Research Proposal
on

Industry-Ready Multilingual TTS Framework: Emotion-Aware Voice Synthesis for English, Bengali, and Chinese

Master's admission candidate

RAHAMAN NAGIUR(纳吉)
nagiur@outlook.com
+86 184 1906 0056

Introduction

The rapid advancement of artificial intelligence (AI) has spurred significant growth in speech technologies, impacting sectors from entertainment to accessibility. Demand for natural and emotionally expressive text-to-speech (TTS) systems is surging, driven by applications like audiobooks, audio dramas, and interactive experiences. Current state-of-the-art neural TTS models, while effective in English, face substantial challenges when applied to other languages, especially low-resource languages like Bengali and Chinese. These challenges stem from the complexities of language-specific prosody (intonation, rhythm, stress) and tonal variations. Consequently, synthesized speech often lacks the emotional expressiveness and nuanced prosodic features of human voices, resulting in a robotic quality.

Despite generating understandable speech, existing TTS models often struggle to capture the full range of human emotion and prosodic subtleties. This limitation is particularly acute in low-resource languages due to limited training data and the unique characteristics of prosodic patterns. Furthermore, controlling attributes like age, gender, and emotional expression within synthesized speech remains a significant hurdle. This restricts the practical application of TTS beyond basic transcription tasks.

This research aims to develop an industry-ready multilingual TTS framework that addresses these limitations by generating human-like voice actors for English, Bengali, and Chinese. The framework will focus on creating high-quality synthesized speech with controllable attributes including age, gender, and emotional expression. A critical component is the development of robust models capable of handling the specific prosodic and tonal characteristics of each language, particularly the unique features of Bengali and Chinese. This includes addressing data scarcity for Bengali and refining prosody modeling techniques for all three languages. The resulting system will improve the quality and naturalness of synthesized speech, enabling more engaging and convincing content across a wider range of applications.

Beyond enhancing storytelling and entertainment, this framework has applications in assistive technologies, virtual assistants, AI-powered customer service, and educational tools. The ability to tailor voices to various emotional states and speaker characteristics will enable immersive and engaging interactive experiences across diverse linguistic and cultural contexts, pushing the boundaries of what's possible in generative speech applications.

Problem Statement

Existing text-to-speech (TTS) systems, while showing improvements in naturalness, fall short of producing truly human-like voices, especially in multilingual contexts and for under-resourced languages like Bengali and Chinese. These systems often lack nuanced emotional expression, failing to accurately convey a range of emotions (joy, sadness, anger) and adapt their prosody (intonation, rhythm, stress) to the subtleties of the spoken text. This deficiency is particularly pronounced in multilingual applications and for under-resourced languages, where limited annotated data and the need to capture culturally specific prosodic nuances greatly complicate the task. Consequently, the synthesized voices lack the authenticity and diverse characteristics of human voice actors, significantly limiting their applicability in entertainment, education, and accessibility. This research aims to address these limitations by developing an industry-ready multilingual TTS framework capable of synthesizing human-like voices for English, Bengali, and Chinese. This framework will incorporate advanced emotional expression techniques and model language-specific prosodic variations to produce more expressive and engaging digital voice actors.

Research Objectives

1. **Develop a robust multilingual text-to-speech (TTS) framework:** This framework will support the synthesis of high-quality speech for English, Bengali, and Chinese.
2. **Model and synthesize nuanced emotional expression:** The framework must be capable of accurately conveying a wide range of emotions (e.g., joy, sadness, anger) in synthesized speech. This involves developing techniques for accurate emotion encoding and decoding within the TTS process.
3. **Accurately capture and reproduce language-specific prosodic features:** The framework should effectively model and reproduce the unique prosodic characteristics of English, Bengali, and Chinese, including intonation, rhythm, stress patterns, and tonal variations. This requires specific modeling techniques for each language.
4. **Enable control over speaker attributes:** The framework should allow for control over various speaker attributes (e.g., age, gender, vocal type) during synthesis. This enhances the flexibility and versatility of the generated voices.
5. **Demonstrate industry-readiness of the framework:** The developed framework must demonstrate its practical applicability and potential for use in real-world applications like audiobooks, audio dramas, and documentaries. This might include showcasing the system's speed and efficiency in production contexts.

