研究提案

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

硕士研究生申请人

RAHAMAN NAGIUR(纳吉)

nagiur@outlook.com

+86 184 1906 0056

引言

人工智能（AI）的快速发展推动了语音技术领域的显著增长，影响着从娱乐到无障碍的各个行业。对于自然且情感丰富的文本转语音(TTS) 系统的需求日益增长，这由诸如有声读物、音频剧和互动体验等应用驱动。尽管当前最先进的神经TTS 模型在英语中表现出色，但在应用于其他语言，特别是资源匮乏的语言（如孟加拉语和中文）时，却面临着巨大的挑战。这些挑战源于语言特定韵律（语调、节奏、重音）和声调变化的复杂性。因此，合成的语音往往缺乏人类声音的情感表达和细微的韵律特征，呈现出机械的品质。

尽管现有的TTS 模型能够生成可理解的语音，但它们通常难以捕捉人类情感和韵律细微差异的全部范围。这种限制在资源匮乏的语言中尤其严重，因为这些语言的训练数据有限，而且韵律模式也具有独特的特征。此外，控制合成语音中的年龄、性别和情感表达等属性仍然是一个巨大的障碍。这限制了TTS 的实际应用，使其无法超越基本的转录任务。

本研究旨在开发一个可投入产业应用的多语言TTS 框架，通过为英语、孟加拉语和中文生成逼真的人声来应对这些限制。该框架将侧重于创建高质量的合成语音，并控制包括年龄、性别和情感表达在内的属性。关键部分是开发能够处理每种语言特定韵律和声调特征的鲁棒模型，特别是孟加拉语和中文的独特特征。这包括解决孟加拉语数据稀缺问题以及改进所有三种语言的韵律建模技术。最终系统将提高合成语音的质量和自然度，从而在更广泛的应用中创建更引人入胜和令人信服的内容。

除了增强讲故事和娱乐功能，该框架还适用于辅助技术、虚拟助手、人工智能驱动的客户服务和教育工具。能够将语音调整为各种情绪状态和说话人特征，将能够在不同的语言和文化背景下提供身临其境的互动体验，从而推动生成语音应用的可能性边界。

问题陈述

现有的文本转语音(TTS) 系统，虽然在自然度方面有所改进，但仍难以生成真正逼真的语音，尤其是在多语言环境和资源匮乏的语言（如孟加拉语和中文）中。这些系统往往缺乏细致的情感表达，无法准确传达各种情绪（快乐、悲伤、愤怒），并且无法根据语音文本的细微差别调整其韵律（语调、节奏、重音）。这种不足在多语言应用和资源匮乏的语言中尤为明显，因为有限的标注数据和捕捉特定文化韵律细微差别的需求极大地增加了任务的复杂性。因此，合成的语音缺乏真实性和人类配音演员的多样化特征，极大地限制了它们在娱乐、教育和无障碍领域的应用。本研究旨在通过开发一个可投入产业应用的多语言TTS 框架来解决这些限制，该框架能够为英语、孟加拉语和中文合成逼真的人声。该框架将整合先进的情感表达技术，并对语言特定的韵律变化进行建模，从而产生更具表现力和吸引力的数字配音演员。

研究目标

1. **开发一个强大的多语言文本转语音(TTS) 框架:** 该框架将支持为英语、孟加拉语和中文合成高质量语音。

2. **建模和合成细致的情感表达:** 该框架必须能够准确地传达合成语音中的各种情感（例如快乐、悲伤、愤怒）。这包括开发在TTS 过程中准确编码和解码情感的技术。

3. **准确捕捉和再现语言特定的韵律特征:** 该框架应该有效地建模和再现英语、孟加拉语和中文的独特韵律特征，包括语调、节奏、重音模式和声调变化。这需要针对每种语言的特定建模技术。

4. **实现说话人属性的控制:** 该框架应允许在合成过程中控制各种说话人属性（例如年龄、性别、声线类型）。这增强了生成语音的灵活性和多功能性。

5. **展示框架的产业适用性:** 开发的框架必须展示其在实际应用（例如有声读物、有声剧和纪录片）中的实用性和潜力。这可能包括展示系统在生产环境中的速度和效率。

研究问题

本研究的研究问题如下:

1. 如何开发一个多语言TTS 框架，有效地处理英语、孟加拉语和中文的独特韵律和声调特征，从而生成自然流畅的合成语音?

