Task 2

Fine-tuning SpeechT5 TTS for a Regional Language -- Hindi

-by Tejasva Maurya

1. Introduction

SpeechT5 is a versatile model from Microsoft that integrates text-to-speech (TTS), automatic speech recognition (ASR), and other speech-related tasks into a unified framework. It leverages both text and speech embeddings, allowing for high-quality speech generation.

This project involves fine-tuning a pre-trained SpeechT5 model on Hindi audio data, normalizing the text, and extracting speaker embeddings to enhance the quality of generated speech. The model's performance will be assessed using objective metrics like inference time and subjective measures such as the Mean Opinion Score (MOS).

2. Dataset Description:

The Hindi dataset in Mozilla Foundation's **Common Voice 17.0** is a comprehensive, crowd-sourced collection of speech data designed to support the development of open-source speech recognition models. The dataset features high-quality recordings contributed by volunteers, ensuring a diverse representation of accents, dialects, and speaking styles within the Hindi language.

The dataset contains audio samples paired with corresponding transcriptions, with each clip representing a range of sentences in Hindi. This diversity aids in creating robust models that can cater to different variations of spoken Hindi, contributing to the inclusiveness of voice technology.

Dataset :- https://huggingface.co/datasets/mozilla-foundation/common_voice_17_0/viewer/hi

Key Features:

- **Language:** Hindi
- **Audio Data:** Thousands of recordings in diverse accents and dialects
- **Text Data:** Accurate transcriptions paired with each audio clip
- **Data Format:** Audio files in .mp3 format and text files in .txt format
- **Use Cases:** Ideal for speech-to-text (ASR) and text-to-speech (TTS) applications, language modelling, and linguistic research.

3. Data Preprocessing Steps

1. **Audio Sampling:** The audio files in the dataset are cast to a uniform sampling rate of 16,000 Hz using the datasets library and the Audio module.
2. **Text Tokenization:** The text data is tokenized using a pre-trained tokenizer (SpeechT5Processor). This tokenizer splits the text into tokens, which are later used to train the model.
3. **Character Extraction:** A function is used to extract all unique characters from the dataset. This helps in building a custom vocabulary for the model to better handle the text input.
4. **Text Normalization:** The is applied to clean and standardize the text data.
5. **Vocabulary Building:** The dataset is mapped using the extracted unique characters to create a vocabulary. This vocabulary helps to identify and address any missing characters or discrepancies between the dataset vocabulary and the tokenizer's vocabulary.

4. Model Hyperparameters

The following hyperparameters were used during training:

- learning_rate: 0.0001
- train_batch_size: 4
- eval_batch_size: 2
- seed: 42
- gradient_accumulation_steps: 8
- total_train_batch_size: 32
- optimizer: Adam with betas=(0.9,0.999) and epsilon=1e-08
- lr_scheduler_type: linear
- lr_scheduler_warmup_steps: 100
- training_steps: 1500
- mixed_precision_training: Native AMP

5. Training and Fine-Tuning

The fine-tuning process was carried out using Hugging Face's `Seq2SeqTrainer` class. The `SpeechT5Processor` was used to tokenize both the text and audio inputs. Gradient checkpointing was enabled to save memory, and the model was trained using mixed precision for efficiency.

The model's performance was tracked using evaluation metrics such as loss and MOS (Mean Opinion Score) calculated from native speakers' feedback on generated speech samples. The training logs captured key performance indicators throughout the training process.

Training Logs :-

Training Loss	Epoch	Step	Validation Loss
0.6856	0.3442	100	0.5976
0.5929	0.6885	200	0.5453
0.5554	1.0327	300	0.5130
0.5407	1.3769	400	0.5052
0.5318	1.7212	500	0.4847
0.5213	2.0654	600	0.4796
0.514	2.4096	700	0.4728
0.5065	2.7539	800	0.4703
0.5046	3.0981	900	0.4684
0.4976	3.4423	1000	0.4621
0.4929	3.7866	1100	0.4583
0.4791	4.1308	1200	0.4550
0.4823	4.4750	1300	0.4529
0.485	4.8193	1400	0.4506

0.4774	5.1635	1500	0.4524

6. Performance Evaluation

The performance of the fine-tuned Hindi Text-to-Speech (TTS) model was assessed using both **objective** and **subjective** evaluation metrics to ensure a comprehensive analysis of the model's quality.

Objective Evaluation:

1. **Inference Time**: The time taken by the model to generate speech from text (in seconds) is measured to evaluate the model's efficiency, particularly for real-time applications.

Subjective Evaluation:

2. **Mean Opinion Score (MOS)**: This is a commonly used subjective metric where human listeners rate the quality of the synthesized speech on a scale of 1 to 5. The MOS provides an insight into how natural and intelligible the speech sounds to users.
3. **Pronunciation Accuracy**: Human evaluators are asked to assess how accurately the model reproduces specific technical terms and general vocabulary in Hindi, ensuring the speech output aligns with the language's phonetics.
4. **Naturalness and Fluency**: Listeners rate the smoothness, rhythm, and prosody of the synthesized speech to assess its naturalness in comparison to human speech.

