INDIAN INSTITUTE OF TECHNOLOGY,
JODHPUR

REPORT
SPEECH UNDERSTANDING
ASSIGNMENT - 1

Submitted By

HIMANSHU
M24CSE009

Analysis of Windowing Techniques

Click here to access all the codes

1 INTRODUCTION

The objective of this study is to analyze the impact of different windowing techniques—Hann, Hamming, and Rectangular—on spectrogram generation using the UrbanSound8K dataset. This dataset consists of urban audio recordings categorized into ten different classes. We employ the Short-Time Fourier Transform (STFT) to generate spectrograms and visually compare the effects of the selected windowing functions.

2 Windowing Techniques

2.1 Hann Window

The Hann window is defined as:

$$w(n) = 0.5 \times \left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right) \tag{1}$$

It smoothly tapers off at both ends, reducing spectral leakage by minimizing discontinuities at segment boundaries.

2.2 Hamming Window

The Hamming window is given by:

$$w(n) = 0.54 - 0.46 \times \cos\left(\frac{2\pi n}{N-1}\right) \tag{2}$$

It is similar to the Hann window but retains slightly higher amplitudes at the edges, offering better frequency resolution at the cost of increased leakage.

2.3 Rectangular Window

The rectangular window is the simplest, defined as:

$$w(n) = 1 \tag{3}$$

for all , meaning no modification is applied. It provides high frequency resolution but causes substantial spectral leakage due to abrupt segment boundaries.

3 Implementation Details

The implementation of this study was carried out using PyTorch and torchaudio. Below are the key steps taken:

3.1 Dataset Processing

- The UrbanSound8K dataset was loaded using its metadata file to access file paths and corresponding class labels.
- The audio files were loaded using torchaudio.load(), which provides a waveform tensor and its corresponding sample rate.
- Each audio sample was resampled to ensure consistency in processing.

3.2 Short-Time Fourier Transform (STFT)

The STFT is computed using the following equation:

$$X(m,k) = \sum_{n=0}^{N-1} x(n + mH) w(n) e^{-j2\pi kn/N} \tag{4}$$

where:

- $X(m, k)$ is the STFT output at time frame and frequency bin k.
- $x(m)$ is the input signal,
- $w(n)$ is the window function,
- N is the window length,
- H is the hop size,
- $e^{-j2\pi kn/N}$ is the Fourier transform basis function.

The STFT was applied to each waveform using different window functions (Hann, Hamming, Rectangular). The function torch.stft() was used with an FFT size of 2048 and a hop length of 512, providing a balance between time and frequency resolution.

3.3 Spectrogram Generation

- The STFT magnitude was computed using torch.abs() to obtain the spectrogram representation.
- The spectrogram was converted to a decibel scale using:

$$S_{dB} = 20 \times \log_{10}(S + \epsilon) \tag{5}$$

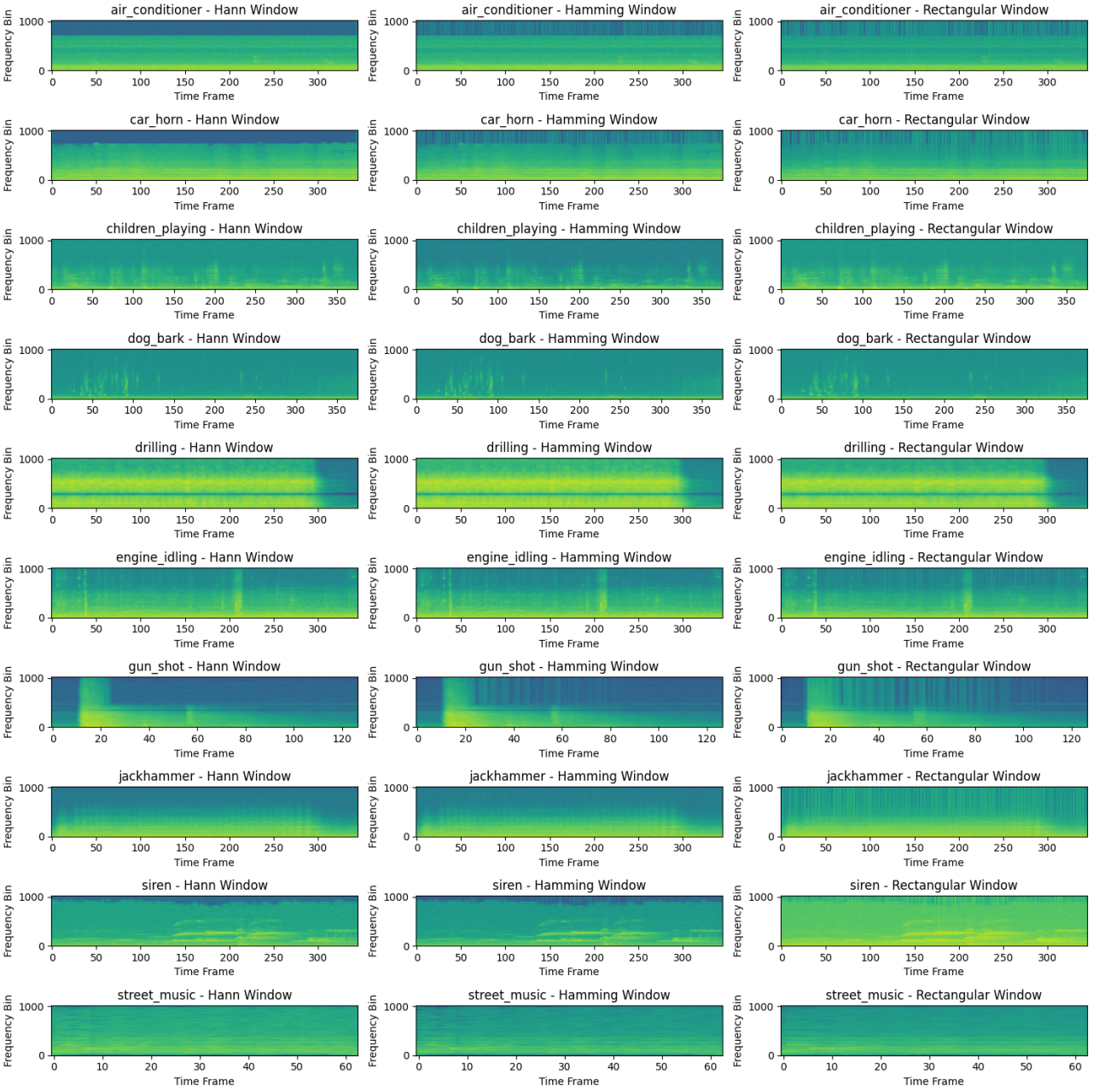
where is the spectrogram magnitude and is a small constant to prevent numerical issues.

3.4 Visualization

- Each spectrogram was plotted using Matplotlib for a visual comparison of different windowing effects.
- The time-axis was labeled in terms of frames, and the frequency-axis was labeled in bins.

4 Visualisation and Correctness

4.1 Visual Differences in Spectrograms



Spectrograms for Different classes with different windowing technique

- **Hann Window**: Provided a balanced trade-off between resolution and leakage, producing a smooth and clear representation of frequency content.
- **Hamming Window**: Retained more energy in frequency bins compared to Hann, showing slightly better amplitude retention at the cost of minor spectral leakage.
- **Rectangular Window**: Exhibited the highest spectral leakage, with smeared frequency components due to the absence of tapering.

Sound Class	Hann Window	Hamming Window	Rectangular Window
Air Conditioner	Smooth, low leakage	Sharp, minor leakage	High leakage
Car Horn	Clear bands, reduced leakage	Similar to Hann	Smearing, high leakage
Children Playing	Balanced, smooth bands	Sharper peaks	Blurred frequencies
Dog Bark	Moderate clarity	Sharp peaks	High leakage
Drilling	Clear harmonics	Slightly better amplitude	Less clarity, high leakage
Engine Idling	Smooth, distinct bands	More amplitude	Frequency smearing
Gun Shot	Clear transients	Sharper details	Loss of sharpness
Jackhammer	Reduced leakage	More amplitude	Less defined bands
Siren	Clear harmonics	Retains energy	Smearing, less detail
Street Music	Smooth transitions	Sharper details	Distorted representation

Table 1: Comparison of Spectrograms Using Different Windowing Techniques

4.2 Analysis of Windowing Effectiveness

- **Correctness of Windowing**: The spectrograms confirm that windowing significantly influences spectral representation. Hann and Hamming windows suppress edge artifacts, whereas the Rectangular window introduces undesired frequency spreading.
- **Comparative Trade-offs**: While the Rectangular window provides fine frequency resolution, its leakage distorts frequency components, making it less suitable for detailed audio classification. The Hann and Hamming windows, however, provide a clearer separation of frequency bands, making them preferable for most practical applications.

5 Neural Network Classification for UrbanSound8K Dataset

5.1 Introduction

The UrbanSound8K dataset consists of audio recordings of urban sounds categorized into ten classes. In this project, we utilized different windowing techniques—Hann, Hamming, and Rectangular—to generate spectrograms from the audio data. A Convolutional Neural Network (CNN) classifier was trained on the spectrograms to evaluate the impact of windowing techniques on model performance.

