Analysis of Windowing Techniques and Spectrograms for Audio Classification

Neermita Bhattacharya

Indian Institute of Technology, Jodhpur, Jodhpur 342037, India
b22cs092@iitj.ac.in

Abstract

Spectrogram analysis of audio files is an essential task in audio signal processing, enabling the visualization of sound frequencies over time. This study focuses on the application of windowing techniques and spectrogram analysis to extract meaningful features from audio signals. Using these features, audio classification is performed to identify patterns and distinguish between different sound sources. Windowing methods, such as Hamming and Hanning, are employed to optimize time-frequency resolution, minimizing spectral leakage and enhancing feature extraction. The spectrogram features, which encapsulate the energy distribution across frequencies, are utilized as inputs for classification algorithms. This approach demonstrates the effectiveness of spectrogram-based feature analysis in various tasks like speech recognition, music genre classification, and sound identification. The results highlight the critical role of windowing functions and spectrogram features in achieving accurate and efficient audio classification. Training machine learning classifiers to distinguish between different audio classes is a non-reference-based task. Two classical models— Support Vector Machine (SVM) and Random Forest (RF)—were used for audio classification. One of the objectives was to analyze whether padding affects classification performance. To investigate this, the SVM model was trained on both padded and non-padded audio clips. The results showed that non-padded data yielded higher accuracy. When trained on padded data, it achieved classification accuracies of 61%, 62%, and 62% using Rectangular, Hamming, and Hann windowing techniques, respectively. In contrast, when trained on non-padded data, the SVM consistently achieved 68% accuracy across all windowing techniques. The Random Forest model, trained only on non-padded data, obtained classification accuracies of 54%, 57%, and 58% for the respective windowing techniques. Additionally, the differences in spectrograms produced by each windowing technique for various audio classes was analyzed and a comparative evaluation of the classification performance is provided. The codes and songs are available at github.com/Neermita18/Speech-Understanding-Labs.

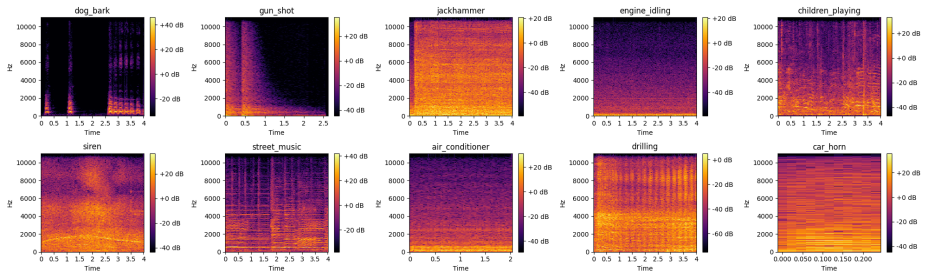

Figure 1: Spectrograms of the 10 classes in UrbanSound8K dataset

Contents

1 Question 2

1.1 Introduction

Speech analysis is the study and processing of spoken language to extract meaningful information about its linguistic, acoustic, and emotional content. It is a cornerstone of speech technology applications such as speech recognition, speaker identification, emotion detection, and language modeling. Moreover, changes in speech patterns can serve as biomarkers for detecting and monitoring a range of medical conditions, especially neurological and mental health disorders. It is essential to pre-process auditory data before any sort of analysis. A global amplitude analysis can be done in the time domain whereas a global frequency analysis can be done in the frequency domain using the Forward Fourier Transform.

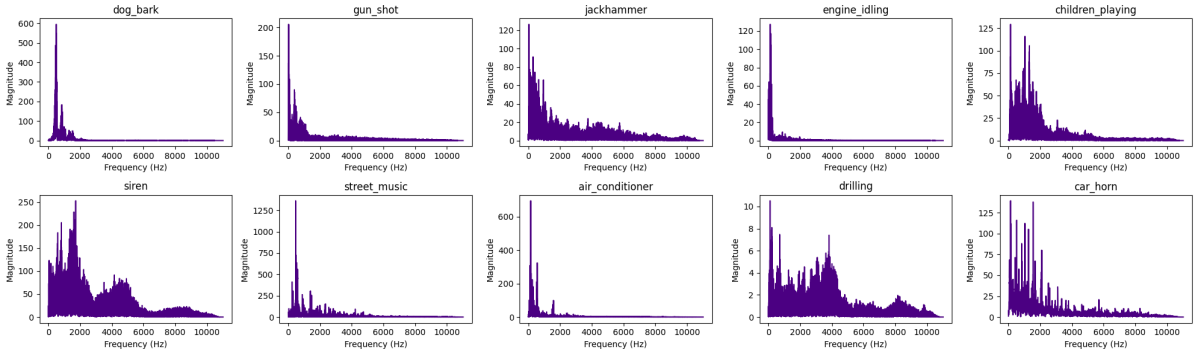

Figure 2: Fast Fourier Transforms of the 10 classes of UrbanSound8k Dataset

However, speech signals are non-stationary, meaning their properties change over time. To analyze such signals effectively, they are divided into small time segments using a process called windowing. Each windowed segment is short enough for better feature extraction (such as pitch and formants). Common windowing functions like Hamming and Hanning are applied to smooth the edges of the segments, reducing spectral leakage during frequency analysis. Short-Term Fourier Transform (STFT) is applied to these overlapping windowed segments. STFT inherently requires windowing for localized time and frequency analysis. The window size (n_fft) is a hyperparameter. A larger n_fft gives better frequency resolution because we analyze a larger portion of the signal, but it results in poorer time resolution. A smaller n_fft gives better time resolution (finer temporal granularity), but poorer frequency resolution.

