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

I. Introduction

The objective of this project is to develop an advanced audiobook system capable of converting text input into high-quality, natural-sounding audio narration. Unlike conventional text-to-speech (TTS) systems, our approach incorporates expressive storytelling by embedding emotions, character-specific accents, and multi-voice narration to enhance listener engagement.The core methodology is built on two primary components: (1) leveraging Large Language Models (LLMs) for annotating the provided text into semantically meaningful and narratively coherent segments, and (2) integrating the annotated output with an advanced TTS system to generate expressive and contextually appropriate audio output. The LLM is tasked with identifying and labelling emotional cues, character identities, and other relevant attributes in the text, ensuring these are accurately translated into audio characteristics. The TTS system then synthesises speech with appropriate prosody, tone, and accents, aligned with the annotated descriptions.This approach aims to bridge the gap between traditional monotonic TTS systems and the more dynamic demands of audiobook narration. By incorporating natural language understanding and expressive speech synthesis, our method presents a novel framework for generating immersive audiobooks. The subsequent sections of this paper detail the methodology, implementation, and evaluation of our system, along with a discussion of its potential applications and future directions for improvement.

II. Approach

Our proposed framework consists of two primary components:

- **LLM-Based Contextual Annotation**: Responsible for segmenting and annotating the story text with speaker and emotional attributes.
- **TTS System**: Generates expressive and natural-sounding voices for multiple speakers based on the annotations.

A. LLM Contextual Annotation

The first step in the framework involves the decomposition and annotation of long-form text into smaller, semantically meaningful segments. This is achieved using advanced Large Language Models (LLMs) to process the story input and produce a structured output containing speaker identification and emotional context for each segment.

1) Text Segmentation and Speaker Identification: To segment long text into manageable sentences with specific emotions and speaker labels, we employed the LLaMa 3.1 405B API. This model was tasked with parsing the input text and generating an annotated file where each sentence was tagged with its respective speaker and emotional state. Due to the unavailability of a publicly accessible dataset linking long text to such detailed annotations, we manually constructed a dataset by annotating five short stories. Through manual evaluation, we observed that the speaker segmentation results generated by LLaMa 3.1 405B closely aligned with human annotations. However, since there was no pre-existing data to facilitate fine-tuning for this specific task, the segmentation process relied on the base model's inherent capabilities.

2) Emotion Prediction: To enhance the accuracy of emotion predictions in the annotations, we fine-tuned the LLaMa 3.1 8B Instruct model (QLoRA finetuning)musing the Multimodal EmotionLines Dataset (MELD) on an Emotion Recognition in Conversation (ERC) task. This step was critical as storytelling often involves multiple speakers and requires accurate identification of emotional tones to maintain the narrative's authenticity and engagement. The MELD dataset provided a robust foundation for training the model to detect nuanced emotions within conversational contexts.

B. Gan Based Emotional Voice Conversion

This step involved converting neutral voice segments we got from our TTS system into appropriate sentiment based voice. to do this we followed a paper on Independent Speaker based Emotional Voice Conversion that used VAW-GAN model to convert voices. The quality of the voices were however subpar since we have limited training data. For training of this GAN we used EmoV-DB.

C. TTS System

1) Adding Multi-Speaker to the Audio: After establishing a mapping between the characters and speaker categories, the next step in our framework involves generating audio files for the segmented text chunks. Each chunk is assigned a specific speaker based on the character-to-speaker mapping, ensuring that the audio narration reflects the character's unique voice profile.

- **Voice Synthesis**:The synthesis of speech for each chunk is performed using the WhisperSpeech TTS model. WhisperSpeech was selected due to its exceptional capability to clone and generate natural-sounding voices with high fidelity. This model allows for precise replication of distinct vocal characteristics, enabling dynamic storytelling that captures the essence of each character and narrator.To make the narration engaging and realistic, we collected voice data for all six predefined speaker categories (child male/female, adult male/female, and elderly male/female) from friends and family members. The data collection process was conducted with proper consent to ensure ethical compliance. These voice samples were then used as references for the WhisperSpeech model to generate personalized and expressive speech for each character.
- **Audio File Assembly:**For each annotated chunk of text, an individual audio file was generated corresponding to the assigned speaker. These files were then combined to create the final audiobook. The assembly process involved concatenating the audio chunks with a 1-second pause between them to ensure smooth transitions and maintain the listener's engagement. The final result is a cohesive audio file that narrates the story in a natural and captivating manner, with distinct voices and emotional depth reflecting the characters and the narrative.

