A PROJECT FOR COMPUTER MUSIC

AUDIO DATA ANALYSIS AND MUSIC GENRE CLASSIFICATION

BY ELEFTHERIA ELLINA

1115201800228

This project intends to analyse audio files and identify their music genre using tools that are suitable for the modern times we live in. The music genre classification problem can be resolved with machine learning/ deep learning techniques. Music genres are class holding categories that represent styles or classifications of music, useful in identifying and organizing similar musical artists or recordings. Genres differ from each other due to certain musical-audio features. The growing amount of music that is available electronically, particularly in the World Wide Web, creates an arising need for automated classification. Among others, I tried to have the best possible classification results using an Artificial Neural Network. At first, a sound is going through analysis and the most important features are extracted for us to see and understand the information that are used for the classification. Afterwards, my model is being trained with 1000 audios, 100 audios of 10 genres, taken from the well-known paper in genre classification “Musical genre classification of audio signals” by G. Tzanetakis and P. Cook in IEEE Transactions on Audio and Speech Processing 2002. At last, five popular songs of different genres are assigned as the test samples and the model identifies their genre, fast and accurately via Python programming language on the Google Colab platform.

PART 1:

Audio Analysis and Processing

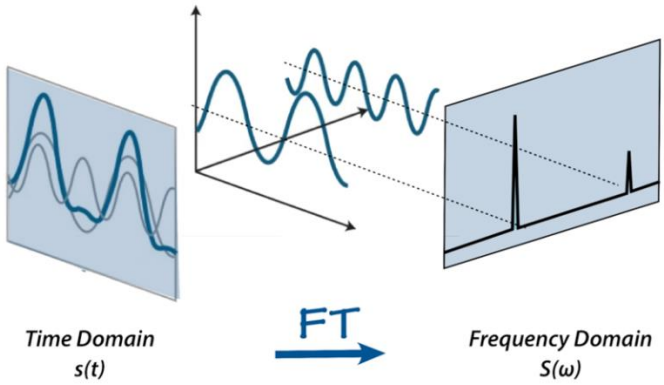
Audio analysis and signal processing have benefited greatly from machine learning and deep learning techniques.

Basics of audio analysing:

Sound travels in waves that propagate through vibrations in the medium the wave is traveling in. Sound, doesn't travel in empty space. Sound waves can be imagined as pressure waves by understanding the idea of compressions and rarefactions. A wavelength is the distance between two consecutive compressions or two consecutive rarefactions. Frequency, or pitch, is the number of times per second that a sound wave repeats itself. Then the velocity of a wave is the product of the wavelength and the frequency of the wave. [1]

Fourier Transform:

An audio signal is a complex signal composed of multiple 'single-frequency sound waves' which travel together as a disturbance(pressure-change) in the medium. When sound is recorded we only capture the resultant amplitudes of those multiple waves. Fourier Transform [2] is a mathematical concept that can decompose a signal into its constituent frequencies. Fourier transform does not just give the frequencies present in the signal, It also gives the magnitude of each frequency present in the signal.



Short-time Fourier Transform:

The Short-time Fourier transform (STFT)[3], is a Fourier-related transform used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time.[1] In practice, the procedure for computing STFTs is to divide a longer time signal into shorter segments of equal length and then compute the Fourier transform separately on each shorter segment. This reveals the Fourier spectrum on *each shorter segment. One then usually plots the changing spectra as a function of time, known as a spectrogram or waterfall plot, such as commonly used in Software Defined Radio (SDR) based* spectrum displays. Full bandwidth displays covering the whole range of an SDR commonly use Fast Fourier Transforms (FFTs) with 2^24 points on desktop computers.

Spectogram:

A spectrogram[2] is a visual way of representing the signal strength, or “loudness”, of a signal over time at various frequencies present in a particular waveform. Not only can one see whether there is more or less energy at, for example, 2 Hz vs 10 Hz, but one can also see how energy levels vary over time. A spectrogram is usually depicted as a heat map, i.e., as an image with the intensity shown by varying the color or brightness.

Zero-crossing Rate:

Zero-crossing rate[4] is a very simple way for measuring smoothness of a signal is to calculate the number of zero-crossing within a segment of that signal. A voice signal oscillates slowly - for example, a 100 Hz signal will cross zero 100 per second - whereas an unvoiced fricative can have 3000 zero crossing per second. To calculate the zero-crossing rate of a signal you need to compare the sign of each pair of consecutive samples.