After STFT of the windowed segments, spectrogram analysis is performed. A spectrogram represents how the frequency content of a speech signal evolves over time. In speech and audio analysis, spectrograms are widely used to visualize and analyze features such as formants, pitch, loudness, etc. Spectrograms provide various insights by visualizing how frequency and amplitude evolve over time in an audio signal. For example, a piano note and a drum sound can be easily differentiated: the piano note shows distinct harmonic frequencies that decay gradually, while due to the transient nature of drum hits, they show up as sharp vertical lines rather than horizontal lines like sustained notes from melodic instruments. This ability to visualize frequency content over time is essential for tasks like sound classification and analysis, as it reveals unique patterns in various types of audio signals.

1.2 Task A

Task A involved utilizing the UrbanSound8k dataset, which consists of 10 distinct audio classes. Three windowing techniques—Rectangular, Hann, and Hamming—were applied to analyze the effect of windowing on spectral representation. A sample from each class was processed using Short-Time Fourier Transform (STFT) to compare the differences introduced by each windowing method.

1.2.1 Analysis of Windowing Techniques

In signal processing, window functions are used to minimize spectral leakage when analyzing signals using the Short-Time Fourier Transform (STFT). Spectral leakage occurs when a signal is windowed before performing STFT, leading to the spreading of frequency components across adjacent frequencies. It happens because of the discontinuities introduced at the edges of a finite windowed segment when the signal is not perfectly periodic within that segment. This leakage causes energy from a single frequency to leak into neighboring frequencies. Three commonly used window functions are the Rectangular, Hann, and Hamming windows.

Rectangular Window
The Rectangular window is the simplest window function, where all samples within the window have equal weights. It is defined mathematically as:

$$w(n) = 1, \quad 0 \leq n \leq N-1 \tag{1}$$

where N is the total number of samples in the window. This window is computationally efficient and preserves the maximum amount of signal information since no modifications are made to the data. However, its abrupt transitions (vertical cut-offs) at the edges result in significant spectral leakage, making it unsuitable for most applications.

Hann Window
The Hann window, also known as the Hanning window, is a smoothly tapered function that reduces spectral leakage by gradually decreasing the amplitude at the edges. It is defined as:

$$w(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right), \quad 0 \leq n \leq N-1 \tag{2}$$

This window significantly reduces spectral leakage compared to the Rectangular window, making it a better choice for frequency analysis. However, due to its tapering, it slightly lowers the amplitude of the signal, leading to potential loss of information. Additionally, its wider main lobe reduces frequency resolution more than Rectangular windowing.

Hamming Window
The Hamming window is a modified version of the Hann window, designed to further reduce spectral leakage. It is mathematically expressed as:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \tag{3}$$

Compared to the Hann window, the Hamming window retains signal amplitude more effectively while providing even better suppression of spectral leakage. However, this improvement comes at the cost of a slightly wider main lobe, which reduces frequency resolution.

The Rectangular window is useful when preserving the original signal strength is a priority, but it suffers from high spectral leakage. The Hann window provides a good trade-off between frequency resolution and leakage, while the Hamming window further reduces leakage at the cost of a slightly wider main lobe.

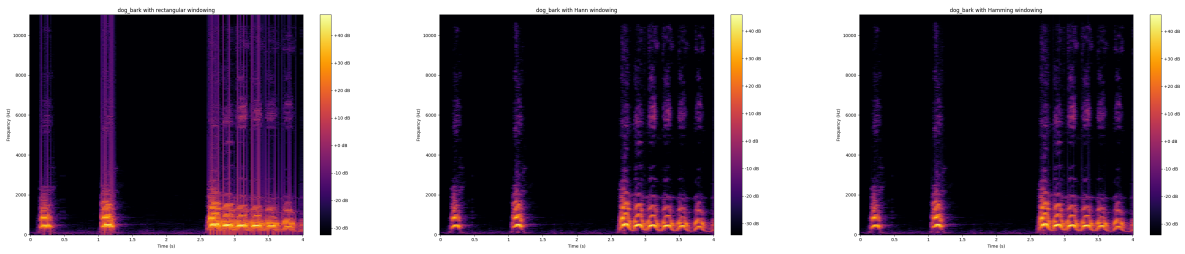

Figure 3: Different windowing techniques on an audio sample (dog_bark)

We notice in the spectrogram of Rectangular windowing, vertical bands of energy extend beyond the main frequency components. Instead of sharply defined frequencies, the energy leaks into neighboring frequencies, making the spectral lines blurry. The dog_bark energy should be concentrated and sharp, but with the rectangular window, it is spread out more than necessary. There is no smooth transitions and no clear frequency components. Also, it includes high frequency artifacts (higher frequencies show moderate energies too). The Hann and Hamming show much less spectral leakage. The gaps between the barks are less abrupt and more smooth. The Hann window shows slightly higher energies at higher frequencies which is expected since it has a smaller main lobe than the Hamming window.

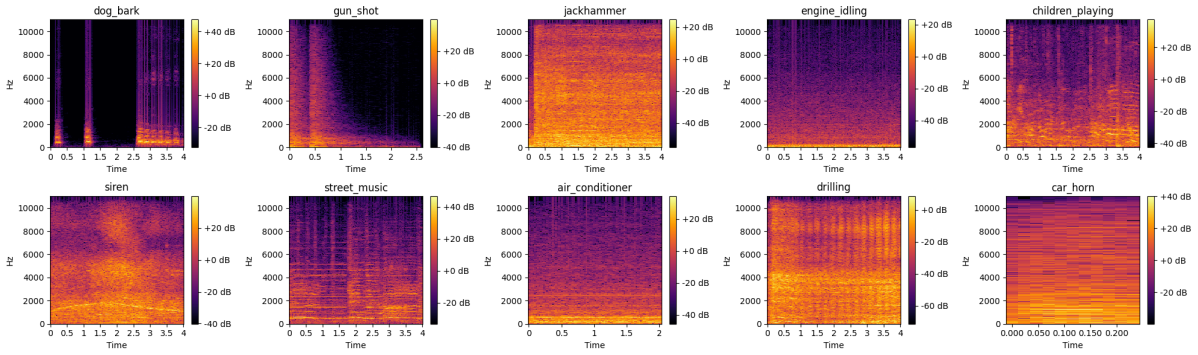

Figure 4: Spectrogram Analysis with Rectangular Windowing

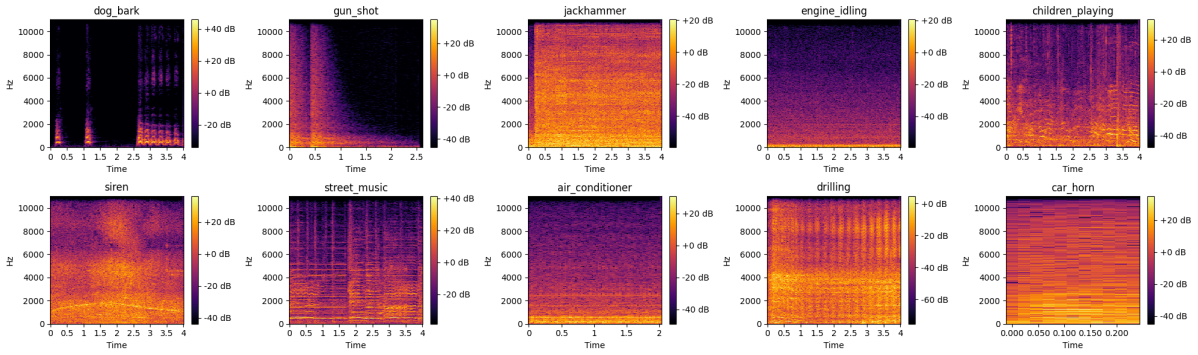

Figure 5: Spectrogram Analysis with Hann Windowing

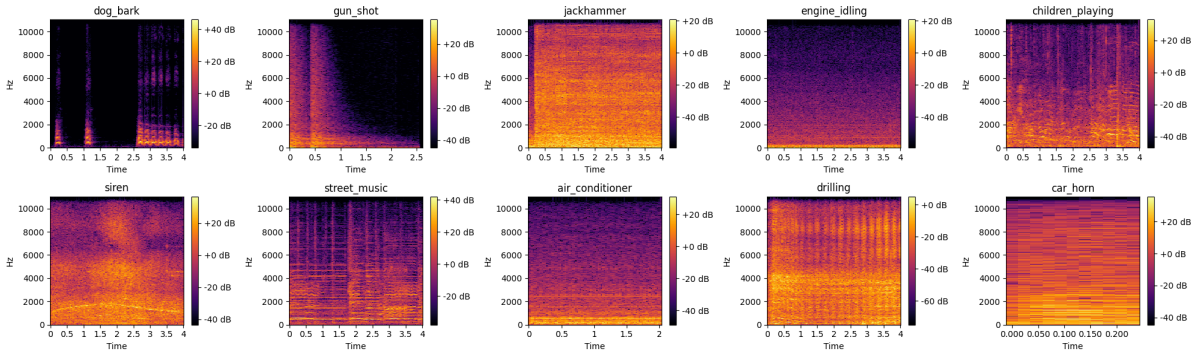

Figure 6: Spectrogram Analysis with Hamming Windowing

Through this comprehensive investigation, we aim to provide insights into the differences in results produced due to the windowing techniques as well as the impact padding the audio clips has on the performance of models.

1.2.2 Classification of Audio Clips with Machine Learning

Two ML models were used- Support Vector Machine (SVM) and Random Forest (RF). SVM finds an optimal decision boundary (hyperplane) that maximizes the margin between different classes. It uses kernel tricks to handle non-linearly separable data by mapping it to a higher-dimensional space. RF is an ensemble of multiple decision trees, where each tree votes on the class, and the final prediction is based on majority voting (for classification) or averaging (for regression). It reduces overfitting by training on random subsets of data and features.

