Speech Understanding
Programming Assignment - 1

Mitesh Kumar
Indian Institute of Technology, Jodhpur
m23mac004@iitj.ac.in

GitHub Link:
https://github.com/mitesh-kr/Speech_Understanding_Assignment_1.git

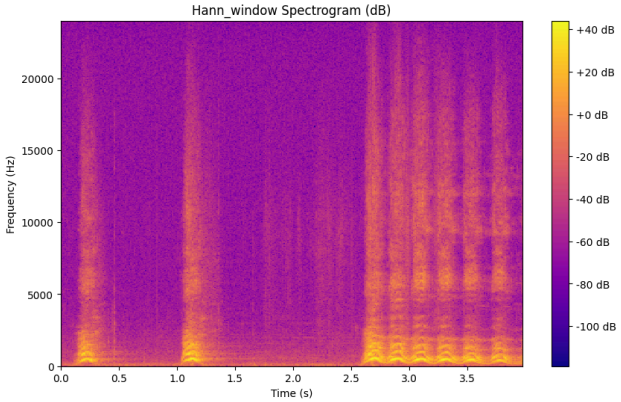
2. Task A. Experimenting with Spectrograms and Windowing Techniques on UrbanSound8k

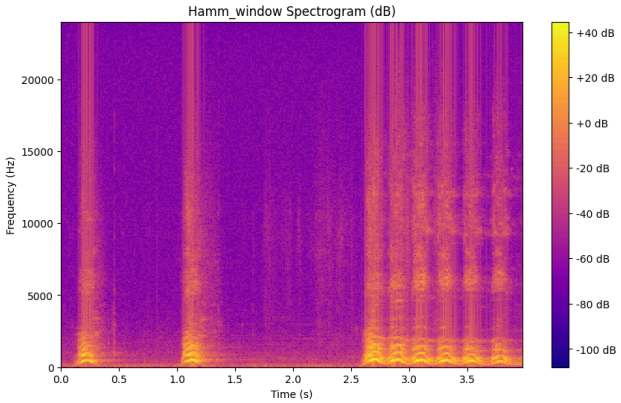
urbanSound8k is an audio dataset containing 8732 labelled audio samples from 10 different classes of urban sounds. For the classification task, a spectrogram of audio has been used, which is obtained by using the Short-Time Fourier Transform (STFT). Three different windowing techniques have been used to obtain spectrogram and there is a detailed analysis and comparison of classification tasks for the three methods.

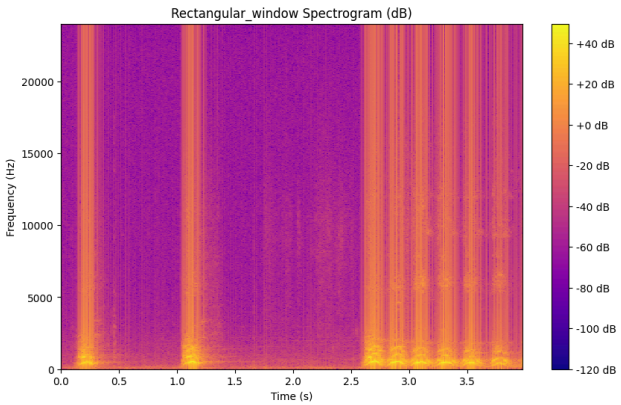
2.1 Data Preprocess pre_process.ipynb

This function computes the Short-Time Fourier Transform (STFT) of an audio file to generate a spectrogram. It starts by loading the audio file and converting it to mono if it has multiple channels.
The waveform is divided into frames using the win_length and hop_length parameters, ensuring that each frame has sufficient data for frequency analysis. For each segment, the function applies the Fast Fourier Transform (FFT) to convert the time-domain signal into frequency components. The magnitudes of these frequency components are stored in a spectrogram. For each audio file, three window techniques Hann Window, Hamming Window and Rectangular Window have been used keeping parameters n_fft=1024, win_length=1024 and hop_length = win_length // 2 to make sure the frames overlap by 50%. To visualise the spectrogram one sample from each class using these three window techniques has been used.

On visual inspection, it has been observed that the sharpness of the spectrogram obtained from Hann is smooth, while for Hamm it is sharper than Hann and the rectangular window produces very sharp frequency bands.







2.2 Feature Extraction

feature_extraction.ipynb

for each audio file spectrogram tensor has been obtained using 3 different windowing techniques resizing the tensors into 224 x 224 and saving them so that they could be used to train a neural network for future tasks.