2. 哪些是最有效的技术，能够准确地将情绪编码和解码到语音中，从而实现目标语言中情感表达的合成语音?

3. 如何设计该框架，以便在合成过程中控制和操纵说话人属性（例如年龄、性别、声线类型），确保灵活性与多功能性?

4. 如何验证该框架，以确保合成语音能够有效地满足行业标准，在自然度、清晰度和情感表达方面达到要求，尤其是在诸如有声读物、有声剧和纪录片等应用中?

文献综述

近年来，文本转语音(TTS) 合成技术取得了显著进展，产生了能够从各种输入（包括短音频提示）生成逼真语音的模型[20, 18, 11, 10] 。这种进步，主要由神经网络架构驱动，对各种应用具有重要意义，从虚拟助手和聊天机器人到音频剧和互动内容创作[8, 15] 。然而，当前系统在推广到不同语言时，尤其是在资源匮乏的环境（如孟加拉语和中文）以及传达细致情感时，仍然面临挑战[3, 5]。

自回归(AR) 模型在TTS 中是常用方法，已实现令人印象深刻的零样本性能。例如，NaturalSpeech 3 [9] 和VALL-E 2 [2] 展示了合成各种语音风格的能力。然而，AR 模型通常会面临推理延迟、

曝光偏差问题以及需要仔细设计的分词器设计[21, 4, 6, 17]。这正是非自回归(NAR) 方法发挥作用的地方，通过并行处理提供了一种有吸引力的替代方案。

扩散模型[7]，特别是那些利用最优传输流匹配(FM-OT) [12] 的模型，在NAR TTS 系统中已被证明非常有效。这些模型，例如最近的Voicebox [13] 和Matcha-TTS [16]，直接对音频特征的连续空间进行建模，通常无需显式预测音素或持续时间。然而，准确地将输入文本与输出合成语音对齐仍然是NAR 模型中的一个重大挑战，尤其是在处理这些方法固有的显著长度差异时[9]。虽然一些模型使用了帧级音素对齐，但最近的研究表明，这种方法对自然度可能不够有效。跳过显式音素级持续时间建模的方法，例如E2 TTS [5] 和Seed-TTS [1],通常表现出更自然的韵律。这些模型通常依赖于模型在推理过程中隐式地从整体序列长度推断持续时间。

文献中尤其强调了文本语音对齐的鲁棒性需求，特别是对于多语言TTS [19]。在资源匮乏的语言（如孟加拉语和中文）中普遍存在的数据稀缺问题，需要创新的模型训练方法。数据增强和改进技术在这类情况下有所帮助。DiTTo-TTS [14] 等模型试图通过预训练语言模型整合语义信息。然而，处理对齐和高效合成需求（尤其是在多语言环境中）的最有效方法仍然是研究的重点。本文提出的研究F5-TTS [3] 旨在通过一种更简单的方法来构建在现有工作之上，避免显式基于音素的持续时间模型，同时在性能上达到相当甚至可能更好的水平，特别是对于鲁棒性。

总之，TTS 领域正快速发展，正朝着更高效和更灵活的NAR 模型转变。然而，强大的文本语音对齐以及解决资源匮乏语言的挑战仍然是需要进一步探索的关键领域。本研究旨在通过开发一个强大的多语言TTS 框架（针对英语、孟加拉语和中文），在合成质量和效率之间取得平衡，特别是考虑到工业级应用需求，来为该领域做出贡献。

研究方法论述

本提案的主要研究成果围绕三个阶段展开：**(1) 数据准备**、**(2) 模型开发**、and **(3) 实际应用测试。**

阶段一数据准备与特征工程

本阶段专注于为模型训练准备和组织数据。它涉及收集每个语言的多样化高质量音频语料库，涵盖各种说话者、年龄和情绪。必要的标注将包含说话者特征、表达的情绪以及对应的文本。语料库的设计将涵盖每种语言内的各种口音、方言和说话风格。将收集特定于不同媒体（例如有声读物、有声剧、纪录片）的平行语料库（音频和相应的脚本），以捕捉语境差异。数据将通过降噪、归一化和将语音分割成单独的单元等方法进行预处理，以解决不一致问题并提高整体质量。自动和手动特征提取程序将识别相关的声学线索（例如音高、能量、频谱特征）和语言特征（例如音素持续时间、停顿持续时间、声调变化、韵律轮廓），这些特征对于建模语言特定的韵律至关重要。提取的特征将以适合模型架构的方式进行精细化表示。

