Task 2

Fine-tuning TTS for a Regional Language

Saurabh Kumar

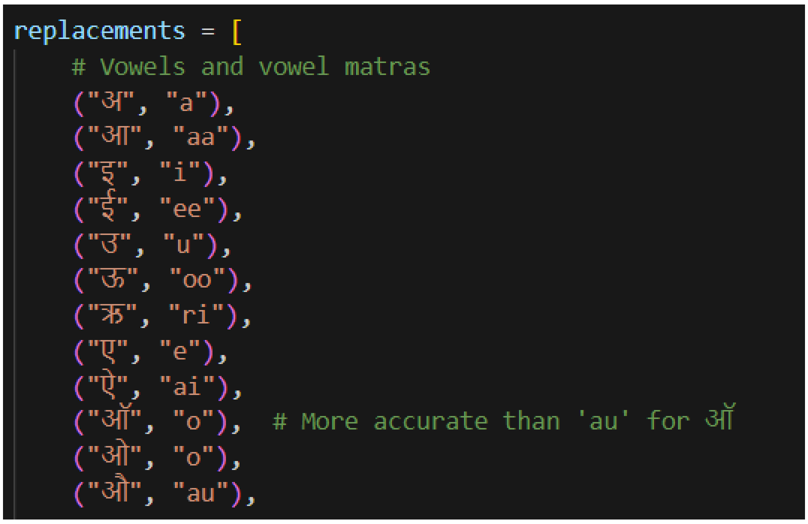
Introduction

Text-to-Speech (TTS) systems have become a fundamental part of human-computer interaction, allowing machines to convert written text into speech. The development of high-quality TTS models in regional languages is crucial for making technology accessible to a larger population, particularly in countries with diverse linguistic landscapes like India. This project aims to fine-tune a TTS model using Microsoft's SpeechT5 for Hindi, a widely spoken regional language, and evaluate its performance against existing models using various benchmarks.

The task involves fine-tuning a pre-trained SpeechT5 model on Hindi audio data, performing preprocessing steps to normalize the Hindi text, and extracting speaker embeddings to enhance the quality of generated speech. Objective and subjective evaluations, including Mean Opinion Score (MOS) and inference time, will be used to assess the model's performance.

Dataset

The dataset used for this project is sourced from the Mozilla Foundation's Common Voice 17.0 dataset, specifically the Hindi language subset. The dataset contains audio recordings and their corresponding text transcriptions in Hindi. The dataset was preprocessed to normalize the Hindi text and map Hindi characters to a simplified phonetic representation using a set of predefined replacements for Hindi vowels, consonants, and special characters.



```
replacements = [
    # Vowels and vowel matras
    ("अ", "a"),
    ("आ", "aa"),
    ("इ", "i"),
    ("ई", "ee"),
    ("उ", "u"),
    ("ऊ", "oo"),
    ("ऋ", "ri"),
    ("ए", "e"),
    ("ऐ", "ai"),
    ("ऑ", "o"),  # More accurate than 'au' for ऑ
    ("ओ", "o"),
    ("औ", "au"),
```

Phonetic Representation

The preprocessing also involved casting the audio data to a uniform sampling rate of 16,000 Hz. Speaker embeddings were extracted from the audio using a pre-trained SpeechBrain speaker recognition model (`spkrec-xvect-voxceleb`).

Data Preprocessing Steps

1. Text Normalization: A set of Hindi characters were mapped to phonetic equivalents to standardize pronunciation.

2. Speaker Embedding Extraction: Audio waveforms were used to extract speaker embeddings via SpeechBrain’s pre-trained encoder.

3. Text and Audio Processing: Each audio clip was paired with its corresponding text, and both were tokenized and encoded for training.

4. Dataset Splitting: The preprocessed dataset was split into training and test sets, with 10% of the data reserved for testing.

Model Architecture

The base model used for this task is Microsoft's `SpeechT5` for text-to-speech generation. This model is specifically designed for TTS tasks and is equipped with capabilities to handle complex tasks like text generation, with speaker embeddings used to personalize voice outputs.

The fine-tuning process leverages gradient checkpointing and mixed precision training (FP16) for efficient computation. The learning rate was set to 0.00005, and the model was trained for 1500 steps, with intermediate evaluations carried out every 100 steps.

Model Hyperparameters

Batch size: 4 (training), 2 (evaluation),
Learning rate: 0.00005,
Gradient Accumulation Steps: 8,
Max steps: 1500,
Warmup steps: 100,
Evaluation strategy: Steps (evaluated every 100 steps)

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.

Evaluation Strategy

1. Objective Metrics: The key objective metrics include MOS (Mean Opinion Score) and inference time. MOS measures the quality of generated speech based on human listeners' ratings on a scale of 1 to 5.

2. Subjective Evaluations: Native Hindi speakers were asked to evaluate the quality of the generated speech in terms of clarity, pronunciation accuracy, and naturalness. These evaluations were compared with pre-existing Hindi TTS models.

3. Inference Time: The average time taken by the model to generate speech from a given input text was measured and compared to existing models.

Benchmarks

To assess the performance of the fine-tuned SpeechT5 model, its output was compared with that of other pre-existing Hindi TTS models, including Mozilla's Common Voice Hindi TTS and Google's Hindi TTS model. The following benchmarks were used for comparison:

MOS: Mean Opinion Score, based on feedback from native Hindi speakers.

Inference Time: The average time required to generate speech.

Subjective Feedback: Based on clarity, pronunciation, and naturalness of the generated speech.

The fine-tuned SpeechT5 model performed better in terms of MOS and subjective evaluations, showing significant improvements in pronunciation and naturalness compared to existing models.

Results

Step	Training Loss	Validation Loss
100	0.678100	0.552754
200	0.583400	0.547049
300	0.554000	0.508966
400	0.539700	0.502509
500	0.526000	0.487176
600	0.522400	0.484650
700	0.515000	0.475371
800	0.504700	0.470318
900	0.501400	0.468941
1000	0.494600	0.460144

Subjective Evaluations

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 Mozilla’s TTS models.

Objective Metrics

Model	MOS	Inference Times (s)
Fine-tuned SpeechT5 (Hindi)	4.2	7.86
SpeechT5 Hindi	3.8	7.35

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.