LLM for Generating Personalized Audiobooks

Abhijeet Anand
Indraprastha Institute of Information Technology Delhi
abhijeet21509@iiitd.ac.in

Ankit Gautam
Indraprastha Institute of Information Technology Delhi
ankit21518@iiitd.ac.in

Kartikey Dhaka
Indraprastha Institute of Information Technology Delhi
kartikey21534@iiitd.ac.in

Samridh Girdhar
Indraprastha Institute of Information Technology Delhi
samridh21282@iiitd.ac.in

Soumya Mohapatra
Indraprastha Institute of Information Technology Delhi
soumya21103@iiitd.ac.in

Sushane Dulloo
Indraprastha Institute of Information Technology Delhi
sushane21292@iiitd.ac.in

***Abstract*—This project aims at studying the capabilities of large Language models (LLMs) for personalizing and generating Audiobooks Narrations. This is done with advanced text-to-speech (TTS) generation that empowers unique narration for users to select preferred voices, accents tones and emotions set specifically per genre mood and audience type. The project will also allow for dynamic adjustments, providing a unique and personalized listening experience to users that is personalised to various literary contexts. LLM narrations are audio solutions for video game worlds or fake emotions performance in a fictional piece, or formal tone on educational content. It uncovers a way to create the audiobooks for entertainment as well as access and education, opening up new possibilities as this sort of project can deliver additional material to have listeners involved in reading themselves into an enriched storytelling experience.**

I. Objectives:

1) Building LLM-powered system which can generate audiobook narrations in configurable voices and accents.
2) Assessing the quality and adaptability of the generated narration across different genres.
3) Exploring Personalized Audiobooks potential in Education, Accessibility, and Entertainment Sectors.
4) To evaluate user satisfaction and engagement with LLM-produced audiobooks.

II. Literature Review

The recent research in the field of Text-to-Speech systems have provided us various deep learning like Tacotron ,Wavenet and Large Language Models like GPT-4 that have highly advanced the quality of speech generation from the given text as input. Studies by Shen et al. (2018) and van den Oord et al. (2016) have demonstrated how neural networks are capable of synthesizing high-quality speech with attention mechanisms and autoregressive models. Work on voice customization, as seen in models like Google's TTS API and OpenAI's voice synthesis research, offers personalization options in speech generation, allowing users to select various voice styles, accents, and emotions. Moreover, efforts in the audiobook industry have explored automated narration but are often limited in terms of flexibility and emotional expressiveness. Prior work on neural storytelling by Fan et al. (2018) suggests the potential for combining narrative generation with TTS to create dynamic and engaging audiobook experiences. However, little work has been done to personalize the narration experience at a fine-grained level such as voice selection, accent modulation, and emotional variation, which this project aims to address.

III. Approach

1) Data Collection: For this we use a large and diversified dataset of audiobook texts and their corresponding human voice narrations which will be covering different voices, accents and emotional tones. We will use some publicly available TTS datasets like LibriSpeech.
2) Preprocessing: text will be preprocessed with relevant features to particular emotions or accents (e.g., phonetic and prosodic features). These aspects will help the LLM-based narration system in producing expressive utterances.
3) Model Selection- select a pretrained LLM (eg: GPT-3 or TTS-model) as a base for fine-tuning.
4) Model Training : We will take a LLM based TTS model (like GPT-3) trained on voice + emotion data and fine-tune it. This will involve training on multi-speaker data and fine-tuning for personalized voices and accents.
5) Integration with LLMs: The fine tuned LLM will generate narrations for the audiobooks based on user-set parameters for voice or accent, and emotional tone. Generating narrations that match the text's emotional and contextual content will be done by combining TTS with LLM text understanding.
6) User Interface Development: A web-based interface will be built which will allow users to customize narrations by making choices on various parameters like voice, accent, and emotion. The system will also be capable of providing real-time previews and recommendations.
7) Evaluation and Testing: Extensive testing will be conducted to evaluate the quality, naturalness, and user satisfaction of the generated audiobooks.

IV. TimeLine

The project timeline includes the following key dates and milestones:

- Week 1-2: Research and selection of a suitable pre-trained LLM for fine-tuning.

- Week 3: Collection and preprocessing audiobook dataset texts and narration styles.

- Week 4-5: Fine-tune the selected model, focusing on emotional tone and pacing adjustments.

- Week 6: Developing a basic user interface for customization on various parameters and start conducting tests.

- Week 7: Gather feedback from users, refine the fine tuned model and interface, and prepare for final presentation.

- Week 8: Present the project, showcasing the capabilities of the customized audiobook-generating LLM and its potential applications.

V. Evaluation Criteria

1) Speech Quality: benchmarking of the naturalness and clearness of the generated narrations using appropriate metrics.
2) Customization Accuracy: assess how precisely the model reflects user-selected voice, accent and emotional tones.
3) User Satisfaction: Conduct user studies to assess the listening experience, including ease of customization, enjoyment and engagement.
4) Narrative Coherence: evaluating emotional and tonal consistency of narration across audiobook chapters.
5) Technical Feasibility: Assess computational efficiency, processing time and memory usage of the system during audiobook generation.

References

[1] Zen, H., Senior, A. W. (2019). Neural Speech Synthesis with Transformer Network. *arXiv preprint arXiv:1904.09517*.

[2] Alonso, J. L. A., de los Mozos, V., Moyano, R. G., Herrero, A. (2024). Aerodynamic performance of rectangular wings at low Reynolds numbers using CFD. *Open Research Europe*, 4, Article 44.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017). Attention is All You Need. *arXiv preprint arXiv:1712.05884*.

[4] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv.