Voice Cloning – Training Report

LLM & Generative AI Project – Phone Restaurant Assistant

Model repository: https://huggingface.co/VianLB/coqui_tts_fine-tuned

1. Overview

This document summarizes the training process of a custom text-to-speech (TTS) voice model developed as part of an LLM and Generative AI school project. The objective was to fine-tune a high-quality voice cloning model suitable for a phone-based restaurant assistant, capable of natural conversational speech in all the common languages.

2. Base Model

The model is based on XTTS v2 from the Coqui TTS framework. XTTS v2 is a multilingual, GPT-based neural TTS architecture supporting voice cloning from short audio samples. The base model was fine-tuned to reproduce a custom voice with stable pronunciation, and natural intonation.

3. Training Dataset

The training dataset consists of 129 high-quality audio samples totaling 9 minutes and 44 seconds of speech. The average sample length is approximately 4.5 seconds. The dataset is bilingual, with a balanced distribution between English (63 samples) and French (66 samples), enabling natural code-switching.

The recorded content focuses on conversational phrases relevant to customer service and restaurant interactions, including greetings, confirmations, short answers, and natural dialogue patterns. All audio samples were manually transcribed to ensure high alignment quality between text and speech.

4. Training Process

Training was conducted across multiple sessions between December 2 and December 17, 2025. In total, 400 epochs were completed, representing approximately 53 hours of cumulative training time over 7 sessions.

5. Hardware and Configuration

Training was performed on a local machine equipped with an NVIDIA RTX 4050 GPU (6 GB VRAM). A batch size of 2 was used together with gradient accumulation (126 steps), resulting in an effective batch size of 252. The optimizer used was AdamW with a learning rate of 5e-6.

6. Training Performance

Validation loss decreased from 4.440 at epoch 0 to 1.234 at epoch 400, representing a total reduction of approximately 72%. Text cross-entropy loss converged towards near-zero values, indicating near-perfect text-to-phoneme alignment, while mel spectrogram loss continued to improve steadily throughout training.

7. Results and Observations

Training remained fully stable with no crashes. Validation loss continued to decrease throughout all sessions, with no evidence of overfitting or performance plateau. Later sessions showed accelerating improvements.

8. Model Deployment

The final 400-epoch model checkpoint was selected based on the lowest validation loss and published on Hugging Face for reproducibility:

https://huggingface.co/VianLB/coqui_tts_fine-tuned

9. Conclusion

The fine-tuned XTTS v2 model successfully meets the project objectives, delivering a stable, bilingual voice suitable for a phone-based restaurant assistant and ready for inference and deployment.

Last updated: January 2026