The UrbanSound8K dataset consists of 10 folds. Each fold contains audio clips of ten classes, namely- dog_bark, gun_shot, jackhammer, engine_drilling, siren, street_music, air_conditioner, and drilling. 9 folds were used for training, and 1 fold was used for testing. The precision, recall and f1-score of all classes were reported each time after testing. The overall accuracy of the specific combination of windowing technique and classifier is also reported. One issue with training with traditional Neural Networks is that they require inputs of the same size. Audio clips are generally not similar in duration and hence, padding and duration clipping are used. This causes loss of information. Padding with zeros artificially reduces variance, leading to:

- Mean, Median, and Energy Drop – Silence lowers energy-related features.
- Skewness & Kurtosis Distortion – More zeros bias the distribution, reducing useful frequency variations.
- MFCCs & Spectral Contrast Become Less Informative – Extra silence can mask real patterns.
- Chroma Features Get Corrupted – Silence introduces unnatural tonal structures.

Classification on padded audio
Audio clips were padded with 0 and each clip had a fixed duration of 3 seconds. The sampling rate was 22050 Hz. n_fft = 1024 and the hop length = 512. The results on fold 10 are as follows-

Rectangular Window Results:

	precision	recall	f1-score	support
air_conditioner	0.69	0.81	0.75	100
car_horn	0.95	0.61	0.74	33
children_playing	0.53	0.74	0.62	100
dog_bark	0.77	0.58	0.66	100
drilling	0.39	0.51	0.44	100
engine_idling	0.71	0.71	0.71	93
gun_shot	0.88	0.72	0.79	32
jackhammer	0.48	0.33	0.40	96
siren	0.55	0.39	0.45	83
street_music	0.69	0.75	0.72	100
accuracy			0.61	837
macro avg	0.67	0.61	0.63	837
weighted avg	0.63	0.61	0.61	837

Hann Window Results:

	precision	recall	f1-score	support
air_conditioner	0.55	0.83	0.66	100
car_horn	0.85	0.52	0.64	33
children_playing	0.52	0.77	0.62	100
dog_bark	0.78	0.58	0.67	100
drilling	0.51	0.55	0.53	100
engine_idling	0.70	0.45	0.55	93
gun_shot	0.88	0.72	0.79	32
jackhammer	0.67	0.59	0.63	96
siren	0.59	0.41	0.48	83
street_music	0.71	0.76	0.73	100
accuracy			0.62	837
macro avg	0.68	0.62	0.63	837
weighted avg	0.65	0.62	0.62	837

Hamming Window Results:

	precision	recall	f1-score	support
air_conditioner	0.56	0.83	0.67	100
car_horn	0.94	0.45	0.61	33
children_playing	0.51	0.77	0.62	100
dog_bark	0.75	0.60	0.67	100
drilling	0.49	0.52	0.50	100
engine_idling	0.67	0.40	0.50	93
gun_shot	0.92	0.72	0.81	32
jackhammer	0.64	0.60	0.62	96
siren	0.60	0.40	0.48	83
street_music	0.69	0.77	0.73	100
accuracy			0.62	837
macro avg	0.68	0.61	0.62	837
weighted avg	0.64	0.62	0.61	837

Figure 7: Results of padding on windowing techniques + SVM classifier

Classification on non-padded audio (raw audio)
The audio clips are of varying lengths. The sampling rate was 22050 Hz. n_fft = 1024 and the hop length = 512. The results on fold 10 are as follows-

Rectangular Window Results:

	precision	recall	f1-score	support
air_conditioner	0.77	0.77	0.77	100
car_horn	0.83	0.61	0.70	33
children_playing	0.56	0.76	0.64	100
dog_bark	0.82	0.65	0.73	100
drilling	0.58	0.52	0.55	100
engine_idling	0.69	0.85	0.76	93
gun_shot	0.93	0.88	0.90	32
jackhammer	0.66	0.68	0.67	96
siren	0.56	0.37	0.45	83
street_music	0.71	0.78	0.74	100
accuracy			0.68	837
macro avg	0.71	0.69	0.69	837
weighted avg	0.69	0.68	0.68	837

Hann Window Results:

	precision	recall	f1-score	support
air_conditioner	0.54	0.82	0.65	100
car_horn	0.85	0.67	0.75	33
children_playing	0.54	0.78	0.64	100
dog_bark	0.82	0.64	0.72	100
drilling	0.72	0.53	0.61	100
engine_idling	0.66	0.51	0.57	93
gun_shot	1.00	0.94	0.97	32
jackhammer	0.77	0.82	0.80	96
siren	0.56	0.39	0.46	83
street_music	0.77	0.79	0.78	100
accuracy			0.68	837
macro avg	0.72	0.69	0.69	837
weighted avg	0.69	0.68	0.67	837

Hamming Window Results:

	precision	recall	f1-score	support
air_conditioner	0.59	0.81	0.68	100
car_horn	0.86	0.73	0.79	33
children_playing	0.54	0.80	0.64	100
dog_bark	0.78	0.64	0.70	100
drilling	0.67	0.52	0.58	100
engine_idling	0.71	0.60	0.65	93
gun_shot	1.00	0.84	0.92	32
jackhammer	0.79	0.80	0.80	96
siren	0.57	0.36	0.44	83
street_music	0.76	0.81	0.79	100
accuracy			0.68	837
macro avg	0.73	0.69	0.70	837
weighted avg	0.70	0.68	0.68	837

