LLM for Generating Personalized Audiobooks

End-Sem Presentation

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Content

- Problem Statement
- Motivation
- Literature review
- The novelty in our idea
- Methodology
- References

Problem Statement

- Develop an advanced audiobook system capable of converting text input into high-quality, natural-sounding audio narration that incorporates expressive storytelling elements.
- The system aims to enhance listener engagement by embedding emotions, character-specific accents, and multi-voice narration. This involves leveraging Large Language Models (LLMs) for annotating the text into semantically meaningful and narratively coherent segments, capturing emotional cues and character identities.
- The annotated text is then integrated with an advanced Text-to-Speech (TTS) system to generate expressive and contextually appropriate audio output, bridging the gap between traditional monotonic TTS systems and dynamic audiobook narration.

Motivation

- The increasing demand for automated, real-time audio processing solutions in various industries.
- Challenges faced in traditional methods, such as handling noise, multiple audio formats and dialect variations.
- Need for a comprehensive solution that integrates multiple advanced tools for seamless workflow.
- Potential impact on sectors like customer service, transcription services and accessibility technologies.

Literature review

- Overview of popular audio processing tools (e.g., Whisper AI, pydub) and their main capabilities.
- Limitations of older tools, such as slower processing speeds and limited accuracy.
- Analysis of recent advancements in AI-driven transcription models and audio processing techniques.
- Comparative analysis highlighting how modern approaches outperform traditional methods in efficiency and results.

The novelty in our idea

- **What we are doing -** LLM for Generating Personalized Audiobooks
- **What currently exists :-**
  - LLMs that generate speech
  - LLMs that that annotate using context
  - Voice Understanding and Generation Foundation Models
  - LLM for Multilingual Speech-to-Text
  - LLM with Strong automatic speech recognition (ASR)
  - Emotion detection in text : https://arxiv.org/abs/1806.00674

Methodology

1. LLM-Based Contextual Annotation

- Text Segmentation and Speaker Identification:
  - Segmented long-form text into manageable sentences.
  - Annotated each sentence with speaker identification and emotional state.

- Emotion Prediction:
  - Fine-tuned LLaMa 3.1 8B Instruct model using QLoRA fine-tuning. MELD Dataset used.

- Character-to-Speaker Mapping:
  - Retrieval-Augmented Generation (RAG) with LLaMa 3.2 3B model.