阶段二模型开发与训练

本阶段将设计一个基于Transformer架构的新型多语言TTS框架。该框架将整合多模态编码器，

以整合语言信息、说话人特征和情感线索。该架构将包含语言特定组件，以处理每种语言（英语、孟加拉语和中文）的独特韵律和语音特征。该框架将使用合适的优化技术和损失函数在准备好的语料库上进行训练。将使用保留数据进行严格验证，以评估框架的泛化能力。在训练过程中持续使用相关指标（如MOS）进行评估，将指导模型的开发和改进。本阶段重点在于迭代模型改进和超参数优化，这些优化将由性能评估指导。

阶段三实际应用测试与评估

本阶段评估所开发框架的实际可用性。训练后的模型将被调整以适应特定的实际应用（例如，为有声读物、有声剧和纪录片生成音频）。将设计用户友好的界面，以实现潜在的部署。将进行全面的评估，结合客观指标（平均意见得分、ASR 错误率）和专家小组（配音演员、语言学家和音频专业人士）的主观评估，以评估该框架在每个应用情境中的有效性和整体质量。该评估将重点关注合成语音的自然度、表达性和文化恰当性，为迭代改进提供关键反馈。

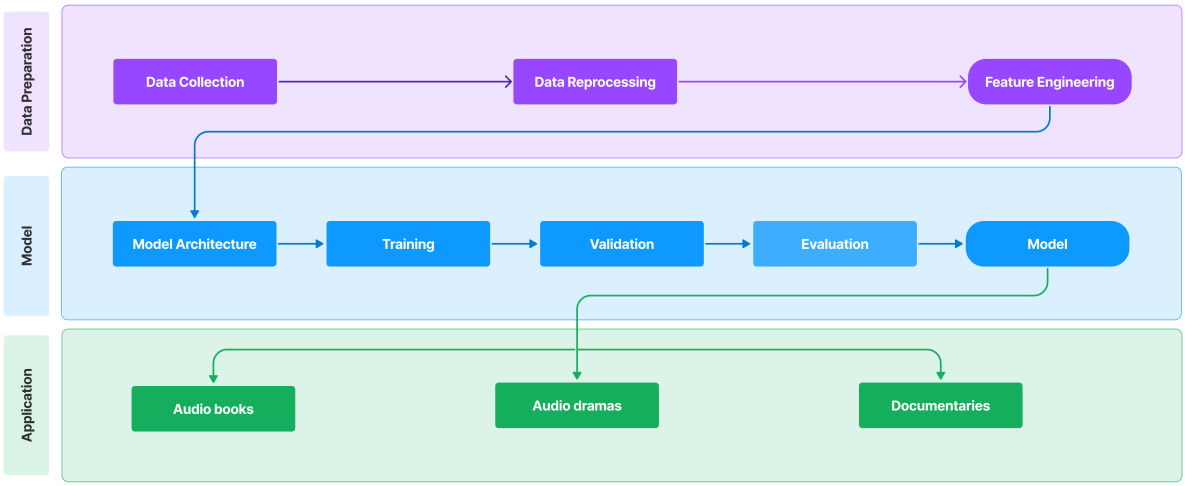

Figure 1: Research flow

该方法采用分阶段的方法开发一个强大的多语言TTS框架。数据准备、模型开发和实际应用测试阶段确保了全面的评估，并使该框架适合于有声读物和纪录片等应用。该方法优先考虑多样化数据、语言特定建模和专家评估，以创建高质量且实用的系统。

预期成果

本研究预期将产生一个强大的、可投入产业应用的多语言文本转语音(TTS) 框架，能够生成高质量、类人语音，并支持英语、孟加拉语和中文。与现有系统相比，该框架预期在合成语音的自然度、清晰度和情感表达方面取得显著改进，尤其是在捕捉细微的韵律特征和传达更广泛的情感方面。这将通过各种评估指标（包括客观指标，例如MOS 分数，和主观评估，例如用户评估）的出色表现得到体现。此外，本研究预计将发表至少一篇高影响力的期刊文章和多篇会议论文，展现该框架的新颖性和广泛应用潜力。

时间线

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

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.