Figure 8: Results of windowing techniques (on raw audio) + SVM classifier

Rectangular Window Results:

	precision	recall	f1-score	support
air_conditioner	0.57	0.74	0.64	100
car_horn	0.45	0.45	0.45	33
children_playing	0.53	0.80	0.63	100
dog_bark	0.72	0.56	0.63	100
drilling	0.34	0.51	0.41	100
engine_idling	0.56	0.41	0.47	93
gun_shot	0.92	0.72	0.81	32
jackhammer	0.54	0.23	0.32	96
siren	0.51	0.25	0.34	83
street_music	0.62	0.75	0.68	100
accuracy			0.54	837
macro avg	0.58	0.54	0.54	837
weighted avg	0.56	0.54	0.53	837

Hann Window Results:

	precision	recall	f1-score	support
air_conditioner	0.50	0.69	0.58	100
car_horn	0.62	0.45	0.53	33
children_playing	0.54	0.84	0.66	100
dog_bark	0.77	0.57	0.66	100
drilling	0.45	0.53	0.48	100
engine_idling	0.42	0.27	0.33	93
gun_shot	0.92	0.75	0.83	32
jackhammer	0.60	0.52	0.56	96
siren	0.59	0.28	0.38	83
street_music	0.65	0.77	0.71	100
accuracy			0.57	837
macro avg	0.61	0.57	0.57	837
weighted avg	0.58	0.57	0.56	837

Hamming Window Results:

	precision	recall	f1-score	support
air_conditioner	0.52	0.78	0.63	100
car_horn	0.67	0.55	0.60	33
children_playing	0.51	0.81	0.62	100
dog_bark	0.79	0.60	0.68	100
drilling	0.46	0.59	0.52	100
engine_idling	0.47	0.26	0.33	93
gun_shot	0.92	0.72	0.81	32
jackhammer	0.65	0.45	0.53	96
siren	0.58	0.27	0.36	83
street_music	0.66	0.78	0.72	100
accuracy			0.58	837
macro avg	0.62	0.58	0.58	837
weighted avg	0.60	0.58	0.57	837

Figure 9: Results of windowing techniques (on raw audio) + RF classifier

1.2.3 Analysing the Effect of Padding & different Windowing Methods

Padding adds artificial silence to shorter audio clips, which lowers variance and distorts feature distributions. This negatively impacts classification because MFCCs and spectral features become less informative due to extra silent regions, energy-related features decrease, making it harder to distinguish between classes. This explains why non-padded data performed better, as it preserved the original signal characteristics.

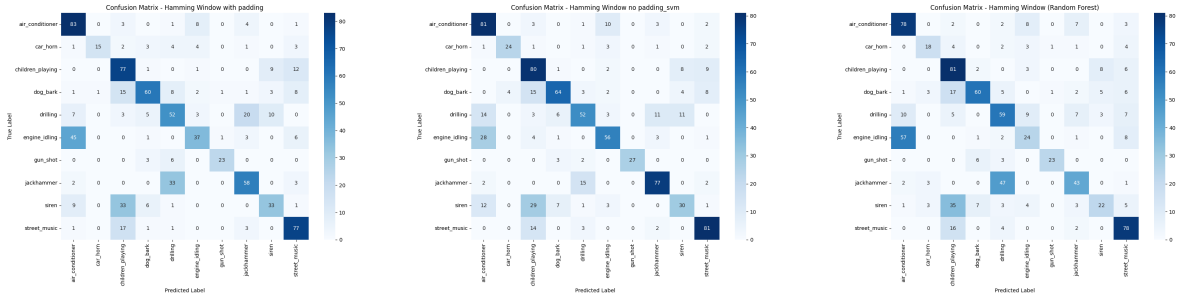

Figure 10: Confusion matrices of the Hamming window applied on padded audio + SVM, non-padded audio + SVM, non-padded audio + RF

The rectangular window causes high spectral leakage, spreading energy across frequencies and reducing classification precision. Hamming & Hann Windows reduce spectral leakage by smoothing edges, leading to slightly better accuracy than Rectangular. Rectangular generally showed a poorer performance than the other windowing techniques. Thus, non-padded data + Hann/Hamming performed best, while padded data limited accuracy across all windows.

1.3 Task B

For this task, [1] was referred to.

Task B centers on the spectrogram analysis of four distinct songs, carefully selected to ensure diversity and variety. The chosen pieces include a melancholic English song, a K-Pop track, an orchestral composition, and an energetic Hindi song.

1.3.1 Instead - Ryan Amador (Melancholic English song)

The song can be downloaded from Instead.mp3.

- Pauses (periods of silence), low frequencies dominating and vertical striations in the spectrogram denote a male singer. The constant horizontal lines with upper harmonics denote piano notes.
- Piano notes are prominent at lower frequencies as this song is indeed a deep-pitched song.
- This song does not have a constant beat and hence, thin vertical stripes are not very significant. This is in deep contrast to the last song- Masakali, where the constant beats are visible.

