Carson Sager
ECEN 5060

Final Project Proposal

For my final project, I have selected the problem of text-to-speech (TTS) synthesis, specifically focusing on generating mel-spectrograms from text input. I mainly chose this problem because my fiancée is a speech-language pathologist and I hope to one day integrate our careers in order to assist individuals with communication disorders from a technological standpoint. There is no concrete goal yet, but I believe this topic will provide a beneficial ground knowledge on speech synthesis and its potential use cases. Additionally, TTS systems are widely used in applications such as virtual assistants, audiobooks, and accessibility tools for the visually impaired. The dataset I will use is the LJSpeech dataset, which contains over 13,000 short audio clips of a single speaker reading passages from books. This dataset is large enough to train a deep neural network and is commonly used in TTS research. I will implement a combination of LSTM and CNN networks to handle both the sequential nature of the text input (via the LSTM) and the feature extraction required to produce mel-spectrograms (via the CNN). The LSTM will help capture the temporal dependencies of the input text, while the CNN will extract meaningful features from the generated intermediate representations. The framework I plan to use is PyTorch because of its flexibility, strong support for research, and the ease of experimenting with various model architectures. To inform my approach, I will refer to the papers associated with the LJSpeech dataset from Papers with Code to understand best practices and existing projects that I will use to improve parts of my network. To evaluate the performance of the model, I will use several metrics: mean squared error (MSE) to measure the difference between predicted and ground-truth mel-spectrograms, spectral distortion (SD) to assess the quality of the generated spectrograms, signal-to-noise ratio (SNR) to evaluate waveform quality, and mean opinion score (MOS) for subjective human evaluation of audio quality. Additionally, Perceptual Evaluation of Speech Quality (PESQ) will be used as an objective metric to assess the speech quality in terms of perceived naturalness and clarity. A rough schedule for the project includes:

Week of 3/31: Data Processing and Setup

Week of 4/7: Model architecture design and training

Week of 4/14: Evaluation and fine-tuning

Week of 4/21: Final testing, documentation, and report writing