languages would facilitate cultural exchange and boost accessibility for millions of speakers. Persian and Bengali communities currently face

limited availability of affordable or freely accessible TTS solutions, which restricts their access to digital content, education, and communication tools. By providing more accurate, natural, culturally relevant, and open-source TTS solutions for these languages, this project seeks to close this gap. The ultimate goal is to democratize access to language technology, promote linguistic diversity, and empower these historically underserved communities.

2 Background

2.1 Current state of Art in Persian and Bengali TTS

The academic study on Persian Text-to-Speech (TTS) technology, highlighted in the proposal, encompasses two significant projects. The DeepMine-Multi-TTS project[1] stands out for its meticulous creation of a Persian speech corpus, featuring 120 hours of high-quality speech from 67 diverse speakers, both male and female. The comprehensive dataset is accompanied by a rigorous data collection process, including data source selection, text generation, and phonetic transcription. This project also extends beyond data collection, incorporating a Tacotron2-based model enriched with a speaker verification module and HiFi-GAN vocoder for natural speech synthesis. The evaluation using Mean Opinion Scores (MOS) confirms the effectiveness of this dataset in training multi-speaker TTS models for Persian. On the other hand, the ArmanTTS dataset [4]presents a valuable single-speaker Persian dataset based on OpenSubtitles, offering over 9 hours of recorded audio. This dataset is designed for phoneme-based input, featuring a mapping table for phoneme representation, and exhibits an MOS of 2.98 for the TTS model in the evaluation. Together, these academic endeavors provide essential resources and insights for advancing Persian TTS technology, including the creation of high-quality datasets, model architectures, and evaluation methodologies the present status of Persian Text-to-Speech (TTS) models citing diverse open-source repositories and commercial products. Various models including VITS, WaveRNN, tacotron2, HifiGAN, glow, and FastSpeech2, are employed across various data set sizes and durations. The need for improvement in Persian TTS multi-speaker models is emphasized for improved TTS system quality and flexibility. Potential advancements are highlighted, which include gathering a single-speaker TTS dataset for both genders in Persian language, beneficial for training and development of TTS models. In the digital age, these initiatives are essential for enhancing Persian language technology in natural language processing.

There has been relatively limited development in Text-to-Speech research in Bengali. A Bengali TTS dataset was curated by Google which is open-sourced to the research community under Bengali AI. It contains multi-speaker high-quality transcribed audio data. The dataset contains two separate directories of audio recordings of the dialects of Bengali that are spoken in Bangladesh and India- adding to possible diversity in application. The Bangladesh Bengali dataset contains 1891 recordings and the Indian Bengali dataset contains 1376 recordings. The largest dataset is the OpenSLR Bengali ASR dataset which contains 196k utterances. Another dataset with more speaker, phoneme, and environmental diversity is the Common Voices Speech dataset. The Common Voice Corpus 15.0 contains 22,879 voices.

Bengali TTS has come a long way from the design and development of TTS systems, focusing on modules such as normalization, phonetic analysis, prosodic

analysis, and waveform synthesis of the early 2010s. General research trends favor Speech-to-Text synthesis, so, Bengali TTS remains a rarely explored field of research, with the majority of work being done in the early 2000s. Much of the existing work in TTS is now outdated. Saha et al proposed a Deep Neural Network based statistical parametric TTS system in 2019 on a custom dataset[1]. It generates two voices - male and female, and the authors performed multifaceted evaluations to show significantly better performance compared to traditional TTS methods. However, there remains a scarcity of open-sourced Bengali TTS work which presents an opportunity for novelty and innovation.

2.2 Dataset and Training Approach

For this project, we will leverage the Common Voice dataset [2] for Persian and OpenSLR [5] for Bengali, which offer a valuable resource for training TTS models in low-resource languages. Our approach involves training Vits [3] multi-speaker for Persian language as well-resourced but complicate language and vits single speaker for Bengali as low resource one, which provides a strong foundation for TTS models.

Subsequently, we trained these models on both the Persian and Bengali datasets, adapting them to the unique linguistic characteristics of these languages. This fine-tuning process will involve optimizing the models for the complexities of Persian and Bengali, including variations in letter pronunciation, the presence of an unwritten Ezafe in Persian, and other linguistic challenges. By doing so, we aim to enhance the naturalness and accuracy of the generated speech output for both languages. The latest model that released by couqi TTS is xtts version2 which brings more natural voice to the TTS , for Persian we are working on it and trying to make it support persian language as well as it is now only support 16 languages and Persian is not one of them.[6]

2.3 What is Vits ?

VITS (Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech) is a revolutionary text-to-speech system that significantly outperforms various single-stage, end-to-end models. It follows an advanced deep learning approach incorporating Generative Adversarial Networks (GANs), Variational AutoEncoders (VAEs), and Normalizing Flows. As a result, VITS generates natural-sounding audio superior to the current two-stage methods. It autonomously learns from the text-to-audio alignment, eliminating the need for external alignment. Another standout feature is the stochastic duration predictor which enables the synthesis of natural speech with diverse rhythms, reflecting the multiple ways a text may be spoken with different pitches and rhythms. VITS showcases remarkable efficiency, translating audio 67 times faster than its real-time counterpart, while ensuring the highest quality. Its superior performance is evident from the positive reviews and scores evaluating the sound quality. Thus, the VITS approach offers single-stage training, faster translation, and top-notch sound quality.