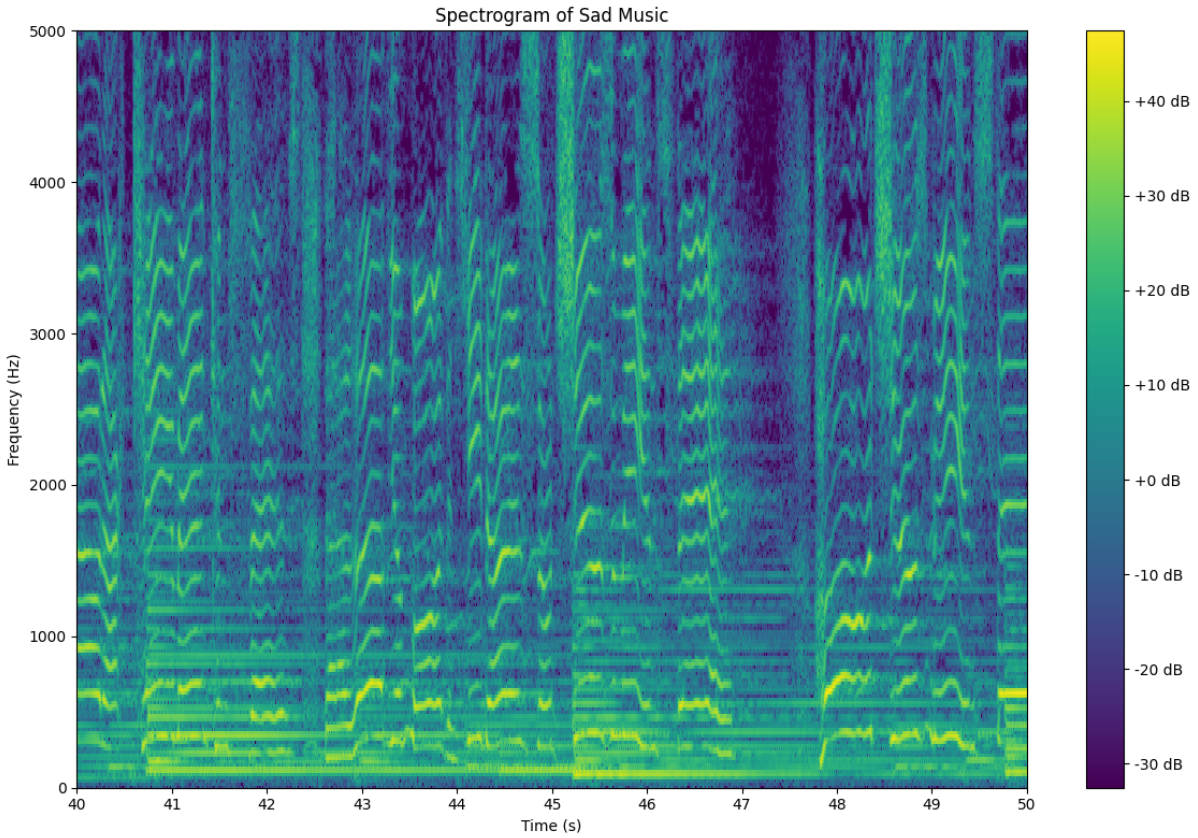

Figure 11: Spectrogram Analysis of Instead - Ryan Amador

1.3.2 Bad News - Kiss of Life (K-Pop song)

The song can be downloaded from Bad News.mp3.

- The spectrogram reveals vertical striations, indicating vocal regions, which aligns with the fact that this song features prominent vocals by the band.
- Guitar riffs are evident as high-intensity patterns at lower frequencies, characterized by vertical and horizontal lines stacked on top of each other with harmonics.

- Unlike the previous song, there are no prominent constant horizontal lines, suggesting that instruments like the piano or violin do not play a significant role in this track.
- Additionally, the absence of significant pauses reflects a consistent tempo throughout the song.

