LLM for Generating Personalized Audiobooks Baseline Results

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 report presents the evaluation of Text-to-Speech (TTS) and Large Language Models (LLMs) for audiobook generation and text segmentation tasks. Two independent evaluations were conducted: (1) assessing the performance of a TTS model trained on the LJSpeech dataset, and (2) evaluating LLMs for their accuracy in segmenting stories into lines, identifying speakers and annotating emotions.**

I. Introduction

This report presents the evaluation of Text-to-Speech (TTS) and Large Language Models (LLMs) for audiobook generation and text segmentation tasks. Two independent evaluations were conducted:

(1) Assessing the performance of a TTS model trained on the LJSpeech dataset

(2) Evaluating LLMs for their accuracy in segmenting stories into ines, identifying speakers and annotating emotions.

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. Text-to-Speech

A. Evaluation

We implemented a TTS system using the ESPnet2 toolkit and evaluated it using the LJSpeech dataset. The system employed a pre-trained model (kan-bayashi/ljspeech-vits) based on the VITS architecture. The evaluation focused on three key metrics: Mel-Cepstral Distortion (MCD), Perceptual Evaluation of Speech Quality (PESQ) and Mean Opinion Score (MOS). MCD measures the difference between the predicted and actual Mel-cepstral features of speech, PESQ evaluates perceived speech quality and MOS is a subjective listener rating of speech quality. An instance of the Text2Speech class by loading a pre-trained model with specific parameters as shown in Fig. 1.:-

- **threshold** defines the attention weight threshold for the model.

- **minlenratio** and **maxlenratio** controls the length of the generated speech relative to the input text, allowing for variable length outputs.

- **use-att-constraint**, **backward-window** and **forward-window** are used to modify attention constraints during the alignment process.

- **speed-control-alpha** adjusts the speaking speed.

- **noise-scale** and **noise-scale-dur** influence the variability of the generated speech in terms of vocal tone and duration.

as shown in Fig. 1. Then, the input sentence is converted into speech using a pre-trained text-to-speech (TTS) model from the ESPnet2 framework. The TTS model processes the sentence and generates a corresponding audio waveform. For evaluation we are using Mel Cepstral Distortion (MCD), Perceptual Evaluation of Speech Quality (PESQ) and Mean Opinion Score (MOS) on "keithito/lj-speech" dataset.

```
tag = "espnet/kan-bayashi_ljspeech_vits"
vocoder_tag = None

text2speech = Text2Speech.from_pretrained(
    model_tag=tag,
    vocoder_tag=None,
    device="cuda" if torch.cuda.is_available() else "cpu",
    threshold=0.5,
    minlenratio=0.0,
    maxlenratio=10.0,
    use_att_constraint=False,
    backward_window=1,
    forward_window=3,
    speed_control_alpha=1.0,
    noise_scale=0.333,
    noise_scale_dur=0.333,
)
```

Fig. 1. TTS Inference Object

```
Sample 1: MCD = 65.41, PESQ = 1.26, MOS = 1.04
Sample 2: MCD = 42.95, PESQ = 1.22, MOS = 1.04
Sample 3: MCD = 48.98, PESQ = 1.19, MOS = 1.04
Sample 4: MCD = 61.43, PESQ = 1.15, MOS = 1.04
Sample 5: MCD = 46.82, PESQ = 1.28, MOS = 1.04
Sample 6: MCD = 56.91, PESQ = 1.19, MOS = 1.04
Sample 7: MCD = 53.59, PESQ = 1.14, MOS = 1.04
Sample 8: MCD = 49.17, PESQ = 1.45, MOS = 1.05
Sample 9: MCD = 61.11, PESQ = 1.23, MOS = 1.04
Sample 10: MCD = 64.66, PESQ = 1.24, MOS = 1.04
Sample 11: MCD = 45.94, PESQ = 1.33, MOS = 1.05
Sample 12: MCD = 53.99, PESQ = 1.17, MOS = 1.04
Sample 13: MCD = 65.57, PESQ = 1.15, MOS = 1.04
Sample 14: MCD = 67.40, PESQ = 1.21, MOS = 1.04
Sample 15: MCD = 50.68, PESQ = 1.19, MOS = 1.04
Sample 16: MCD = 50.90, PESQ = 1.21, MOS = 1.04
Sample 17: MCD = 46.87, PESQ = 1.23, MOS = 1.04
Sample 18: MCD = 50.30, PESQ = 1.21, MOS = 1.04
Sample 19: MCD = 72.33, PESQ = 1.16, MOS = 1.04
Sample 20: MCD = 65.10, PESQ = 1.24, MOS = 1.04
Sample 21: MCD = 59.62, PESQ = 1.24, MOS = 1.04
Sample 22: MCD = 51.50, PESQ = 1.29, MOS = 1.05
Sample 23: MCD = 62.75, PESQ = 1.29, MOS = 1.05
Sample 24: MCD = 59.33, PESQ = 1.20, MOS = 1.04
Sample 25: MCD = 56.11, PESQ = 1.17, MOS = 1.04
...
Sample 2619: MCD = 45.38, PESQ = 1.37, MOS = 1.05
Sample 2620: MCD = 56.13, PESQ = 1.30, MOS = 1.05

Average MCD: 55.72, Average PESQ: 1.25, Average MOS: 1.04
```

