**Voice Cloning Assignment Report**

**1. Project Overview**

This project involved fine-tuning text-to-speech models for voice cloning, specifically using the Sesame CSM (1B) and Orpheus (3B) models. The goal was to clone a voice from the MrDragonFox/Elise dataset on Hugging Face.

**2. Setup and Environment**

* **Platform**: Google Colab with GPU (Tesla T4)
* **Libraries Used**:
  + Unsloth (for efficient fine-tuning)
  + Transformers (v4.52.3 for Sesame, v4.55.4 for Orpheus)
  + Soundfile, IPython.display (for audio playback)

**3. Implementation Attempts**

**a. Sesame CSM (1B) Model**

* **Status**: Successfully implemented
* **Dataset Used**: MrDragonFox/Elise from Hugging Face
* **Key Features**:
  + Single-speaker voice cloning
  + English language support
  + Lower computational requirements

**b. Orpheus (3B) Model**

* **Status**: Not implemented
* **Reason**: Unsloth library requires NVIDIA or Intel GPUs, which were not available in the current environment (message: “You cannot currently connect to a GPU due to usage limits in Colab.”)
* **Planned Features**:
  + Multi-lingual support
  + Multi-voice generation

**4. Errors Encountered**

**a. Invalid Notebook Error**

* **Error**: Missing 'state' key in 'metadata.widgets'
* **Solution**: Fixed using ‘metadata\_cleaning.py’ script to clean notebook metadata.

**b. GPU Compatibility Issue**

* **Error**: NotImplementedError: Unsloth currently only works on NVIDIA GPUs and Intel GPUs
* **Impact**: Could not run Orpheus (3B) model
* **Workaround**: Proceeded with Sesame CSM (1B) model

**c. Multilingual Support Limitation**

* **Issue**: Sesame CSM (1B) doesn't support multilingual voice generation
* **Impact**: Could not complete bonus task for different languages

**5. Time Taken**

* **Environment Setup**: ~15 minutes
* **Model Training (Sesame CSM)**: ~2 hours
* **Troubleshooting**: ~1 hour
* **Documentation**: ~1 hour

**6. GPU/Resource Requirements**

* **Sesame CSM (1B)**:
  + GPU: Tesla T4 (15GB VRAM)
  + RAM: ~12GB
  + Training Time: ~2 hours (for 1000 steps)
* **Orpheus (3B)**:
  + Requires higher-end GPU (A100/H100 recommended)
  + My Tesla T4 was out of limit usage.

**7. Results and Audio Quality**

* **Output Location**: outputs/Sesame\_CSM\_(1B)\_TTS/
* **Files**:
  + original.wav: Original voice sample
  + cloned.wav: Generated voice clone
* **Quality Assessment**:
  + The cloned voice maintains good intonation and pacing
  + Slight robotic artifacts are present
  + Pronunciation is mostly accurate but has minor inconsistencies
  + Overall quality is acceptable but could be improved with more training data

**8. Challenges Faced**

1. **Hardware Limitations**:
   * Could not run the more advanced Orpheus (3B) model
   * Limited by Tesla T4 GPU capabilities
2. **Multilingual Support**:
   * Sesame CSM (1B) only supports English
   * Could not complete the bonus task for different languages
3. **Training Data**:
   * Used a smaller dataset than recommended
   * Impact on voice quality and naturalness

**9. Future Improvements**

1. **Hardware Upgrade**:
   * Access to A100/H100 GPUs for running larger models
   * Enable Orpheus (3B) for better multilingual support
2. **Data Collection**:
   * Use the full 3-hour or more dataset for better voice quality and performance
   * Include more diverse speech samples, with multiple voices
3. **Model Optimization**:
   * Experiment with different hyperparameters
   * Try different fine-tuning approaches
4. **Multilingual Support**:
   * Implement Orpheus (3B) for multilingual capabilities
   * Include language adaptation layers

**10. Conclusion**

The project successfully implemented voice cloning system using the Sesame CSM (1B) model. While limited by hardware constraints, the implementation demonstrates the core concepts of voice cloning. The cloned voice shows promise but would benefit from more training data and computational resources for production-quality results.