Spectral Centroid:

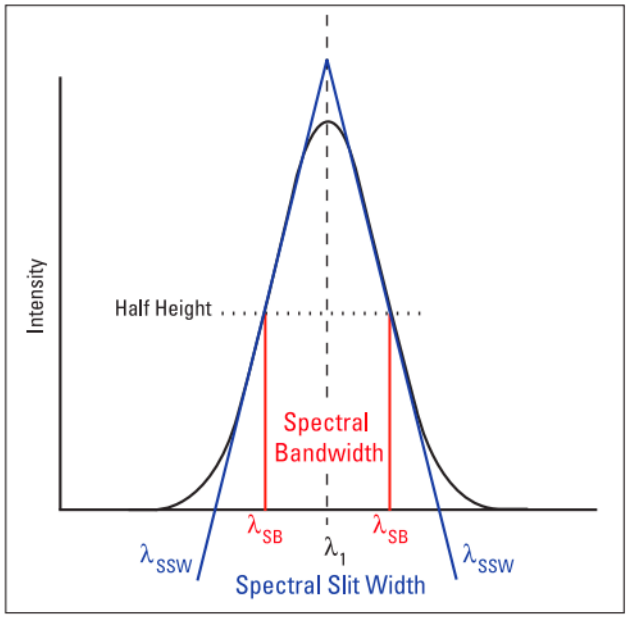
The spectral centroid indicates at which frequency the energy of a spectrum is centered upon or in other words It indicates where the ”center of mass” for a sound is located.

Spectral Rolloff:

The spectral rolloff is a measure of the shape of the signal. It represents the frequency at which high frequencies decline to 0. To obtain it, we have to calculate the fraction of bins in the power spectrum where 85% of its power is at lower frequencies.

Spectral Bandwidth:

Spectral bandwidth[5] is the difference between the upper and lower frequencies in a continuous band of frequencies. As we know the signals oscillate about a point so if the point is the centroid of the signal then the sum of maximum deviation of the signal on both sides of the point can be considered as the bandwidth of the signal at that time frame.



MFCCs:

MFCCs or Mel Frequency cepstral coefficients [4] have become a popular way of representing sound. In a nutshell, MFCCs are calculated by applying a pre-emphasis filter on an audio signal, taking the STFT of that signal, applying mel scale-based filter banks, taking a DCT (discrete cosine transform), and normalizing the output. In other words MFCCs are a small set of features (usually about 10–20) which concisely describe the overall shape of a spectral envelope. It models the characteristics of the human voice.

Chroma Feature:

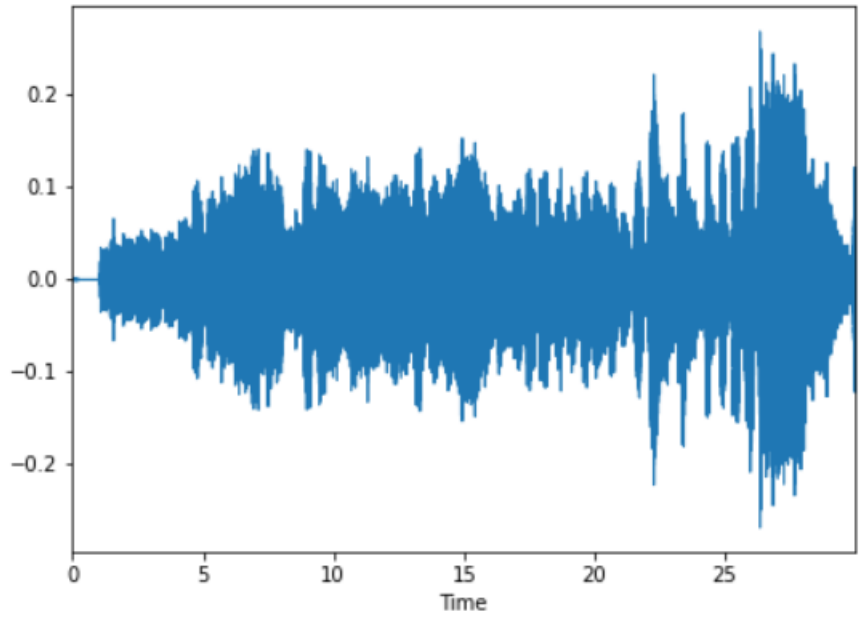
Chroma features[6] are an interesting and powerful representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave. Since, in music, notes exactly one octave apart are perceived as particularly similar, knowing the distribution of chroma even without the absolute frequency (i.e. the original octave) can give useful musical information about the audio -- and may even reveal perceived musical similarity that is not apparent in the original spectra.