Benchmarking: The model's performance is compared against other base model and bark tts model and baselines, including multilingual TTS models and existing Hindi TTS solutions. The comparison is based on both MOS scores.

By combining these metrics, the performance of the fine-tuned Hindi TTS model was evaluated comprehensively, ensuring that both the quality of the synthesized speech and its efficiency meet the project objectives. The evaluations reveal how well the model generalizes to various Hindi dialects while maintaining a natural-sounding voice.

7. Result :-

Objective Metrics

```
Metrics of finetuned and base model inference time
finetuned Inference Time: 9.0837 seconds
base model Inference Time: 4.5315 seconds
bark model Inference Time: 95.7600 seconds
```

Model	MOS (1-5)	Inference Times (s)
Fine-tuned SpeechT5 (Hindi)	4.3	9.0837
Base SpeechT5 (Hindi)	1.5	4.5315
Bark-Small (hindi)	4.7	95.7600

Subjective Evaluations

Clarity: The fine-tuned SpeechT5 model achieved a clarity score of 4.3, higher than base models

Pronunciation: Feedback indicated that SpeechT5 handled complex pronunciations in Hindi better, especially with aspirated consonants and vowel matras.

Naturalness: The generated speech from the SpeechT5 model was rated more natural compared to Bark-Small TTS model.

Audio Samples From Both The Pre-Trained (Bark) And Fine-Tuned Models (SpeechT5 TTS) For Comparison

Mean Opinion Score (MOS): This is a used subjective metric where human listeners rate the quality of the synthesized speech on a scale of 1 to 5.

Sentence	**Bark-Small Model**		**SpeechT5 fine-tuned Model**	
	File name in bark_TTS folder	**MOS**	**File name in hindi_TTS folder**	**MOS**
मेरा नाम अमित है।	output_bark_tts_1.wav	5	output_hindi_tts_1.wav	4.9
आज मौसम बहुत अच्छा है।	output_bark_tts_2.wav	4.9	output_hindi_tts_2.wav	4.9
मैं स्कूल जा रहा हूँ।	output_bark_tts_3.wav	2.5	output_hindi_tts_3.wav	4.5
यह किताब बहुत रोचक है।	output_bark_tts_4.wav	2	output_hindi_tts_4.wav	4
क्या तुमने खाना खाया?	output_bark_tts_5.wav	2	output_hindi_tts_5.wav	5
उसे हिंदी बोलनी आती है।	output_bark_tts_6.wav	2	output_hindi_tts_6.wav	5
तुम कहाँ रहते हो?	output_bark_tts_7.wav	2	output_hindi_tts_7.wav	4.6
मैं ठीक हूँ, धन्यवाद।	output_bark_tts_8.wav	5	output_hindi_tts_8.wav	4
यह मेरा पसंदीदा गाना है।	output_bark_tts_9.wav	3.6	output_hindi_tts_9.wav	4.3
क्या आप मुझे मदद कर सकते हैं?	output_bark_tts_10.wav	3.6	output_hindi_tts_10.wav	3
चाय तैयार है।	output_bark_tts_11.wav	2.5	output_hindi_tts_11.wav	3.2
मेरे पास एक कुत्ता है।	output_bark_tts_12.wav	3	output_hindi_tts_12.wav	3.2
हम कल फिल्म देखने जाएंगे।	output_bark_tts_13.wav	3	output_hindi_tts_13.wav	4
वह स्कूल से लौट रहा है।	output_bark_tts_14.wav	5	output_hindi_tts_14.wav	4
यह बहुत मज़ेदार है।	output_bark_tts_15.wav	2	output_hindi_tts_15.wav	5
समय बहुत महत्वपूर्ण है।	output_bark_tts_16.wav	3.6	output_hindi_tts_16.wav	5
मैं जल्दी घर पहुँच जाऊँगा।	output_bark_tts_17.wav	3.6	output_hindi_tts_17.wav	5
वह एक अच्छा विद्यार्थी है।	output_bark_tts_18.wav	3.5	output_hindi_tts_18.wav	4.8
हमें शांति से काम करना चाहिए।	output_bark_tts_19.wav	3.5	output_hindi_tts_19.wav	4.8

यह किताब किसकी है?	output_bark_tts_20.wav	3.2	output_hindi_tts_20.wav	5
Average MOS :-		**3.095**		**4.410**

Conclusion

This project successfully fine-tuned the SpeechT5 model for Hindi text-to-speech tasks using the Mozilla Common Voice Hindi dataset. The model was evaluated against existing Hindi TTS models, demonstrating improvements in both objective metrics (MOS and inference time) and subjective evaluations from native Hindi speakers. The fine-tuned model showed particular strengths in clarity, pronunciation, and naturalness of speech.