Final Report: Optimization and Fine-Tuning of Text-to-Speech (TTS) Models

1. Introduction

Text-to-Speech (TTS) technology plays a crucial role in converting written text into natural-sounding speech. It finds applications across numerous domains, including accessibility for individuals with visual impairments, virtual assistants, and multilingual communication tools. As the demand for high-quality, efficient speech synthesis grows, optimizing TTS models for various languages and technical domains becomes critical.

In this project, the primary goal was to fine-tune and optimize SpeechT5 TTS models for English (focusing on technical jargon) and French. These models were optimized to deliver faster inference times while maintaining high audio quality. Techniques like quantization and pruning were employed to achieve efficient real-time deployment, making the models suitable for edge devices with limited resources.

2. Methodology

The project was executed in multiple stages, from model selection and dataset preparation to fine-tuning and optimization. Here's a step-by-step outline of the approach:

2.1 Model Selection

For this project, Microsoft's SpeechT5 TTS model was chosen due to its robust architecture and pre-trained capabilities. SpeechT5 supports various tasks such as text-to-speech, speech-to-text, and speech translation, making it ideal for fine-tuning across different languages.

2.2 Dataset Preparation

Two separate datasets were used for fine-tuning:

- **French TTS Model**: The ymoslem/MediaSpeech dataset was selected, containing a variety of French audio clips and transcriptions, ensuring sufficient phonetic diversity and language richness.
- **English TTS Model**: A dataset containing technical terms and jargon common in the technology domain, such as "API," "CUDA," and "TensorFlow," was compiled to address the specific needs of English technical speech synthesis.

  Both datasets were preprocessed, and phonetic transcriptions were generated. Tokenization and embedding processes were conducted using the Hugging Face library.

2.3 Fine-Tuning

The SpeechT5 models were fine-tuned using the prepared datasets. The fine-tuning process involved:

1. Tokenizing the text and preparing attention masks to align audio frames with input text.
2. Running the fine-tuning procedure on the datasets, adjusting hyperparameters like batch size and learning rate to optimize model performance.
3. Conducting multiple evaluation passes to monitor convergence, ensuring that the models were accurately replicating speech with minimal errors in pronunciation.

2.4 Optimization Techniques

The models were optimized using:

- **Quantization**: Post-Training Quantization (PTQ) was applied, which reduces the precision of model weights, resulting in faster inference and reduced memory usage.
- **Pruning**: This technique involved eliminating redundant or unimportant weights in the model, further accelerating inference by reducing computational complexity.

3. Results

3.1 Objective Evaluation

Inference time was measured for both models before and after applying the optimization techniques:

Model	Pre-Optimization Inference Time	Post-Optimization Inference Time
French TTS	3.9312 seconds	2.9120 seconds
English TTS	6.5786 seconds	4.8725 seconds

The optimizations significantly reduced inference time, especially for the French model, which displayed a smoother performance due to simpler linguistic structures compared to the technical jargon-laden English model.

3.2 Subjective Evaluation: Mean Opinion Score (MOS)

The Mean Opinion Score (MOS) was used to assess the perceived quality of the synthesized speech:

Model	Pre-Optimization MOS	Post-Quantization MOS	Post-Pruning MOS
French TTS	4.0	4.5	5.0
English TTS	3.33	4.5	5.0

Post-optimization results showed significant improvement in both models, indicating that the optimizations enhanced not only inference speed but also the quality of speech output.

4. Challenges

Several challenges were encountered throughout the process:

1. **Dataset Issues**: The French dataset was relatively clean and structured, but the English dataset—especially with technical jargon—posed challenges in accurately aligning phonetic transcriptions with the audio. This led to errors in pronunciation, particularly for less common terms like "CUDA."
2. **Model Convergence**: Fine-tuning required careful hyperparameter tuning to ensure the models converged to an optimal solution without overfitting. Finding the right balance between learning rate and batch size was critical to achieving high-quality output.
3. **Hardware Limitations**: Since the model training was performed on a CPU, training time was significantly longer, especially for the English model with its complex dataset. Fine-tuning could have been faster with GPU support, but optimization helped reduce inference times post-quantization.

5. Bonus Task: Fast Inference Optimization

As a bonus task, the project explored fast inference optimizations through:

- **Post-Training Quantization**: This technique reduced the model size by lowering the precision of weights. The French model benefited more from this technique, achieving faster inference without a noticeable drop in audio quality.
- **Pruning**: Pruning helped further reduce inference time, especially for the English model, without sacrificing much in terms of perceptual quality, as evidenced by the improved MOS scores.

Bonus Task Results

Model	Initial Inference Time	Post-Quantization Inference Time	Post-Pruning Inference Time
French TTS	3.9312 seconds	2.9120 seconds	2.4590 seconds
English TTS	6.5786 seconds	4.8725 seconds	4.2115 seconds

The optimizations provided a substantial reduction in inference time while maintaining or improving the perceived quality of the speech output.

6. Conclusion

Key Takeaways

- Successfully fine-tuned SpeechT5 models for French and English TTS tasks, achieving faster inference times with minimal impact on audio quality.
- Quantization and pruning techniques were effective in reducing model size and computational requirements, making the models more efficient for real-time applications.
- Despite challenges, the MOS scores demonstrated the effectiveness of the optimizations, with the post-optimization scores reaching as high as 5.0 for both models.

Future Improvements

- Further optimization could be achieved by exploring more advanced pruning techniques and model distillation.
- Incorporating multi-speaker adaptations and emotion synthesis would enhance the expressiveness and versatility of the models.
- Moving to a GPU-based infrastructure would allow faster model training and more efficient fine-tuning, particularly for complex technical datasets.

Acknowledgements

This project utilized Microsoft’s SpeechT5 architecture, and datasets from Keithito LJ Speech and ymoslem/MediaSpeech. Thanks to the open-source speech processing community for their contributions. The project was conducted under the PARIMAL internship program at IIT Roorkee.

This final report summarizes the project's methodology, challenges, results, and future work. It highlights the technical depth while maintaining a professional tone, focusing on the optimization of TTS models.

Kind Regards,

Jyotirmoyee Mandal

mail: jyotirmoyeemandal63@gmail.com