Creating a Vocoder Using Machine Learning Techniques

Chaitanya Kulkarni MT24028 | **Shaurya Bhatnagar** MT24226 | **Viren Variya** MT24102 | **Priyanshu Gupta** MT24130 | **Raman Chola** MT24072

Abstract

This report presents an investigation into the development of a vocoder utilizing **machine learning techniques** for audio synthesis. The study employs spectrograms and Mel spectrograms as intermediate representations and evaluates various models, methodologies, and metrics for reconstructing high-quality audio.

Dataset Description

The dataset utilized in this study consists of 15 short audio recordings. Key characteristics of the dataset are as follows:

- Audio clip duration: 4-61 seconds
- File format: WAV
- Sampling rate: 16 kHz
- Preprocessing: Short-Time Fourier Transform (STFT) was applied to create spectrograms and Mel spectrograms

Methodology

1. Feature Extraction

The raw audio data was converted into a format suitable for modeling using the following steps:

- Mel Spectrograms: Librosa, a Python library, was employed to convert the audio data into Mel spectrograms. This representation emphasizes frequencies crucial for human hearing.
- Linear Spectrograms: The Mel spectrograms were subsequently converted into linear spectrograms using Non-Negative Least Squares (NNLS). This step helped denoise the spectrograms, enhancing their utility for the vocoder.

2. Training & Reconstruction

The spectrograms were then utilized to train the vocoder and reconstruct the audio using machine learning techniques:

- Griffin-Lim Algorithm: This algorithm was employed to estimate the missing phase information in the spectrogram. By filling in the missing phase, the Griffin-Lim algorithm facilitated the reconstruction of the audio from the spectrogram.

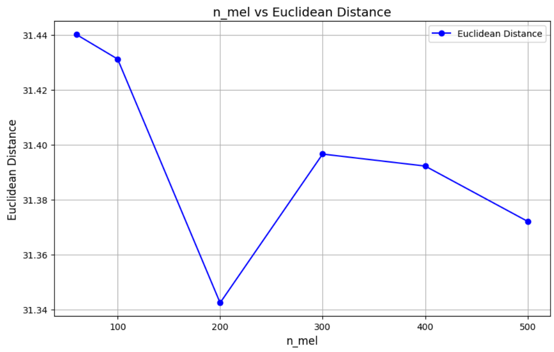
Models / Algorithm

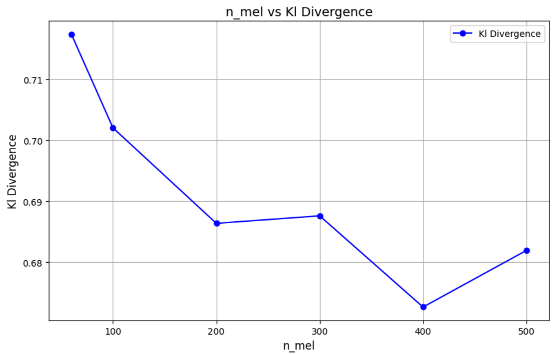
This vocoder project leveraged machine learning techniques to reconstruct the audio. The focus was on enhancing the spectrograms using the following methods:

1. **Griffin-Lim Algorithm:** This algorithm estimated the missing phase information from the spectrogram, enabling the recreation of the audio.
2. **Non-Negative Least Squares (NNLS):** This method denoised the spectrograms, rendering them more accurate for audio reconstruction.

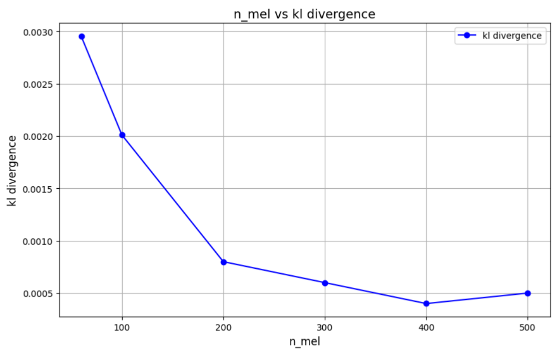
Analysis

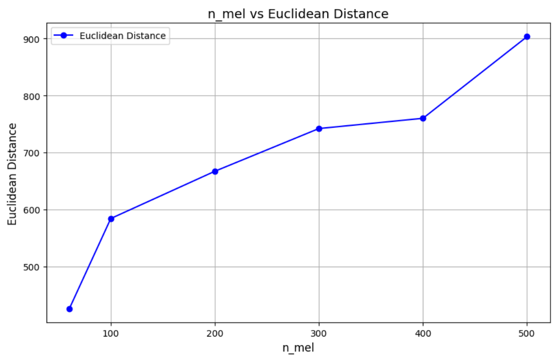
For Audio Files:





For Mel Files:





The performance of the vocoder was evaluated using both Euclidean Distance and KL Divergence, two key metrics that measure the similarity between generated and original audio or spectrograms.

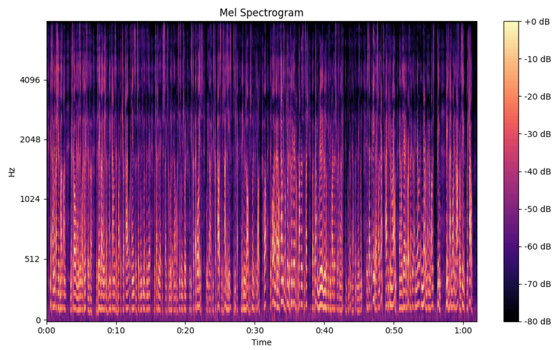
Results

The evaluation of the vocoder's performance in reconstructing audio signals from spectrograms using the Griffin-Lim algorithm and Non-Negative Least Squares (NNLS) denoising. Metrics like reconstruction quality, audio clarity, and computational efficiency were analyzed.

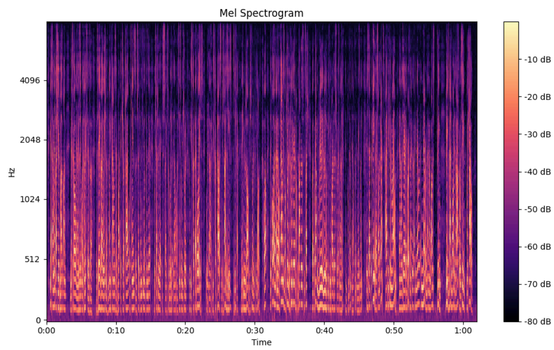
The reconstructed audio maintained high perceptual clarity with minimal distortions

Spectrogram Visualization:

Side-by-side visual comparisons of the original and reconstructed spectrograms to demonstrate the fidelity of the reconstruction process



Original Audio Mel Spectrogram



Reconstructed Audio Mel Spectrogram(500)