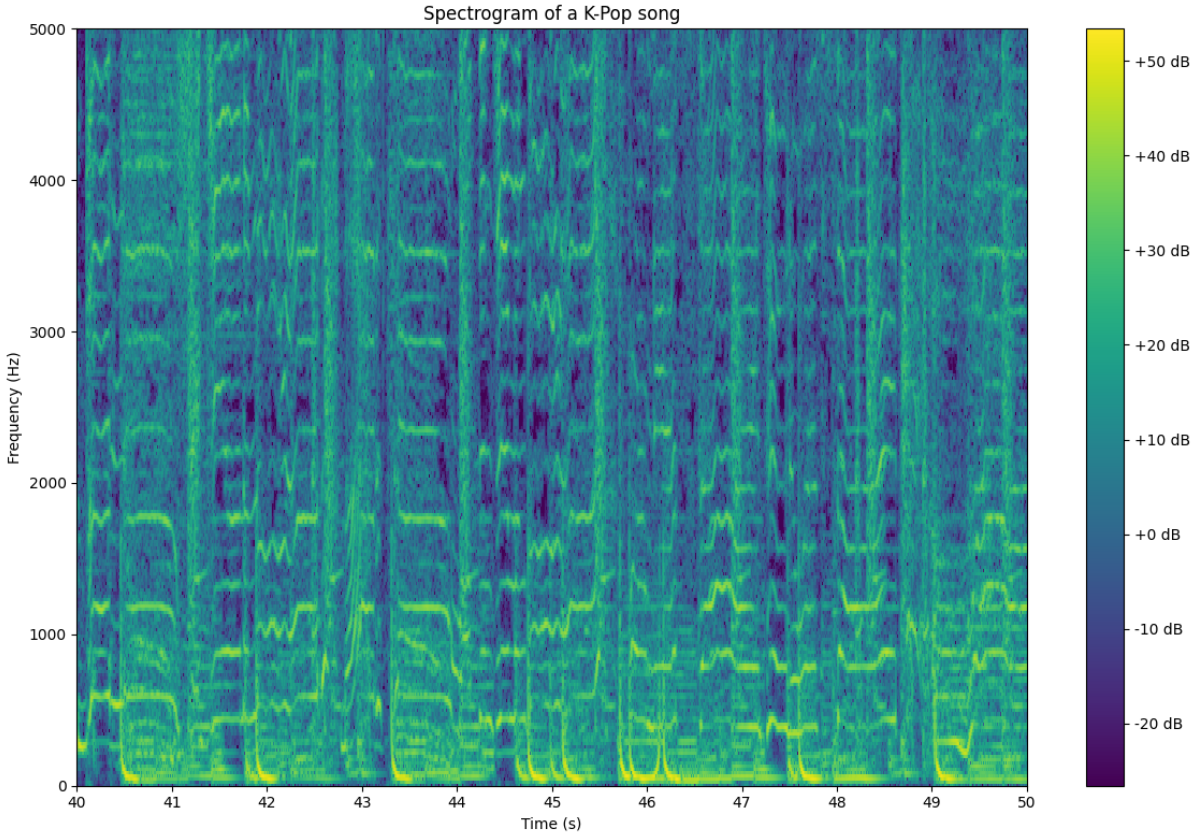

Figure 12: Spectrogram Analysis of Bad News - Kiss of Life

1.3.3 Romantic Flight - John Powell (Orchestral piece)

The song can be downloaded from Romantic Flight.mp3.

- The spectrogram reveals sharp rises in frequency, likely indicating sudden changes in pitch. This is characteristic of orchestral pieces, where scale shifts often occur. In this case, being from a movie soundtrack, the pitch increases as the mood transitions to an uplifting tone (when Hiccup and Astrid fly across the horizon).
- Continuous horizontal bands at various frequencies denotes violin and piano parts. The shorter horizontal stretches are likely from the piano, as piano notes typically have shorter durations compared to violin's.
- Additionally, low-frequency lines stacked vertically may indicate guitar or bass sections. Notably, the absence of vertical striations, which were prominent in the previous two songs, confirms that this piece lacks vocal elements and consists solely of instrumental sounds.
- Around the 50–60 second mark, the spectrogram shows higher intensities at high frequencies, which likely corresponds to the uplifting mood shift in the movie scene.
- This spectrogram is the most different from all others, mainly because of the lack of vocal chords, and electric music. The effect of drums is much less than the violin and piano (no vertical stripes).