2.4 What is XTTS ?

The XTTS model is an open-source text-to-speech model, known for its multilingual speech generation capabilities. It can perform language transformations in multiple

languages and is also equipped for cross-language voice cloning. The model is built on a GPT2 backbone and uses a HiFiGAN model for computing the final audio signals, resulting in improved efficiency and reduced latency. The XTTS model is appreciated for its ability to deliver high-quality audio outputs and its potential to perform with low latency. It can deliver unique and expressive voices while maintaining a balance between speed and performance.

2.5 Evaluation Methodology

The primary method for evaluating the performance of our Persian and Bengali TTS models is through a loss calculation by comparing the model's TTS output to the text generated by the Google Speech API. The following steps will be taken to perform this evaluation:

Data Collection: We will collect a diverse set of 6 audio samples in both Persian and Bengali from our dataset for evaluation purposes.

TTS Generation: We will use our TTS models to generate speech for the selected audio samples, resulting in 6 TTS-generated audio samples in WAV format for each language.

Google Speech API: We will pass each of the 6 TTS-generated audio samples in Persian and Bengali to the Google Speech API for automatic transcription into text.

Loss Calculation: To evaluate the performance, we will calculate the loss between the original text (from our dataset) and the text generated by the Google Speech API for both languages. This loss will provide us with a quantitative measure of the accuracy of our TTS models for each language.

Analysis and Improvement: Based on the loss values obtained, we will analyze the areas where our TTS models can be improved. This may involve fine-tuning the models, adjusting parameters, or exploring alternative methods to enhance the quality of the output for both Persian and Bengali.

3 Significance and Impact

The significance of this project lies in addressing the pressing need for accurate, natural, and open-source TTS solutions for both the Persian and Bengali languages. By evaluating and improving existing TTS models for these languages, we aim to bridge gaps in language technology, facilitate language preservation, and unlock new possibilities for Persian and Bengali speakers worldwide.

In conclusion, we believe that this project will contribute to the advancement of Persian and Bengali language AI and TTS technology, ultimately benefiting the respective language-speaking communities and the broader field of language technology.

3.1 Supportive Documents

- Github repository for Persian: `https://github.com/barghavanii/Text-To-Speech-`
- Hugging face demo for Persian: `https://huggingface.co/spaces/saillab/ZabanZad_PoC`

4 References

References

[1] M. Adibian, H. Zeinali, and S. Barmaki, "Deepmine-multi-tts: A persian speech corpus for multi-speaker text-to-speech," *Available at SSRN 4530203*.

[2] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[3] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[4] M. H. Shamgholi, V. Saeedi, J. Peymanfard, L. Alhabib, and H. Zeinali, "Armantts single-speaker persian dataset," *arXiv preprint arXiv:2304.03585*, 2023.

[5] K. Sodimana, K. Pipatsrisawat, L. Ha, M. Jansche, O. Kjartansson, P. D. Silva, and S. Sarin, "A Step-by-Step Process for Building TTS Voices Using Open Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, Gurugram, India, Aug. 2018, pp. 66–70. [Online]. Available: http://dx.doi.org/10.21437/SLTU.2018-14

[6] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, *et al.*, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *arXiv preprint arXiv:2303.03926*, 2023.

References

[1] M. Adibian, H. Zeinali, and S. Barmaki, "Deepmine-multi-tts: A persian speech corpus for multi-speaker text-to-speech," *Available at SSRN 4530203*.

[2] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[3] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[4] M. H. Shamgholi, V. Saeedi, J. Peymanfard, L. Alhabib, and H. Zeinali, "Armantts single-speaker persian dataset," *arXiv preprint arXiv:2304.03585*, 2023.

[5] K. Sodimana, K. Pipatsrisawat, L. Ha, M. Jansche, O. Kjartansson, P. D. Silva, and S. Sarin, "A Step-by-Step Process for Building TTS Voices Using Open Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, Gurugram, India, Aug. 2018, pp. 66–70. [Online]. Available: http://dx.doi.org/10.21437/SLTU.2018-14

[6] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, *et al.*, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *arXiv preprint arXiv:2303.03926*, 2023.

References

[1] M. Adibian, H. Zeinali, and S. Barmaki, "Deepmine-multi-tts: A persian speech corpus for multi-speaker text-to-speech," *Available at SSRN 4530203*.

[2] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[3] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[4] M. H. Shamgholi, V. Saeedi, J. Peymanfard, L. Alhabib, and H. Zeinali, "Armantts single-speaker persian dataset," *arXiv preprint arXiv:2304.03585*, 2023.

[5] K. Sodimana, K. Pipatsrisawat, L. Ha, M. Jansche, O. Kjartansson, P. D. Silva, and S. Sarin, "A Step-by-Step Process for Building TTS Voices Using Open Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, Gurugram, India, Aug. 2018, pp. 66–70. [Online]. Available: http://dx.doi.org/10.21437/SLTU.2018-14

[6] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, *et al.*, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *arXiv preprint arXiv:2303.03926*, 2023.