Final Report: Fine-tuning Text-to-Speech (TTS) Models for English Technical Speech and Regional Languages

Introduction

Text-to-Speech (TTS) systems are essential in various applications such as virtual assistants, audiobooks, and accessibility solutions. Fine-tuning pre-trained TTS models is critical for improving their performance in specific use cases, like technical speech synthesis or regional language adaptation. This report outlines the process of fine-tuning two TTS models. The first model was optimized to handle technical jargon used in English technical interviews, and the second model was fine-tuned for a regional language.

Methodology

The fine-tuning process involved selecting appropriate base models, preparing datasets, and adjusting the models' hyperparameters for optimal performance. Two tasks were undertaken:

1. English Technical Speech Fine-Tuning: Coqui TTS(XTTS model) was chosen as the base model due to its flexibility. A custom dataset containing general English sentences and technical terms like 'API', 'CUDA' was created for fine-tuning.

2. Regional Language Fine-Tuning: Coqui TTS was selected for fine-tuning on the regional language. A dataset of natural language sentences covering a wide range of phonemes was sourced to train the model.

Results

The fine-tuned models were evaluated using subjective feedback:

1. English Technical Speech Model: The model demonstrated improved pronunciation of technical terms

2. Regional Language Model: Native speakers found the speech to be natural and intelligible.

Challenges

Several challenges were encountered during the fine-tuning process:

1. Dataset Issues: For technical English, gathering enough examples of technical terms in natural speech was challenging, requiring collection of data from YouTube using few python scripts.

2. Model Convergence: Fine-tuning the regional language model required careful tuning of hyperparameters to ensure convergence without overfitting, given the limited dataset size.

Bonus Task: Fast Inference Optimization

Model quantization techniques were explored to reduce model size and increase inference speed. Post-Training Quantization (PTQ) was applied to the Hindi technical speech model, reducing the model size by 64.9%, but this model Is not loading properly and hence, not used in inference. Whereas trimming was applied to the model, which reduced its size by 7.2%, and the model was successfully loaded and used for inference.while maintaining the output speech quality.

Conclusion

Fine-tuning the TTS models significantly improved their performance in both technical English speech and regional language synthesis. The results demonstrate that careful dataset preparation and model adjustment can yield high-quality outputs.

Future improvements could include expanding the dataset for technical terms and exploring more aggressive optimization techniques like pruning or distillation to further enhance model efficiency.