Methodology

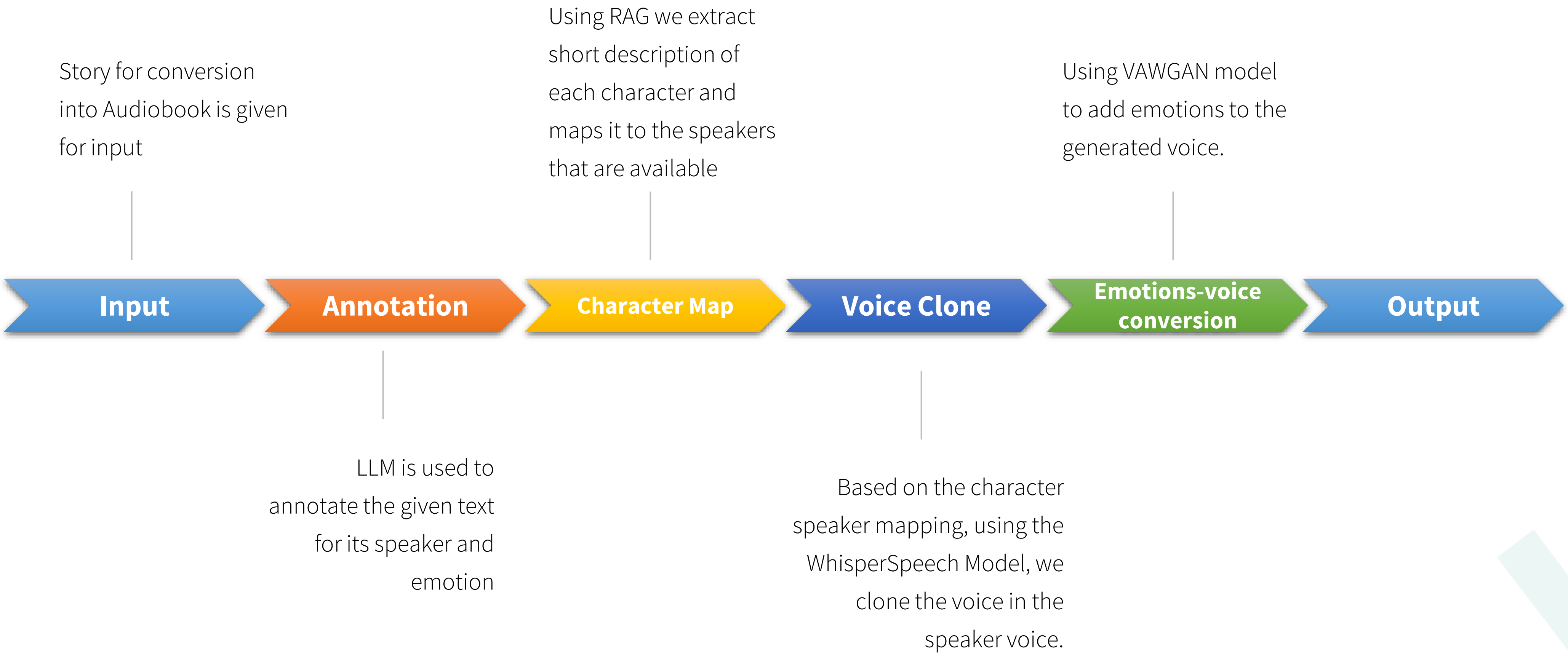
2. Text-to-Speech (TTS) System

- **Voice Synthesis**:
- **Model Used**: WhisperSpeech TTS model.
- **Process**:
  - Performed speech synthesis for each text chunk based on assigned speakers.
  - Collected voice data for all six speaker categories from friends and family with proper consent.

Audio file Assembly:

- **Methodology**:
  - Generated individual audio files for each annotated text chunk.
  - Combined files with 1-second pauses to ensure smooth transitions.

Model HLD



Input

Story for conversion into Audiobook is given for input

Annotation

LLM is used to annotate the given text for its speaker and emotion

Character Map

Using RAG we extract short description of each character and maps it to the speakers that are available

Voice Clone

Based on the character speaker mapping, using the WhisperSpeech Model, we clone the voice in the speaker voice.

Emotions-voice conversion

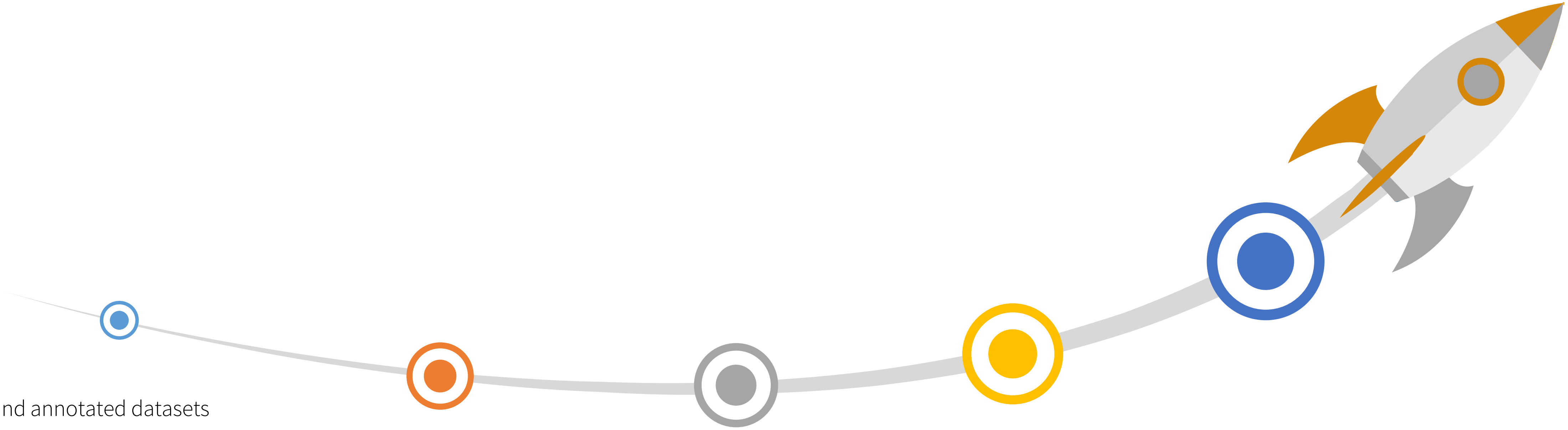
Using VAWGAN model to add emotions to the generated voice.

Output

Limitations

- many existing models lose speaker identity during voice conversion, resulting in non-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.
- Imbalance in MELD dataset affecting emotion prediction.
- Require high computation resources.

Future work



Expand annotated datasets for better fine-tuning

Incorporate more diverse voice recordings

Explore advanced EVC models requiring less data

Accents of each voice can be implemented to include various ethnicities as well

References

- https://github.com/collabora/WhisperSpeech
- https://github.com/jiaaro/pydub
- https://github.com/blessontomjoseph/TTS_with_emotion
- https://github.com/ZET-Speech/ZET-Speech-Demo

Thank You

Open to Questions!