5.2 Data Preprocessing

5.2.1 Spectrogram Generation:

- Applied Short-Time Fourier Transform (STFT) using different windowing functions: Hann, Hamming, and Rectangular.
- Used an FFT size of 512 and a hop length of 256.

5.2.2 Dataset Splitting:

- The dataset was split into training (80%) and testing (20%) sets.
- Data was loaded using PyTorch's DataLoader with a batch size of 32.

5.3 Model Architecture

A simple CNN was designed with the following layers:

- Three convolutional layers with ReLU activation and max-pooling.
- Adaptive average pooling to ensure fixed-sized feature maps.
- A fully connected layer with ten output neurons corresponding to the sound classes.
- Cross-entropy loss function and Adam optimizer with a learning rate of 0.001.

5.4 Training and Evaluation

Each model was trained for five epochs on a GPU-enabled environment. The performance was assessed using training loss, validation loss, and accuracy.

Window Type	Training Accuracy (%)	Validation Accuracy (%)	Training Loss	Validation Loss
Hann	53.31	52.09	1.3804	1.5031
Hamming	53.97	51.46	1.3558	1.4933
Rectangular	52.51	53.23	1.3991	1.5337

Table 2: Comparison of CNN performance using different windowing techniques

5.5 Results and Analysis

5.5.1 Impact of Windowing Techniques:

- The Hann and Hamming windows produced slightly better training accuracy than the Rectangular window.
- Validation accuracy was highest for the Rectangular window (53.23%), but the difference was marginal.

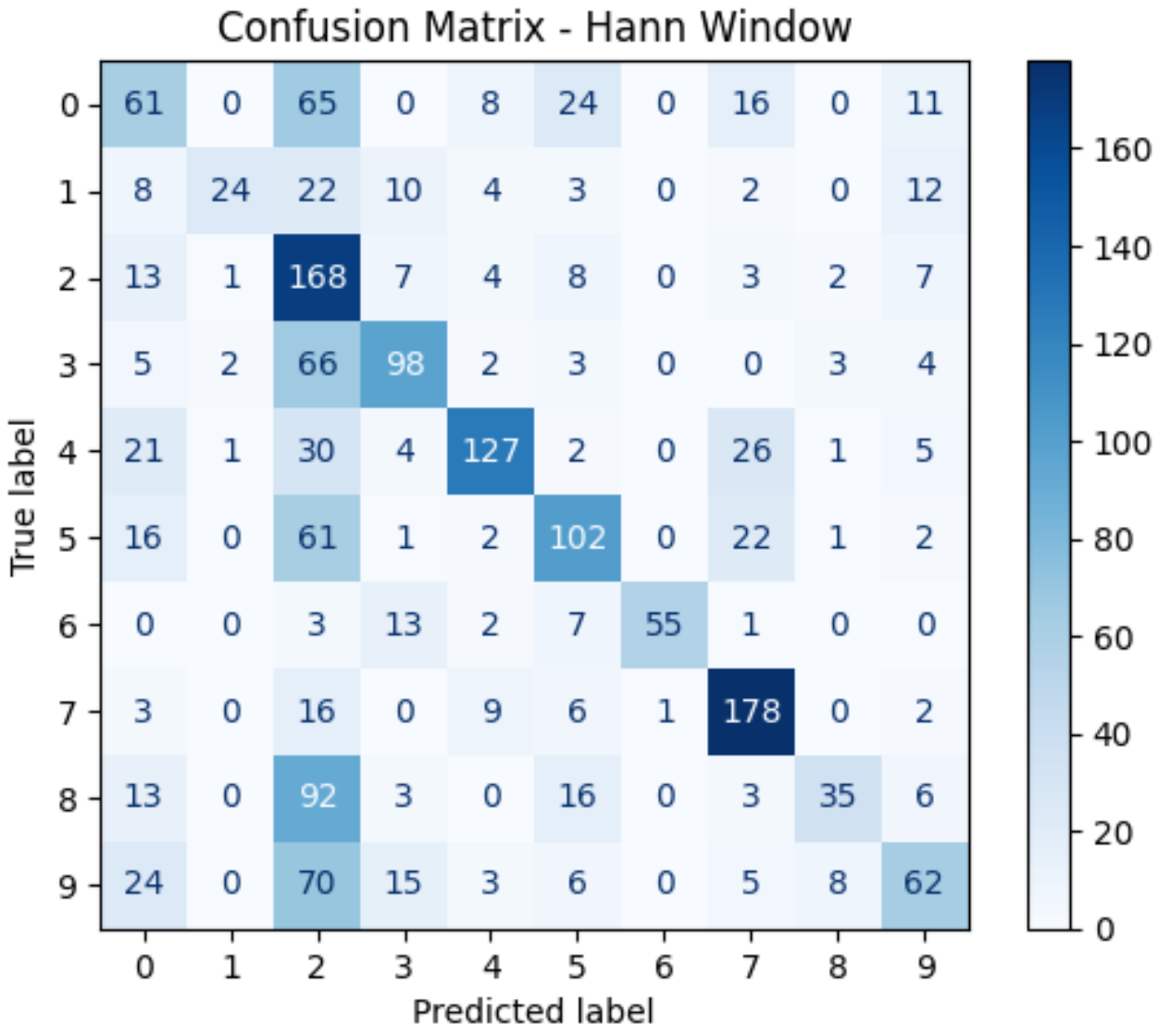
5.5.2 Loss Analysis:

- The training loss for all models steadily decreased, indicating effective learning.
- The Hann and Hamming windows resulted in similar validation losses, whereas the Rectangular window had slightly higher validation loss.

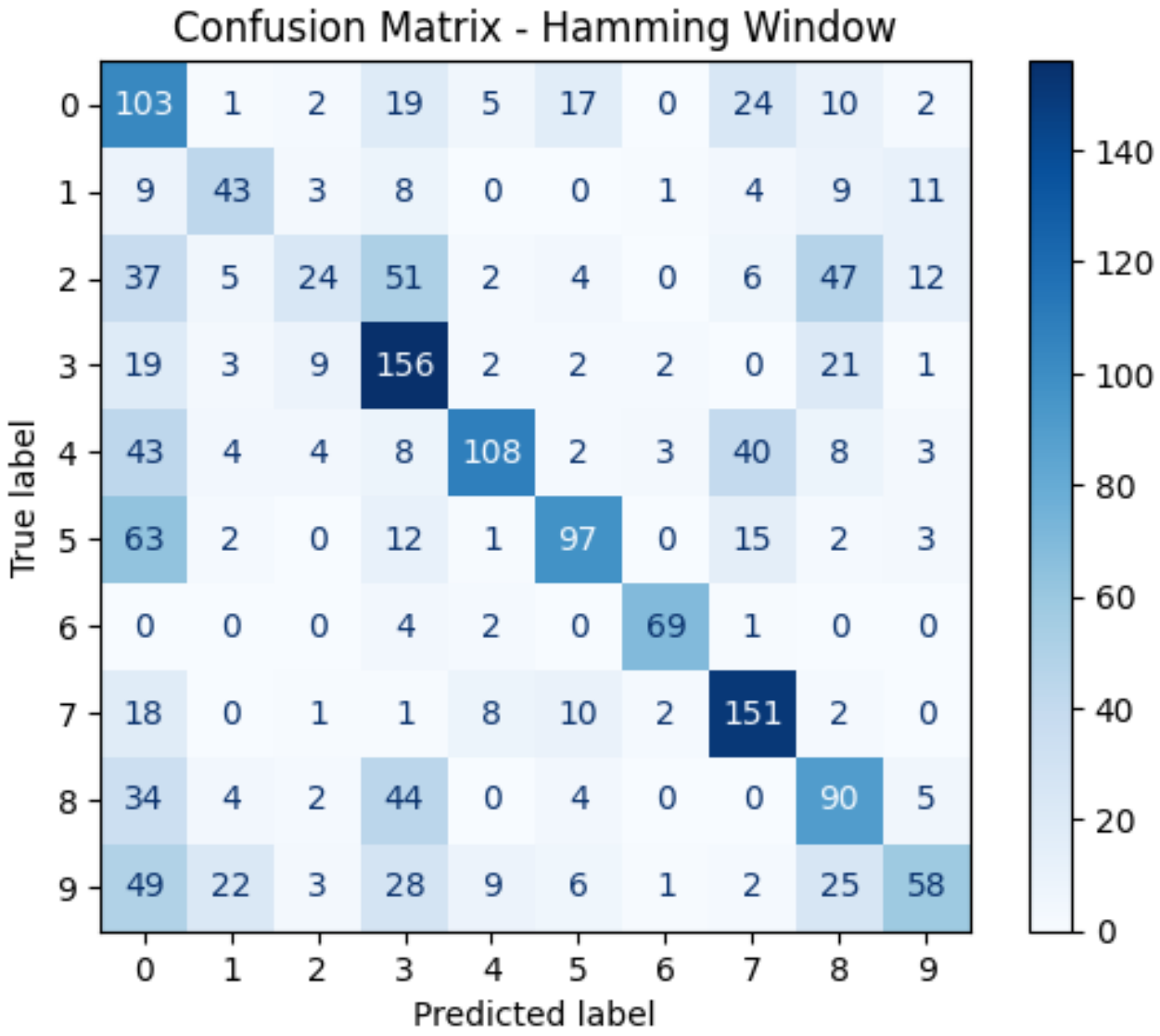
5.5.3 Overall Trends:

- The model showed consistent performance across different windowing techniques, with only minor variations.
- Further hyperparameter tuning (e.g., increased training epochs, deeper networks) may enhance performance.

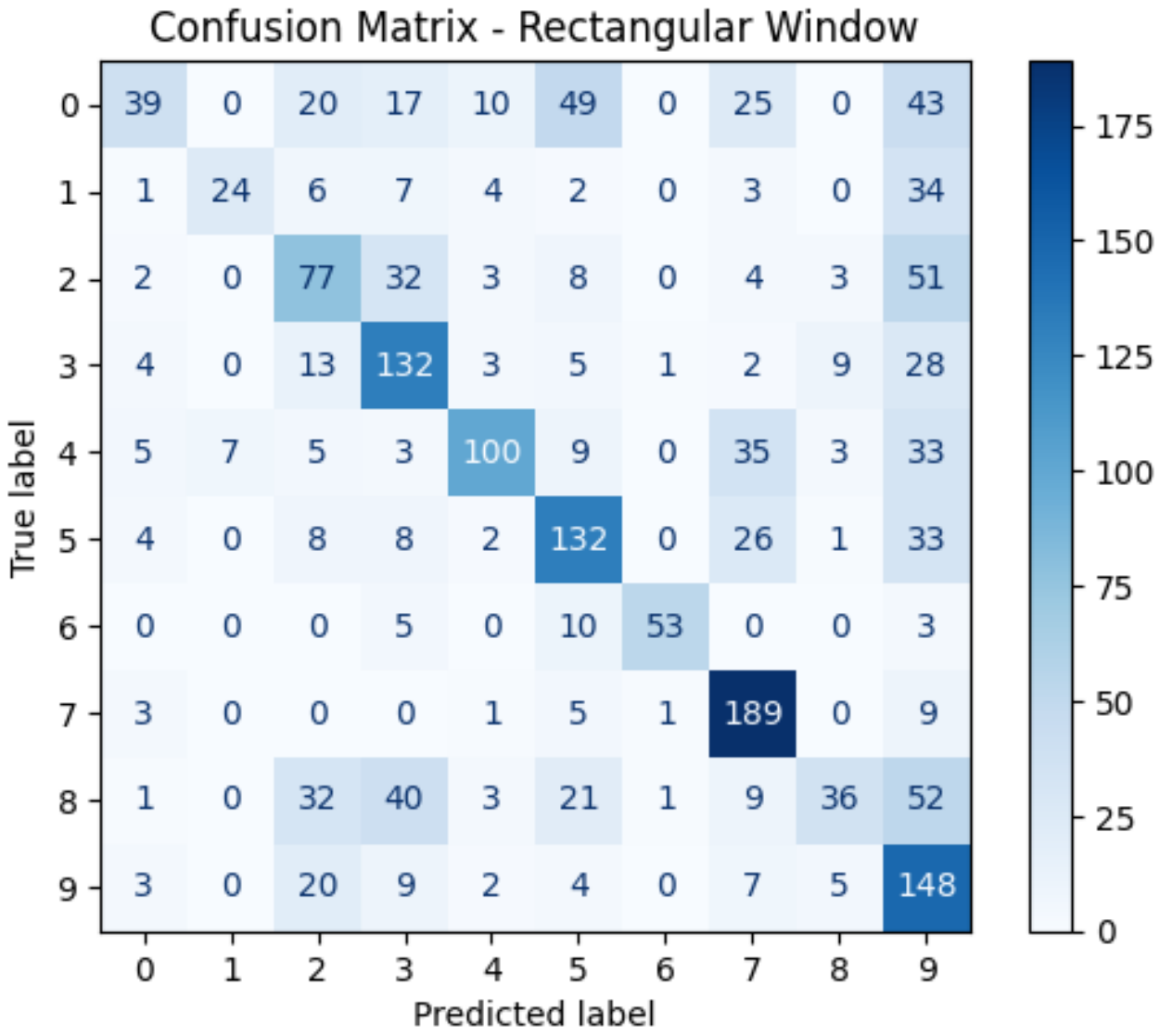
5.5.4 Confusion Matrix for Different Windowing Technique



Confusion Matrix for hann window



Confusion Matrix for hamming window



Confusion Matrix for rectangular window

The choice of windowing function had a subtle impact on classification performance. The Hann and Hamming windows provided slightly better training accuracy, whereas the Rectangular window achieved the highest validation accuracy. These results suggest that windowing techniques play a role in spectrogram-based classification, though their effect is not substantial in this specific model. Future work could explore advanced feature extraction methods and deeper neural architectures to improve accuracy.

