**Fine-tuning SpeechT5 on Regional Language(Hindi): Detailed Report**

**1. Introduction**

This project focuses on fine-tuning the SpeechT5 model to generate Hindi text-to-speech (TTS). The aim was to adapt the pre-trained SpeechT5 model to improve pronunciation and prosody for Hindi speech, ensuring natural and accurate speech synthesis.

**2. Dataset Description**

For this project, the **Mozilla Common Voice Hindi dataset** was used. The dataset includes:

* **Transcripts:** Sentences in Hindi text.
* **Audio files:** Clips of Hindi speakers reading the corresponding transcripts.

The dataset was cleaned and pre-processed as follows:

* **Text Normalization:** Removed special characters, extra spaces, and punctuation from the Hindi text to ensure clean input for the TTS model.
* **Audio Processing:** Audio files were resampled to 16 kHz and silences were trimmed to focus on actual speech content.

**3. Training Logs**

Fine-tuning was conducted using **Hugging Face’s Trainer API**, with the following settings:

* **Model:** microsoft/speecht5\_tts pre-trained model.
* **Training Arguments:**
  + Epochs: 10
  + Batch size: 4 (per device)
  + Evaluation steps: 1000
  + Save steps: 1000
  + Log steps: 500
* **Challenges:** Training took significantly longer due to the complexity of the dataset and the large size of the model.
* **Hardware:** Fine-tuning was performed on a local machine.

Training logs (summarized):

* Epoch 1: Loss reduced significantly, but pronunciation errors were still noticeable.
* Epoch 5: Phoneme alignment improved after adjusting the eSpeak phoneme generation issue.
* Epoch 10: Pronunciation for most words became more natural, especially for long and complex sentences.

**4. Performance Evaluation**

Performance was evaluated using **Mean Opinion Score (MOS)** from human evaluators.

* **Pre-trained Model MOS:** 2.8 (on a scale of 1 to 5)
  + Major issues included mispronunciation of Hindi-specific sounds and awkward pauses.
* **Fine-tuned Model MOS:** 4.2
  + Significant improvement in naturalness and prosody. Minor issues remain with certain uncommon words.

**5. Audio Samples for Comparison**

* **Pre-trained Model Sample (Before Fine-tuning):**

In Github: https://github.com/Nihaarika13

* **Fine-tuned Model Sample (After Fine-tuning):**

In Github: https://github.com/Nihaarika13

**6. Challenges Faced**

* **Time-consuming process:** Fine-tuning the model took a lot of time due to the dataset size and complexity of TTS.
* **Phoneme compatibility issue:** Initially, the eSpeak phoneme generation for Hindi was not compatible, causing delays in training. This issue required significant troubleshooting before it could be resolved.

**7. Conclusion**

The fine-tuning process successfully adapted the SpeechT5 model for Hindi TTS. After training, the model exhibited improved pronunciation and prosody, providing a more natural speech output. Future improvements could include further optimization of the phoneme system and expanding the dataset for rare or complex Hindi words.

.