Fig. 2. Results Obtained

B. Results

For evaluation metric chosen are MOS, MCD and PESQ. Below in Fig. 2. are the results we got:

- MCD: Results varied between 42.95 and 72.33, with an average of 55.72, indicating room for improvement in matching synthetic speech to real speech.

- PESQ: Scores ranged from 1.14 to 1.45, with an average of 1.25, suggesting noticeable artifacts and unnaturalness in synthesized speech.

- MOS: The values were consistently low, averaging around 1.04-1.05, showing that listeners rated the speech quality poorly.

The overall performance of the TTS model indicates significant scope for improvement in reducing artifacts and enhancing naturalness, prosody and emotion in speech synthesis.

Story No.	No. of Sentences	Meta Llama (%)	Gemma 2B (%)	MistralAI (%)
1	85	96	78	85
2	110	94	76	84
3	100	95	77	85
4	90	97	79	86
5	115	93	75	83
Average	**100**	**95**	**77**	**84**

Fig. 3. Table 1

IV. LLMs for Story Segmentation and Annotation

A. Evaluation

The second evaluation focused on LLMs performance in segmenting stories into lines, identifying speakers and annotating emotions. Three LLMs were evaluated: Meta Llama 3.1, Google Gemma 2B and MistalAI 8x22B.

Dataset

Five stories were selected for evaluation, each containing between 80 and 120 sentences. The ground truth dataset included:

• **Sentence Segmentation:** The correct division of the story into individual lines.

• **Speaker Identification:** Attribution of each line to the correct speaker.

• **Emotion Annotation:** Correct labeling of the expressed emotions from the 7 root emotions: [sadness, happiness, fear, anger, surprise, disgust and neutral]

B. Results

Evaluation Techniques

Due to variations in sentence segmentation by different models, traditional word-to-word accuracy metrics were insufficient. We employed the following method to evaluate accuracy:

LLM-assisted Evaluation:

○ Leveraged reliable LLMs to compare outputs with the ground truth.

○ These LLMs aligned the model outputs semantically with the ground truth.

○ Addressed discrepancies arising from different sentence structures with manual evaluation.

Results

Table 1: Accuracy of LLMs on Story Segmentation and Annotation

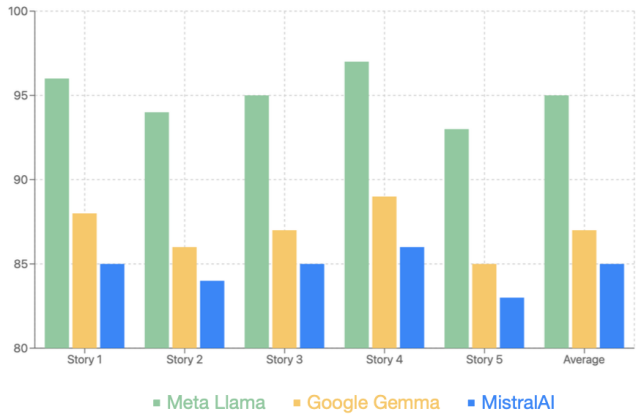

Fig. 4. Comparison Graph

REFERENCES

[1] ESPnet: End to End Speech Processing Toolkit by Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, Tsubasa Ochiai *https://arxiv.org/abs/1804.00015*.