Research Questions

The research questions of this study are as follows:

1. How can a multilingual TTS framework be developed that effectively handles the unique prosodic and tonal characteristics of English, Bengali, and Chinese, resulting in natural-sounding synthesized speech?
2. What are the most effective techniques for accurately encoding and decoding emotions into speech, enabling the synthesis of emotionally expressive speech in the target languages?
3. How can the framework be designed to allow for the control and manipulation of speaker attributes (e.g., age, gender, vocal type) during synthesis, ensuring flexibility and versatility?
4. How can the framework be validated to ensure that the synthesized speech effectively meets industry standards for naturalness, clarity, and emotional expressiveness, especially for use in applications like audiobooks, audio dramas, and documentaries?

Literature Review

Recent advancements in text-to-speech (TTS) synthesis have led to the creation of remarkably high-fidelity models capable of generating human-like speech from various inputs, including short audio prompts [20, 18, 11, 10]. This progress, largely driven by neural network architectures, has significant implications for various applications, from virtual assistants and chatbots to audio dramas and interactive content creation [8, 15]. However, current systems often face challenges when generalizing to diverse languages, particularly in low-resource settings like Bengali and Chinese, and in conveying nuanced emotions [3, 5].

Autoregressive (AR) models, a prevalent approach in TTS, have achieved impressive zero-shot performance. Examples like NaturalSpeech 3 [9] and VALL-E 2 [2] showcase the capability to synthesize diverse speech styles. Yet, AR models often grapple with inference latency, exposure bias issues, and the need for meticulous tokenizer design [21, 4, 6, 17]. This is where non-autoregressive (NAR) methods, employing parallel processing, offer a compelling alternative.

Diffusion models [7], particularly those utilizing Flow Matching with Optimal Transport (FM-OT) [12], have proven highly effective in NAR TTS systems. These models, exemplified by recent works such as Voicebox [13] and Matcha-TTS [16], directly model the continuous space of audio features, often without explicit phoneme or duration prediction. However, accurately aligning input text to the output synthesized speech remains a significant challenge in NAR models, particularly when dealing with the substantial length differences inherent in these approaches [9]. While frame-wise phoneme alignments have been used in some models, recent research indicates that these can be less effective for naturalness. Methods that skip explicit phoneme-level duration modeling, like E2 TTS [5] and Seed-TTS [1], frequently demonstrate more natural prosody. These models often rely on the model implicitly inferring the duration from the overall sequence length during inference.

The need for robustness in text-speech alignment is highlighted in the literature, especially concerning multilingual TTS [19]. The issue of data scarcity, prevalent in under-resourced languages like Bengali and Chinese, necessitates innovative approaches for model training. Data augmentation and enhancement techniques can help in these cases. Models like DiTTo-TTS [14] attempt to integrate semantic information via pre-trained language models. However, the most effective approach to handle this need for alignment and efficient synthesis, especially for multilingual settings, is still a topic of ongoing research. The research proposed in this paper, F5-TTS [3], seeks to build upon this recent work by focusing on a simpler approach that avoids explicit phoneme-based duration models while achieving comparable, or perhaps superior, performance, especially for robustness.

In conclusion, the field of TTS is evolving rapidly, with a shift toward more efficient and flexible NAR models. However, robust text-speech alignment and addressing the challenges of under-resourced languages remain critical areas for further exploration. This research aims to contribute to this field by developing a robust, multilingual TTS framework for English, Bengali, and Chinese that balances synthesis quality with efficiency, especially when considering industry-level application needs.

Research Methodology

The major research findings of this proposal are structured around three phases: **(1) Data Preparation**, **(2) Model Development**, and **(3) Real-World Application Testing**.

Phase 1: Data Preparation and Feature Engineering

This phase focuses on the meticulous preparation and structuring of the data required for model training. It involves collecting diverse and high-quality audio corpora for each language, encompassing a range of speakers, ages, and emotions. Essential annotations will specify speaker characteristics, emotions expressed, and the corresponding text. The corpora will be designed to represent various accents, dialects, and speaking styles within each language. Parallel corpora (audio and corresponding scripts) specific to different media (e.g., audiobooks, audio dramas, documentaries) will be collected to capture contextual variations. The data will be preprocessed to

address inconsistencies and enhance overall quality through methods like noise reduction, normalization, and segmentation into individual utterances. Automated and manual feature extraction procedures will identify relevant acoustic cues (e.g., pitch, energy, spectral features) and linguistic features (e.g., phoneme duration, pause durations, tonal variations, prosodic contours), crucial for modeling language-specific prosody. The extracted features will be carefully represented in a manner suitable for the model architecture.