III. Limitations

Here's a refined and structured explanation of the challenges and limitations you faced:

Challenges and Limitations While developing our framework, we encountered several challenges that highlighted the limitations of existing resources and methodologies. These issues are discussed below:

- **LLM Annotation**:A significant challenge in the annotation process was the lack of a publicly available dataset that maps long text to speaker and emotion annotations. This forced us to manually annotate five short stories to create a reference dataset for speaker segmentation. Since no fine-tuning was performed for this specific task due to the unavailability of data, the model relied on its default capabilities. This manual approach was time-consuming and limited the scalability of the system for larger datasets.
- **Emotion Recognition in Conversation (ERC)**: For the ERC task, we fine-tuned the LLaMa 3.1 8B Instruct model on the MELD (Multimodal EmotionLines Dataset). However, the MELD dataset is highly imbalanced, with neutral and joyful emotions being significantly overrepresented. This imbalance impacted the model's ability to accurately predict less frequent emotions. While there are benchmarks available for MELD, they often require substantial computational resources for training. The LLaMa 3.1 8B model was selected for its compatibility with Google Colab's limited computational capabilities, but this trade-off resulted in suboptimal emotion prediction performance.
- **Limited Speaker Diversity**:The speaker mapping process involved categorizing characters into six predefined classes: child male/female, adult male/female, and elderly male/female. However, each class was represented by only one voice recording. This approach limited the diversity of voices, especially in scenarios where multiple characters belonged to the same class. In such cases, the narration lacked auditory distinction between characters. A more robust solution would involve using multiple voice recordings for each class and dynamically selecting alternate voices when one has already been used.To address this limitation, a scalable approach could involve creating a manually annotated dataset that maps voice characteristics to descriptions of the person speaking. This dataset could then be used to train LLMs for character-specific voice assignments based on textual descriptions.
- **Limitations in Existing EVC Models**: While exploring Expressive Voice Conversion (EVC) models, we encountered a significant drawback: many existing models lose speaker identity during voice conversion, resulting in a uniform voice output. Although speaker-independent EVC models exist, they require extensive parallel voice datasets from a large number of speakers, which are not readily available. This limitation made it infeasible for us to incorporate such models into our framework at this stage.

IV. Conclusion

In this research, we proposed a novel framework for transforming textual stories into expressive audiobooks, integrating advanced Large Language Models (LLMs) and state-of-the-art Text-to-Speech (TTS) technologies. By leveraging LLMs for contextual annotation and WhisperSpeech for natural voice synthesis, we successfully demonstrated a system capable of generating engaging and dynamic audio narration that incorporates character-specific voices and emotional nuances. Our framework showcases the potential for automated audiobook creation, bridging the gap between traditional text-to-speech systems and the creative demands of storytelling.The methodology addressed key challenges in text segmentation, emotion recognition, and character-to-speaker mapping, employing innovative solutions such as fine-tuning LLMs on ERC tasks and designing a systematic speaker mapping strategy. While the results highlighted the system's ability to produce coherent and natural-sounding audio, several limitations emerged, including the lack of annotated datasets, imbalanced emotion data, and constrained speaker diversity. These challenges underscore the need for richer datasets, scalable annotation methods, and more advanced models to enhance performance.Despite these limitations, the proposed system demonstrates significant promise in automating audiobook creation. Future work could focus on expanding the diversity of speaker representations, integrating speaker-independent expressive voice conversion models, and developing scalable datasets to enable fine-tuning and improved annotation capabilities. The insights gained from this research contribute to advancing the field of expressive speech synthesis and highlight the transformative potential of LLMs and TTS systems in the realm of creative audio applications.

- **Speaker Segmentation:** Dividing the text according to different speakers in the dialogue.
- **Emotion Annotation:** Assigning emotional labels to each segment to capture the nuances in the conversation.

This annotated text serves as input for the TTS engine. Since there is no existing pipeline that maps raw text to segmented text with speaker and emotion annotations, we synthesized the required processed text using LLMs. This approach allows us to generate a large amount of annotated data necessary for training robust models.

Initially, we employed the LLaMA 3.1 405B API for annotation through prompt engineering. While the model performed well in speaker segmentation, it showed limitations in emotion recognition. Recognizing that standard emotion classification models typically consider only the current sentence, we aimed to develop a model that accounts for preceding context and emotional shifts to achieve more natural transitions.

To enhance emotion recognition, we explored various fine-tuning techniques on the LLaMA 3.2 1B model, including:

- **Prompt Tuning**
- **Prefix Tuning**
- **Low-Rank Adaptation (LoRA)**

For experimental purposes and to facilitate local testing, we opted to use the smaller LLaMA 3.2 1B model to assess the capabilities of fine-tuned smaller models

A. *Text-to-Speech Engine*

The TTS engine uses the annotated text to generate speech. By incorporating speaker and emotion annotations, the TTS system aims to produce more expressive and contextually appropriate audio outputs.

References

[1] *https://github.com/collabora/WhisperSpeech*

[2] *https://github.com/jiaaro/pydub*

[3] *https://github.com/blessontomjoseph/TTS*$_{w}ith_{e}motion$

[4] *https://github.com/ZET-Speech/ZET-Speech-Demo*

[5] *https://arxiv.org/abs/1806.00674*

[6] *https://www.researchgate.net/publication/341388058-Converting-Anyone's-Emotion-Towards-Speaker-Independent-Emotional-Voice-Conversion*