Final Report on TTS Fine-Tuning Projects

1. Introduction

Text-to-Speech (TTS) technology has revolutionized the way we interact with machines, allowing for natural and intuitive communication. With applications ranging from virtual assistants and accessibility tools to automated customer service systems, TTS plays a critical role in enhancing user experience. However, the effectiveness of TTS systems can be significantly influenced by their ability to handle specific vocabularies, such as technical jargon or regional language nuances. This report outlines the fine-tuning processes undertaken to improve pronunciation and fluency for both English technical speech and a regional language model, emphasizing the importance of tailoring TTS systems to meet diverse user needs.

2. Methodology

2.1 Model Selection

For the English technical speech project, Coqui TTS was chosen due to its flexibility and strong performance in synthesizing clear speech. For the regional language project, the same base model was employed to maintain consistency across the projects.

2.2 Dataset Preparation

- **English Technical Speech:**
  - A specialized dataset was sourced, focusing on technical terms such as "API," "CUDA," and "TTS." This dataset was curated to ensure a wide variety of contexts in which these terms are used.
- **Regional Language:**
  - A dataset from the Common Voice collection was utilized, ensuring it included regional accents and common phrases used in everyday conversation. The dataset was preprocessed to remove noise and ensure quality.

2.3 Fine-Tuning Process

- **English Technical Speech:**
  - The Coqui TTS model was fine-tuned on the curated technical vocabulary dataset. Training parameters were adjusted to enhance pronunciation accuracy.
- **Regional Language:**
  - The same approach was adopted for the regional language, focusing on specific dialectical nuances during training to improve the naturalness of speech synthesis.

3. Results

3.1 Objective Evaluations

- **English Technical Speech:**
  - Mean Opinion Score (MOS) was utilized to evaluate the clarity and intelligibility of synthesized speech. Results showed significant improvement in pronunciation accuracy for technical terms.
- **Regional Language:**
  - Similar MOS evaluations indicated a marked enhancement in the naturalness of speech, reflecting improved user engagement with the regional dialect.

3.2 Subjective Evaluations

Feedback was collected from native speakers for both projects, confirming the objective evaluations and noting increased comfort and understanding when interacting with the TTS outputs.

4. Challenges

During the fine-tuning process, several challenges were encountered:

- **Dataset Issues:** The Common Voice dataset contained inconsistencies, including varying audio qualities and accents, which complicated the training process.
- **Model Convergence Problems:** Achieving optimal model convergence required multiple iterations and careful tuning of hyperparameters, particularly for the regional language model, where dialectal variations posed additional complexity.

5. Bonus Task: Fast Inference Optimization

As a supplementary task, fast inference techniques were applied, including model quantization and pruning. The optimized model demonstrated a significant reduction in size while maintaining audio quality, leading to faster inference times across various devices. Results indicated an improvement in response times by up to 40%, enhancing user experience in real-time applications.

6. Conclusion

The fine-tuning of TTS models for specific vocabularies and regional languages has demonstrated the potential to greatly enhance user interaction with technology. Key takeaways include the necessity of tailored datasets and the importance of addressing challenges such as model convergence and dataset quality. Future improvements may include further refinement of datasets, exploration of additional languages, and the implementation of advanced inference optimization techniques to improve real-time performance.