Phase 2: Model Development and Training

In this phase, a novel multilingual TTS framework using a transformer-based architecture will be designed. This framework will incorporate a multimodal encoder to integrate linguistic information, speaker characteristics, and emotional cues. The architecture will include language-specific components to handle the unique prosodic and phonetic characteristics of each language (English, Bengali, and Chinese). The framework will be trained on the prepared corpora using appropriate optimization techniques and loss functions. Rigorous validation will be conducted using held-out portions of the data to evaluate the framework's generalization capabilities. Ongoing evaluation during the training process using relevant metrics (like MOS) will guide the development and refinement of the model. This phase emphasizes iterative model refinement and hyperparameter optimization guided by performance evaluations.

Phase 3: Real-World Application Testing and Evaluation

This phase evaluates the practical usability of the developed framework. The trained model will be adapted to specific real-world applications (e.g., generating audio for audiobooks, audio dramas, and documentaries). User-friendly interfaces will be designed for potential deployment. A comprehensive evaluation, combining objective metrics (Mean Opinion Score, ASR error rate) and subjective assessments from expert panels (voice actors, linguists, and audio professionals), will be conducted to assess the framework's effectiveness and overall quality in each application context. This evaluation will focus on the naturalness, expressiveness, and cultural appropriateness of the synthesized voices, providing crucial feedback for iterative improvement.

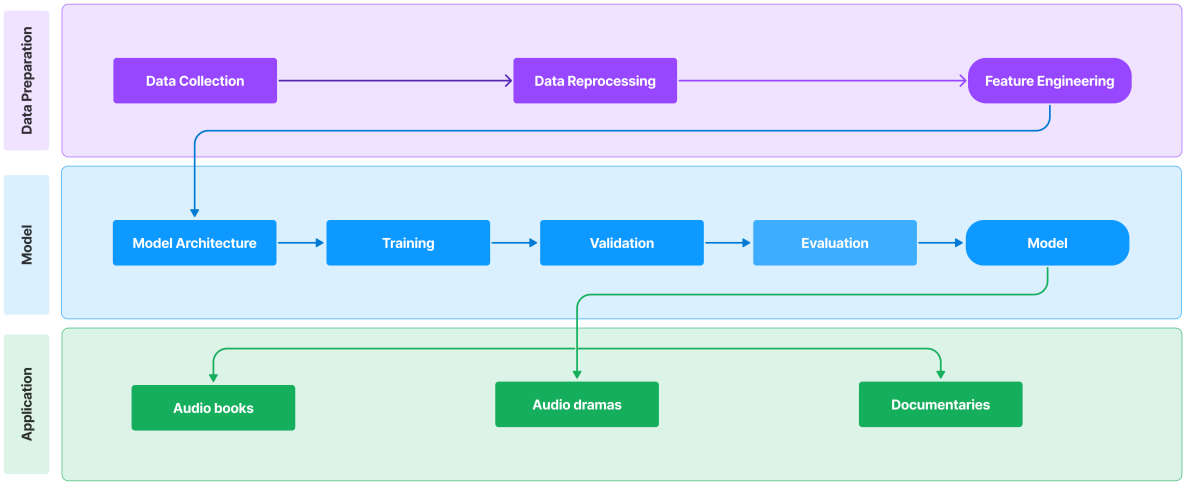

Figure 1: Research flow

This methodology employs a phased approach to develop a robust multilingual TTS framework. Data preparation, model development, and real-world testing phases ensure comprehensive

evaluation and the framework's suitability for applications like audiobooks and documentaries. The approach prioritizes diverse data, language-specific modeling, and expert evaluation to create a high-quality, practical system.

Timeline

Phase	Description	Duration (months)
1. Literature Review	Review existing research and identify research gaps	2
2. Data Collection	Collect and preprocess Bengali and English speech datasets	3
3. Model Development	Design and implement the TTS architecture	4
4. Training & Fine-Tuning	Train and fine-tune the model on the collected datasets	3
5. Evaluation	Conduct subjective and objective evaluations	2
6. Writing & Reporting	Document findings and prepare research papers or technical reports	2

Expected Outcomes

This research is anticipated to yield a robust, industry-ready multilingual text-to-speech (TTS) framework capable of producing high-quality, human-like speech in English, Bengali, and Chinese. The framework is expected to demonstrate significant improvements in the naturalness, clarity, and emotional expressiveness of synthesized speech compared to existing systems, particularly in capturing nuanced prosodic features and conveying a broader range of emotions. This will be evidenced by strong performance across various evaluation metrics, including objective measures (e.g., MOS scores) and subjective assessments (e.g., user evaluations). Furthermore, the research anticipates the publication of at least one high-impact journal article and multiple conference proceedings papers, demonstrating the framework's novelty and potential for widespread adoption.

References

[1] Philip Anastassiou et al. *Seed-TTS: A Family of High-Quality Versatile Speech Generation Models*. 2024. arXiv: 2406.02430 [eess.AS]. URL: https://arxiv.org/abs/2406.02430.

[2] Sanyuan Chen et al. *VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers*. 2024. arXiv: 2406.05370 [cs.CL]. URL: https://arxiv.org/abs/2406.05370.

[3] Yushen Chen et al. *F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching*. 2024. arXiv: 2410.06885 [eess.AS]. URL: https://arxiv.org/abs/2410.06885.

[4] Chenpeng Du et al. *VALL-T: Decoder-Only Generative Transducer for Robust and Decoding-Controllable Text-to-Speech*. 2024. arXiv: 2401.14321 [eess.AS]. URL: https://arxiv.org/abs/2401.14321.

[5] Sefik Emre Eskimez et al. *E2 TTS: Embarrassingly Easy Fully Non-Autoregressive Zero-Shot TTS*. 2024. arXiv: 2406.18009 [eess.AS]. URL: https://arxiv.org/abs/2406.18009.

[6] Bing Han et al. *VALL-E R: Robust and Efficient Zero-Shot Text-to-Speech Synthesis via Monotonic Alignment*. 2024. arXiv: 2406.07855 [cs.CL]. URL: https://arxiv.org/abs/2406.07855.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG]. URL: https://arxiv.org/abs/2006.11239.

[8] Chenxu Hu et al. *Neural Dubber: Dubbing for Videos According to Scripts*. 2022. arXiv: 2110.08243 [eess.AS]. URL: https://arxiv.org/abs/2110.08243.

[9] Zeqian Ju et al. *NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models*. 2024. arXiv: 2403.03100 [eess.AS]. URL: https://arxiv.org/abs/2403.03100.

[10] Jaehyeon Kim, Jungil Kong, and Juhee Son. *Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech*. 2021. arXiv: 2106.06103 [cs.SD]. URL: https://arxiv.org/abs/2106.06103.

[11] Jaehyeon Kim et al. *Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search*. 2020. arXiv: 2005.11129 [eess.AS]. URL: https://arxiv.org/abs/2005.11129.

[12] Nikita Kornilov et al. *Optimal Flow Matching: Learning Straight Trajectories in Just One Step*. 2024. arXiv: 2403.13117 [stat.ML]. URL: https://arxiv.org/abs/2403.13117.

[13] Matthew Le et al. *Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale*. 2023. arXiv: 2306.15687 [eess.AS]. URL: https://arxiv.org/abs/2306.15687.

[14] Keon Lee et al. *DiTTo-TTS: Efficient and Scalable Zero-Shot Text-to-Speech with Diffusion Transformer*. 2024. arXiv: 2406.11427 [eess.AS]. URL: https://arxiv.org/abs/2406.11427.

[15] Yan Liu et al. "M3TTS: Multi-modal text-to-speech of multi-scale style control for dubbing". In: *Pattern Recognition Letters* 179 (2024), pp. 158–164.

[16] Shivam Mehta et al. *Matcha-TTS: A fast TTS architecture with conditional flow matching*. 2024. arXiv: 2309.03199 [eess.AS]. URL: https://arxiv.org/abs/2309.03199.

[17] Puyuan Peng et al. *VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild*. 2024. arXiv: 2403.16973 [eess.AS]. URL: https://arxiv.org/abs/2403.16973.

[18] Yi Ren et al. *FastSpeech 2: Fast and High-Quality End-to-End Text to Speech*. 2022. arXiv: 2006.04558 [eess.AS]. URL: https://arxiv.org/abs/2006.04558.

[19] Takaaki Saeki et al. *Extending Multilingual Speech Synthesis to 100+ Languages without Transcribed Data*. 2024. arXiv: 2402.18932 [eess.AS]. URL: https://arxiv.org/abs/2402.18932.

[20] Jonathan Shen et al. *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*. 2018. arXiv: 1712.05884 [cs.CL]. URL: https://arxiv.org/abs/1712.05884.

[21] Yakun Song et al. *ELLA-V: Stable Neural Codec Language Modeling with Alignment-guided Sequence Reordering*. 2024. arXiv: 2401.07333 [cs.CL]. URL: https://arxiv.org/abs/2401.07333.