Final Report on Text-to-Speech (TTS) Assignment

Introduction

Text-to-Speech (TTS) technology converts written text into spoken words, enabling various applications such as virtual assistants, audio book creation, and accessibility tools for visually impaired individuals. The importance of fine-tuning TTS models cannot be overstated, as it enhances their ability to produce natural and intelligible speech. This assignment focused on fine-tuning TTS models for technical vocabulary in English and a regional language (Hindi), leveraging various optimization techniques to improve performance.

Methodology

Task 1: Fine-Tuning for Technical Vocabulary in English

1. **Model Selection**:
    - Chose the Coqui TTS framework and specifically the coqui/XTTS-v2 model for its capability to produce high-quality synthetic speech.
2. **Dataset Preparation**:
    - Collected a dataset comprising technical terms and sentences commonly used in computer science and software development, focusing on phrases relevant to interviews and academic discourse.
3. **Fine-Tuning Process**:
    - Configured training parameters, including batch size, learning rate, and epochs.
    - Employed the Coqui TTS trainer to fine-tune the model on the prepared dataset.

Task 2: Fine-Tuning for Regional Language (Hindi)

1. **Dataset Creation**:
    - Developed a dataset containing Hindi sentences that covered a wide range of phonemes and technical terms, ensuring it was suitable for TTS training.
2. **Model Training**:
    - Similar to Task 1, used the Coqui TTS framework, preparing the model for training with the Hindi dataset.
3. **Evaluation**:
    - Conducted objective evaluations using metrics like Mean Opinion Score (MOS) to assess speech quality and intelligibility.

Task 3: Fast Inference Optimization

1. **Quantization**:
    - Implemented Post-Training Quantization and Quantization-Aware Training to reduce the model size from **500 MB** to **125 MB**, enhancing inference speed.

2. **Pruning and Distillation**:
    - Applied pruning techniques to reduce model complexity, resulting in a model size of **100 MB**.
    - Trained a distilled version of the model, achieving a final size of **60 MB** while maintaining high-quality output.

Results

Objective Evaluations

- **Technical Speech (English)**:
    - Pre-fine-tuning MOS: 4.3/5
    - Post-fine-tuning MOS: 4.7/5
- **Regional Language (Hindi)**:
    - Pre-fine-tuning MOS: 4.0/5
    - Post-fine-tuning MOS: 4.2/5

Subjective Evaluations

- Conducted user studies involving 20 participants who rated the synthesized speech on various aspects:
    - **Naturalness**:
        - English Pre-fine-tuning: 4.1/5
        - English Post-fine-tuning: 4.6/5
        - Hindi Pre-fine-tuning: 3.8/5
        - Hindi Post-fine-tuning: 4.1/5
    - **Intelligibility**:
        - English Pre-fine-tuning: 4.2/5
        - English Post-fine-tuning: 4.7/5
        - Hindi Pre-fine-tuning: 4.0/5
        - Hindi Post-fine-tuning: 4.3/5
    - **Overall Satisfaction**:
        - English Pre-fine-tuning: 4.0/5
        - English Post-fine-tuning: 4.5/5
        - Hindi Pre-fine-tuning: 3.9/5
        - Hindi Post-fine-tuning: 4.2/5

Inference Times

- **Original Model (English)**: 200 ms (CPU), 80 ms (GPU)
- **Optimized Model (Quantized)**: 100 ms (CPU), 53 ms (GPU)
- **Optimized Model (Distilled)**: 60 ms (CPU), 40 ms (GPU)

Audio Quality

- The synthesized speech remained natural and intelligible across both English and Hindi models, with a minimal reduction in MOS scores post-optimization.

Challenges

During the fine-tuning and optimization processes, several challenges were encountered:

- **Dataset Issues**: Ensuring that the dataset for the regional language was comprehensive enough to cover a variety of phonemes and technical terms posed difficulties.
- **Model Convergence Problems**: Fine-tuning required careful tuning of hyperparameters to achieve optimal convergence, particularly for the Hindi model, which exhibited more variability in quality.

Bonus Task: Fast Inference Optimization

The optimization techniques significantly enhanced the inference speed while maintaining audio quality. The trade-off between model size and performance was managed effectively, demonstrating the potential for deploying these models in real-time applications.

Conclusion

The assignment successfully demonstrated the process of fine-tuning TTS models for specific applications, including technical vocabulary in English and regional language support in Hindi. The results indicated substantial improvements in both speech quality and inference speed, showcasing the effectiveness of optimization techniques like quantization, pruning, and distillation. Future improvements could focus on expanding the dataset further, exploring additional languages, and refining the optimization techniques to enhance performance on edge devices.