6 Comparative Analysis of Spectrograms from Different Music Genres

6.1 Introduction

Spectrograms visually represent the frequency content of a signal over time. In this study, we analyze and compare spectrograms of four songs from different genres: heavy metal, jazz, pop, and rock. The goal is to identify distinct spectral characteristics that define each genre.

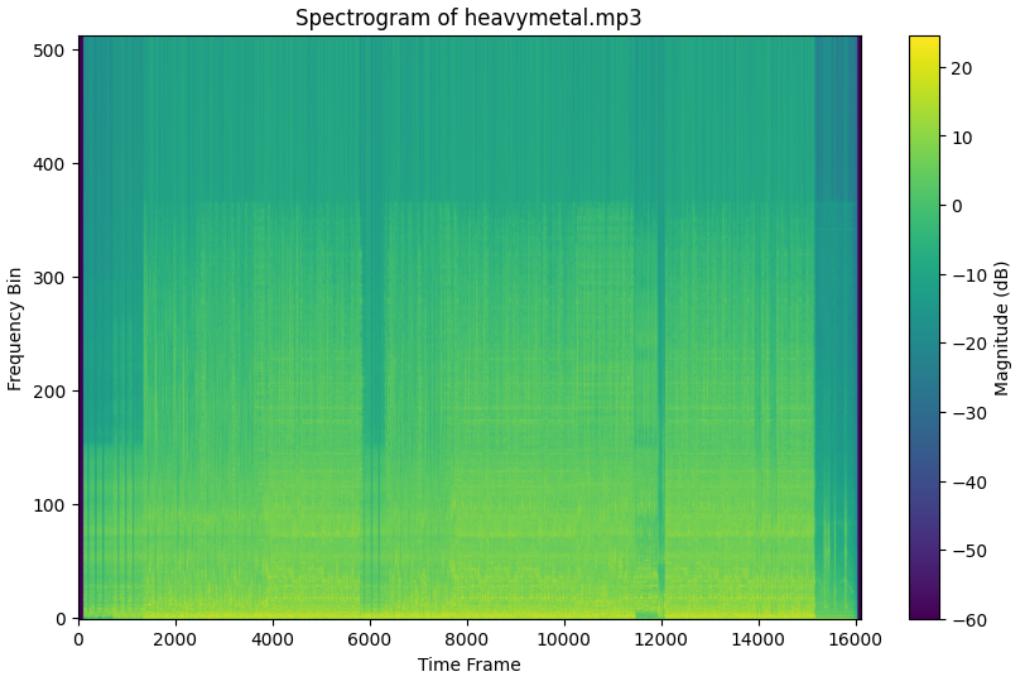
6.1.1 Songs Selected

- Awake and Alive – Skillet (Heavy Metal)
- Just the Two of Us – Bill Withers Grover Washington, Jr. (Jazz)
- Bye Bye Bye – NSYNC (Pop)
- Hotel California – Eagles (Rock)

Each song's spectrogram has been analyzed to observe frequency distribution, intensity, patterns, and speech-like elements.

6.2 Spectrogram Analysis by Genre

6.3 Heavy Metal (Awake and Alive - Skillet)



Spectrogram of Awake and Alive Song by Skillet (Heavy Metal)

Frequency Distribution:

- High energy is distributed across a broad frequency range (50 Hz to 16 kHz).
- Strong low-frequency components from distorted guitars and bass.

Intensity and Texture:

- High-intensity, dense spectrogram with little silence.
- Sudden changes in amplitude and frequency shifts due to aggressive instrumentation.