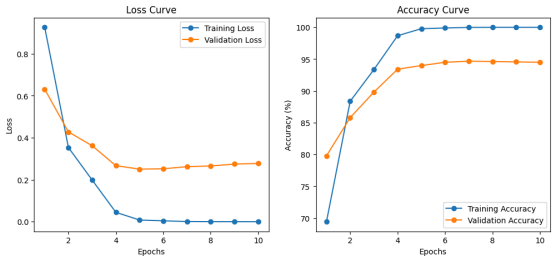
2.3 Classification task

for the classification task pretrained model vit_tiny_patch16_224 has been used. It is trained on the ImageNet dataset. Below are the results for the three window techniques-based classification task.

Hann window-based classification

hann_window_based_classification.ipynb

Results after 10 epochs

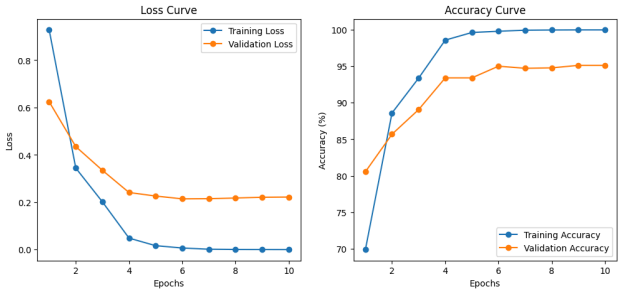


Train Loss	Train Acc(%)	Val Loss	Val Acc(%)
0.0003	100.00	0.2779	94.50

Hamming window-based classification

hamming_window_based_classification.i…

Results after 10 epochs

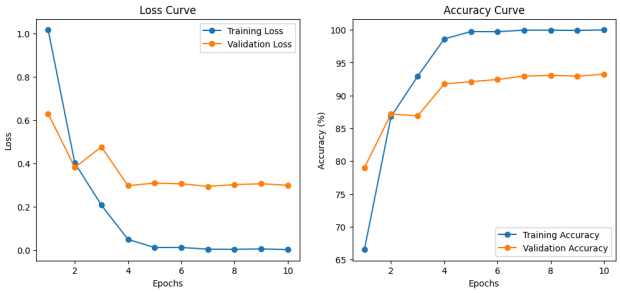


Train Loss	Train Acc(%)	Val Loss	Val Acc(%)
0.0003	100.00	0.2221	95.13

Rectangular window-based classification

Rectangular_window_based_classification.ipynb

Results after 10 epochs



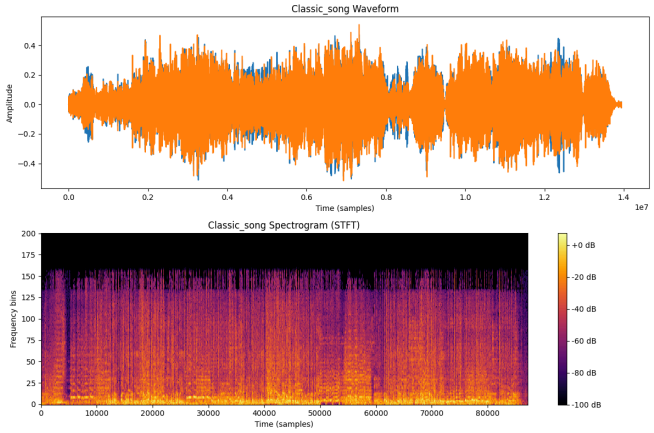
Train Loss	Train Acc(%)	Val Loss	Val Acc(%)
0.0009	100.00	0.2980	93.25

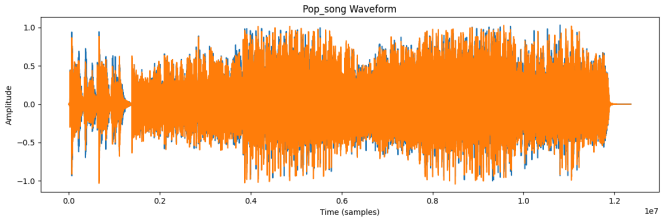
Based on the above analysis of three methods. classification result obtained using Hamm window-based is generating the best result with validation accuracy of **95.13 %** followed by Hann window-based accuracy of **94.50%** and then **93.23 %** accuracy generated by the rectangular window-based

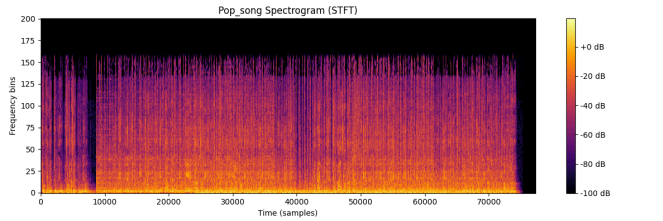
2. Task B. Spectrograms analysis of 4 different songs from different genres