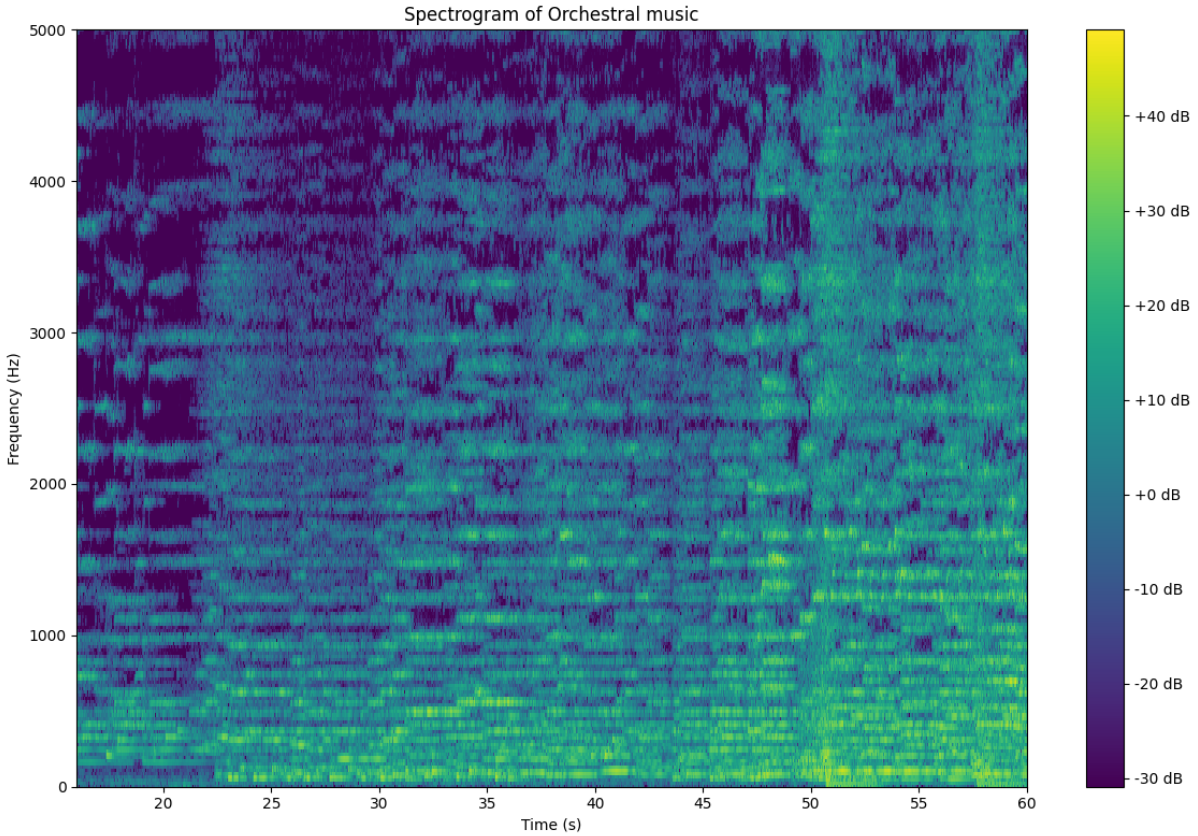

Figure 13: Spectrogram Analysis of Romantic Flight - John Powell

1.3.4 Masakali - A.R. Rahman, Mohit Chauhan (Upbeat Hindi song)

The song can be downloaded from Masakali.mp3.

- Vertical striations appear once again as people sing in this song.
- Horizontal stretches may represent sounds produced by synthesizers, piano, violin, or sarangi. The dark intensities indicate pauses in the song.
- There is a constant hum in the lower frequencies, with consistent horizontal frequencies and their harmonics possibly corresponding to piano or violin notes.
- The constant beat of the drums and the tempo of electronic music is very noticeable, as vertical stripes spanning various frequencies appears prominently.
- Around 37 seconds, a sustained hum becomes noticeable, which could signify a continuous instrumental sound or a steady drumming noise. The energy of percussive sounds like drums tends to be spread across the entire frequency range.

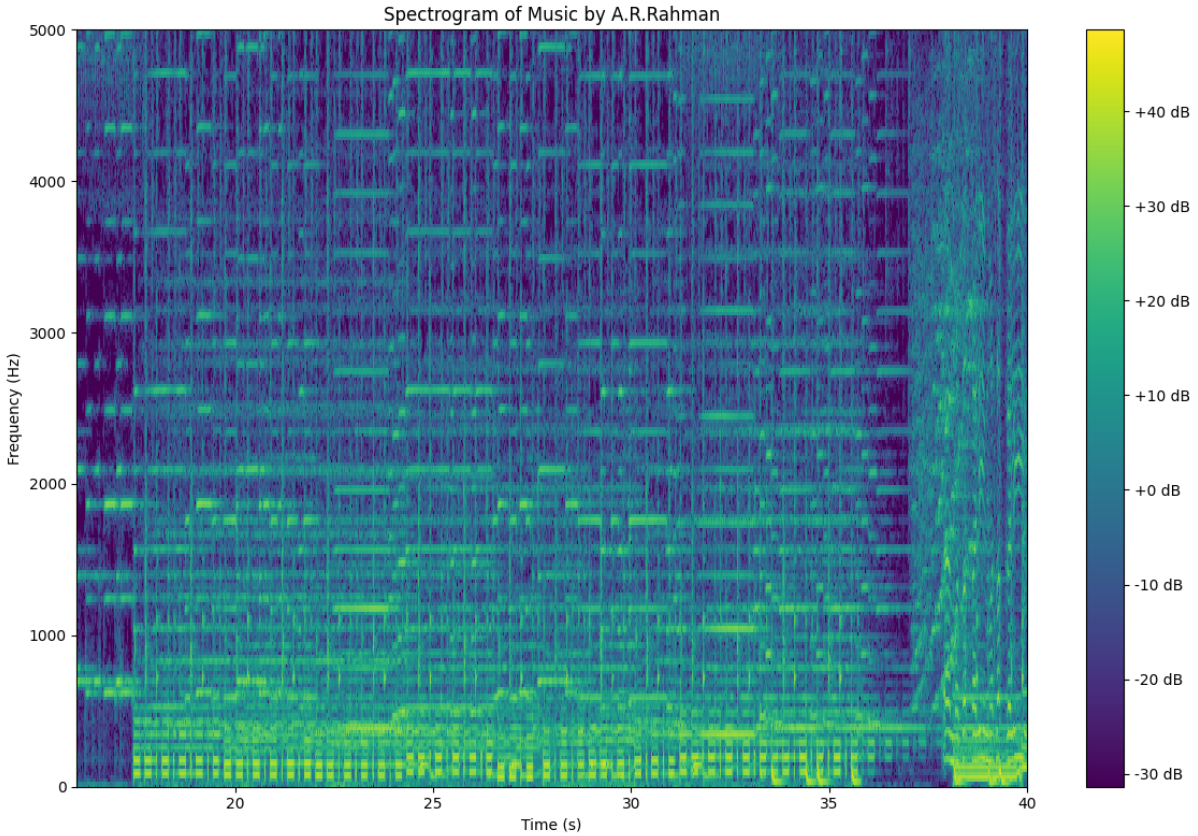

Figure 14: Spectrogram Analysis of Masakali - A.R. Rahman

References

[1] Zulfidin Khodzhaev. A practical guide to spectrogram analysis for audio signal processing. *arXiv preprint arXiv:2403.09321*, 2024.