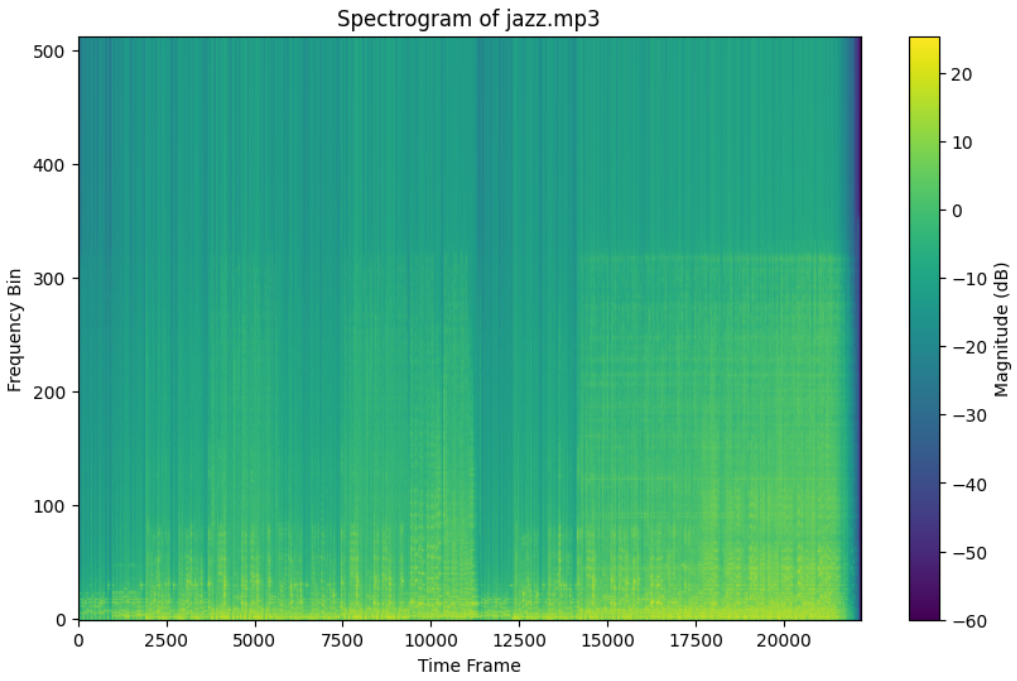
Rhythmic and Speech Components:

- Limited speech-like characteristics; vocals are highly processed.
- Frequent transitions between high-energy segments, typical of metal.

Observation:

- Heavy metal spectrograms tend to be highly saturated across all frequencies.
- Little to no silent regions due to the continuous presence of distorted guitars and intense drumming.

6.4 Jazz (Just the Two of Us - Bill Withers Grover Washington, Jr.)



Spectrogram of Just the Two of Us Song by Bill Withers Grover Washington, Jr. (Jazz)

Frequency Distribution:

- Smooth frequency spread with a strong presence in mid-range (200 Hz – 5 kHz).
- Saxophone and vocals dominate upper midrange (1 kHz – 5 kHz).

Intensity and Texture:

- Noticeable dynamic range, with soft and loud passages.
- Gaps between instrument notes, reflecting improvisation.

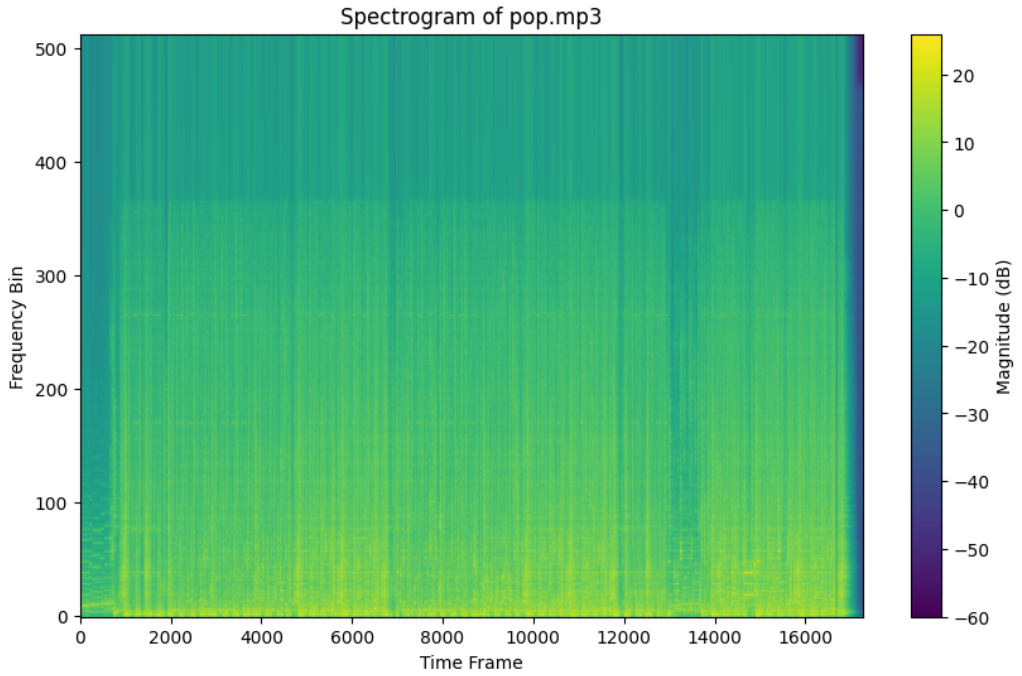
Rhythmic and Speech Components:

- Strong speech-like characteristics due to smooth and expressive vocals.
- Clear separation between instruments (e.g., bass, saxophone, and piano).