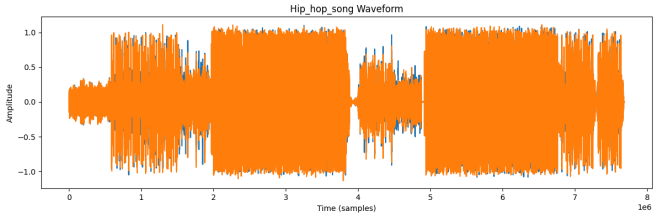
Q2_TaskB.ipynb

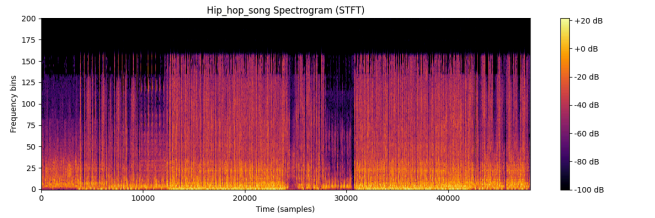
one song from each genre classical, Hip-hop, Pop and rock has been chosen for spectrogram analysis.
all four songs have sample rate of 48000 Hz

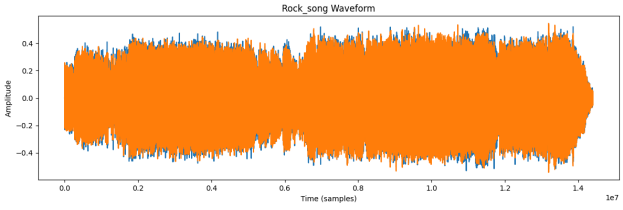


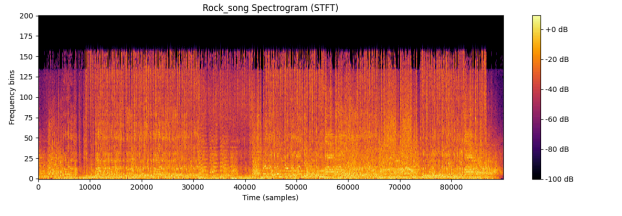












On visual inspection of these four spectrograms. Below are the observations table

Classical Music

Classical music has a wide range of volumes, from quiet to loud, with smooth transitions. The low frequencies (0-5000 Hz) have deep bass but with more variation than other genres. In the mid-range (5000-13000 Hz), the music shows subtle changes, with different instruments creating a rich sound. The high frequencies (13000-24000 Hz) are soft and delicate, adding clarity to the music. Classical songs have natural flow and changes, giving them a detailed, organic feel.

Hip-hop

Hip-hop features strong bass in the low frequencies (0-5000 Hz), creating a solid foundation. In the mid-range (5000-13000 Hz), there are rhythmic patterns and breaks that keep the music energetic. The high frequencies (13000-24000 Hz) are sharp, with snare drums and vocals standing out. The dynamic range is moderate, with clear breaks in the music. Hip-hop has regular rhythms and patterns that keep the track lively and engaging.

Pop Music

Pop music has balanced bass in the low frequencies (0-5000 Hz), with a consistent sound throughout. The mid-range (5000-13000 Hz) stays steady, supporting catchy melodies. In the high frequencies (13000-24000 Hz), the sound is bright and clear, with sparkling highs. Pop music has a smooth, consistent volume throughout, making it easy to listen to. The song structure is simple and repetitive, making pop music easy to follow and enjoy.

Rock Music

Rock music has strong, continuous energy across all frequencies. The bass (0-5000 Hz) provides a solid foundation, while the mid-range (5000-13000 Hz) has a thick sound from guitars and drums. The high frequencies (13000-24000 Hz) are aggressive, with sharp cymbals and guitar solos. Rock music has a dense, powerful sound with fewer quiet parts, and the song structure is continuous, keeping the energy high throughout.

Comparison Report

When comparing the four genres, classical music is the most detailed and complex with natural changes. Hip-hop has strong rhythms and energetic breaks that keep the music moving. Pop music is smooth and easy to follow, with consistent sounds. Rock music is powerful and intense, with lots of energy and fewer quiet moments. Each genre has its own unique style, offering different listening experiences based on how the music is made and the sounds used.

References

[1] Salamon, J., Jacoby, C., & Bello, J. P. (2014). *A dataset and taxonomy for urban sound research.* In *Proceedings of the 22nd ACM International Conference on Multimedia* (pp. 1041–1044). ACM.