Observation:

- Jazz spectrograms show distinct note separation with smoother transitions, unlike the continuous dense structure in heavy metal. The improvisational elements are visible through varied intensity.

6.5 Pop (Bye Bye Bye - NSYNC)



Spectrogram of Bye Bye Bye Song by NSYNC (POP)

Frequency Distribution:

- Vocals dominate midrange frequencies (300 Hz – 6 kHz).
- Bass and drum beats concentrated in lower range (50 Hz – 500 Hz).

Intensity and Texture:

- Balanced intensity with structured patterns due to digital production.
- Repetitive sections with clear verse-chorus segmentation.

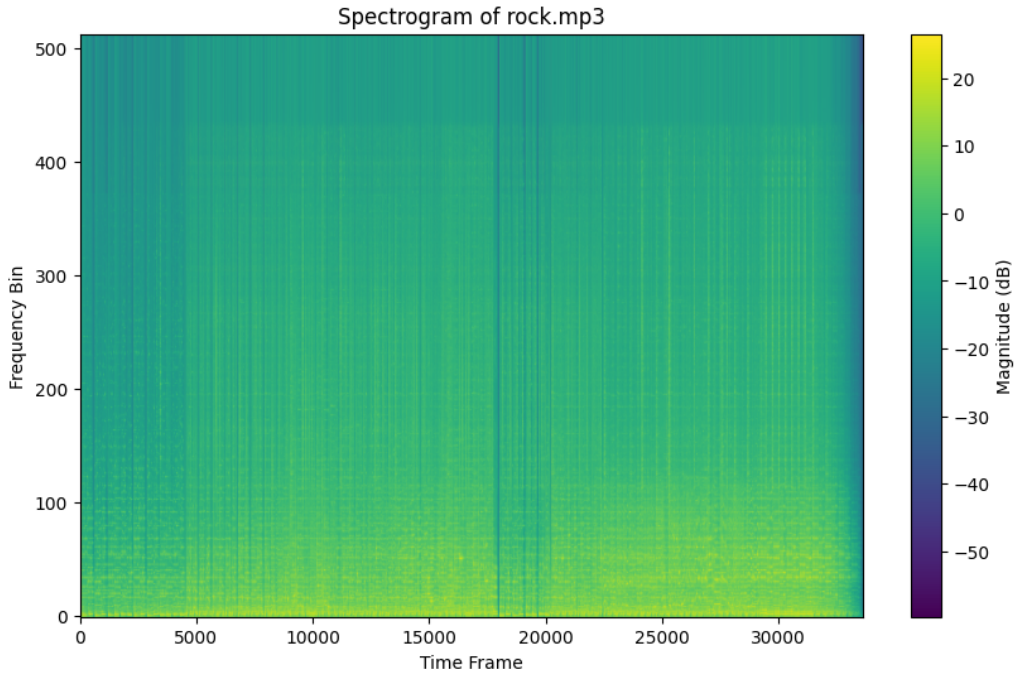
Rhythmic and Speech Components:

- High speech-like characteristics due to prominent, auto-tuned vocals.
- Distinct rhythmic patterns visible in beats and transitions.

Observation:

- Pop spectrograms show well-defined patterns and structured rhythms. The prominence of vocals and percussion reflects a balance between musical elements.

6.6 Rock (Hotel California - Eagles)



Spectrogram of Hotel California Song by Eagles (Rock)

Frequency Distribution:

- Strong midrange presence (200 Hz – 6 kHz).

- Guitar solos and vocals dominate higher frequencies (2 kHz – 8 kHz).

Intensity and Texture:

- Balanced mix of continuous and separated elements.
- Moderate intensity compared to heavy metal but more structured than jazz.

Rhythmic and Speech Components:

- Clear vocal articulation with melodic variations.
- Instrumental solos (guitar) introduce variations in frequency.

Observation:

- Rock spectrograms exhibit a structured yet dynamic intensity, blending vocal clarity with instrumental prominence.