Results of Audio Analysis:

Audio Used:

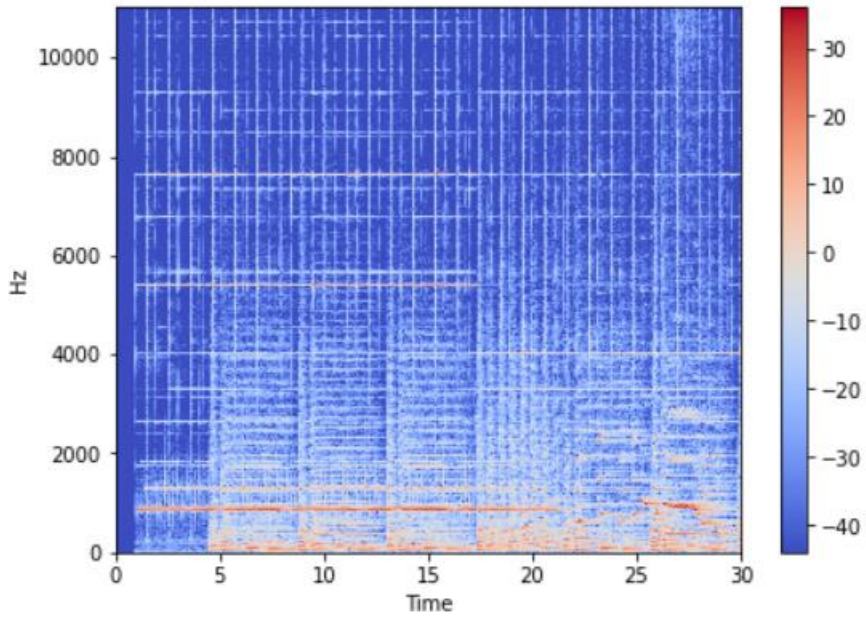
pink_panther.wav, aka "The pink panther theme", classified in jazz genre

Sound Wave:



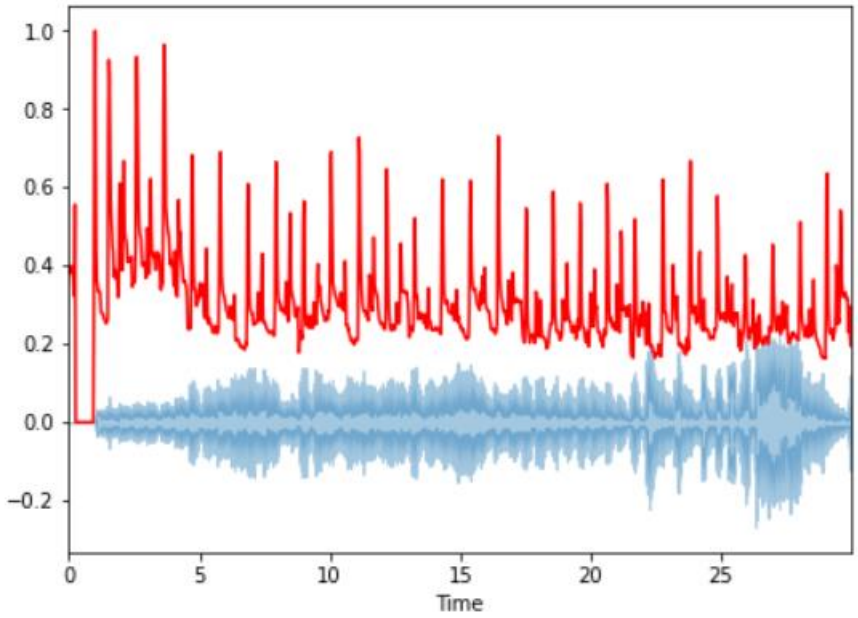
Spectogram

Using Short-Time Fourier Transform:

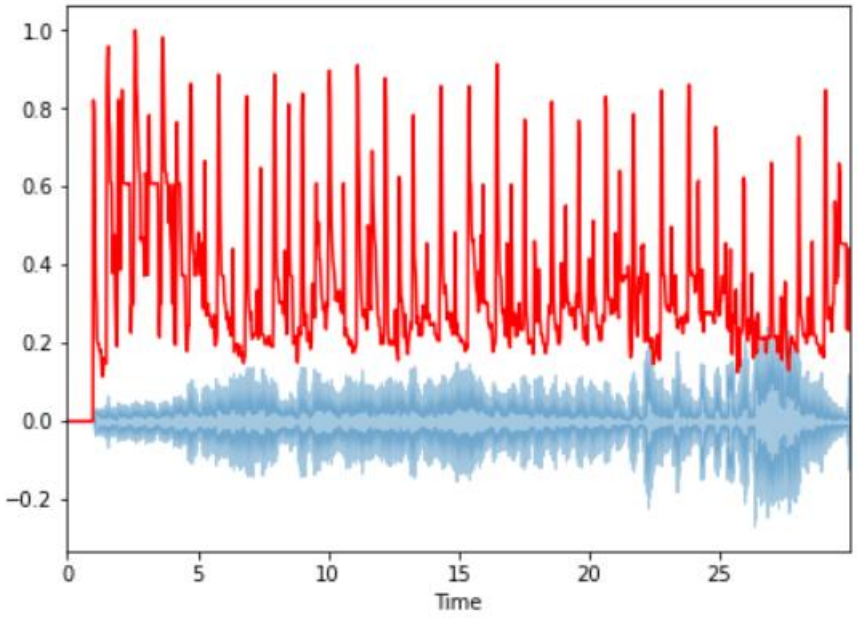


Extracted Features:

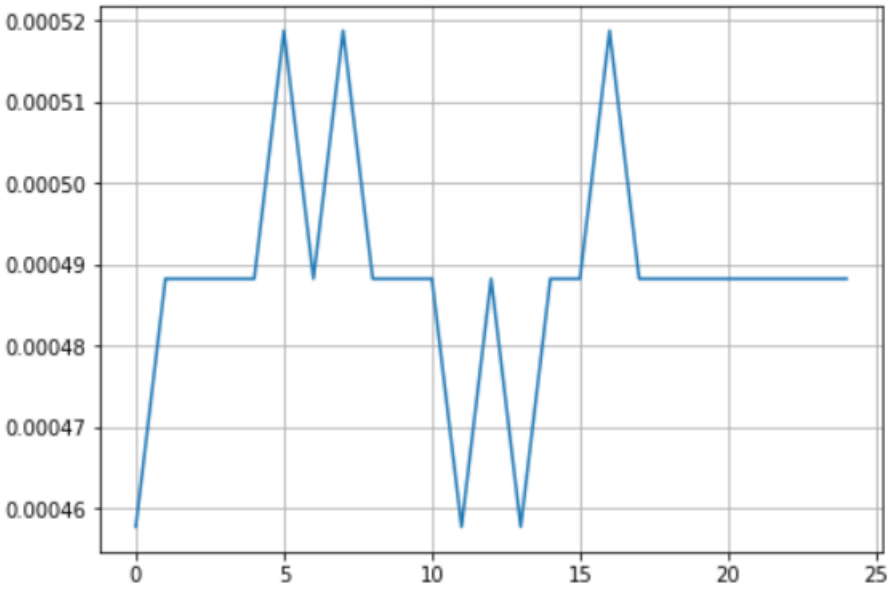
Spectral Centroid



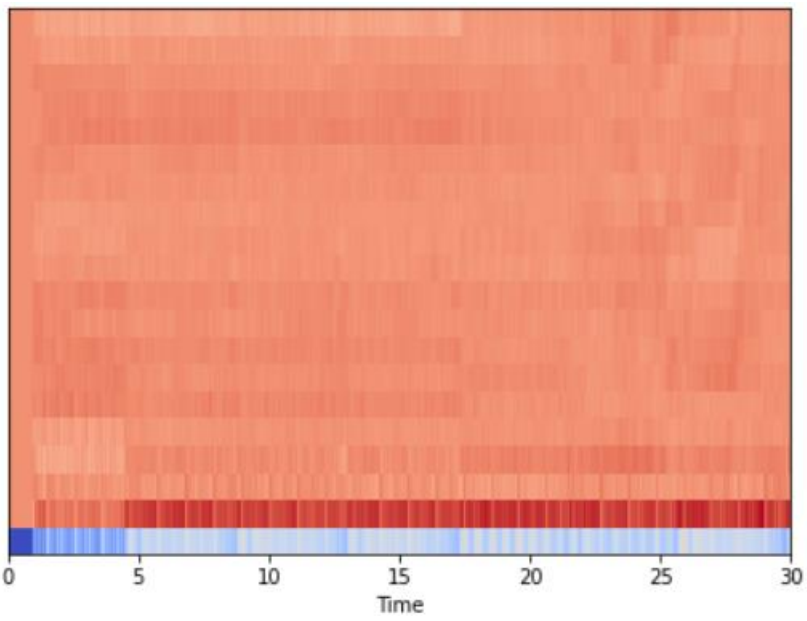
Spectral Roll-off



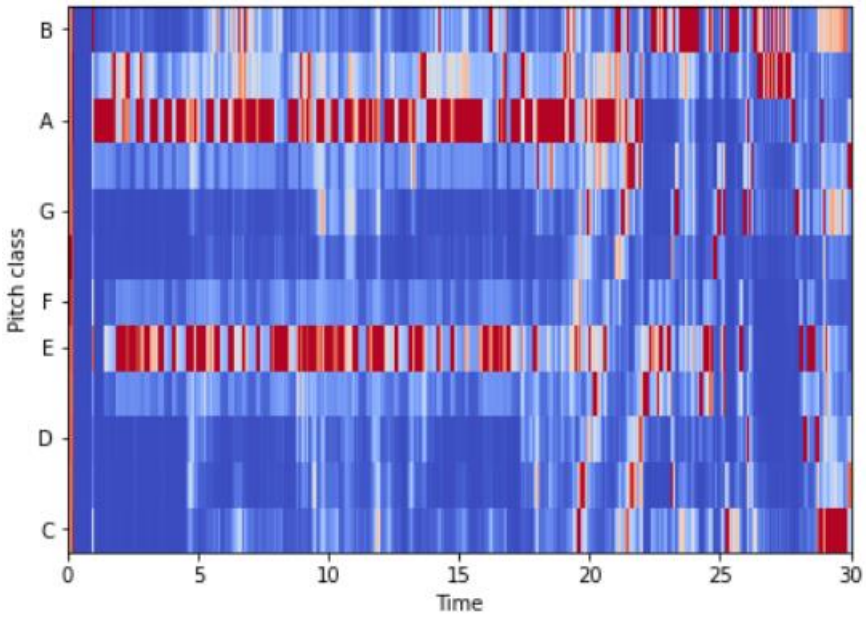
Zero-Crossing Rate



MFCCs



Chroma



PART 2:

Music Genre Classification

Artificial Neural Network:

ANN models[7] are the extreme simplification of human neural systems. An ANN comprises of computational units analogous to that of the neurons of the biological nervous system known as artificial neurons. Mainly, the ANN model constitutes of three layers, viz., input, hidden, and output. Each neuron in the *n*th layer is interconnected with the neurons of the (*n* + 1)th layer by some signal. Each connection is assigned a weight. The output may be calculated after multiplying each input with its corresponding weight. The output passes through an activation function to get the final ANN output. The ANN may be useful in solving different engineering and science problems. As such, the ANN has various applications, i.e. image compression, function approximation, differential equations, stock market prediction, medical diagnosis, and signal processing.

Model Preparation and Training:

1. Creating a header for the CSV file that will hold the features for every audio.
2. Extracting features with librosa library: Mel-frequency cepstral coefficients (MFCC), Spectral Centroid, Zero Crossing Rate, Chroma Frequencies, and Spectral Roll-off.
3. Inserts all features of every audio to the CSV file.
4. Importing the songs that will be assigned as the single tests of the model.
5. Doing the same procedure for the single tests.
6. Data preprocessing: loading CSV data, label encoding, feature scaling and data split into training and test set.
   This means that the model will be trained with a part of the data set, and the test set that is created here is for the general accuracy calculation of the model.
7. Building an ANN model.
8. Training the model with 20 epochs and batch size 128.
9. Evaluating the model and predicting the genre of the imported songs.

Results of Music Genre Classification:

Prediction Results

Test Loss	0.5497
Test Accuracy	80.00%

Classification Report

The Classification Report confirms the prediction on test set results. It is verified that the model classifies the test set audio files with almost 80% accuracy. This is a decent accuracy score, which implies the model will most probably correctly predict the genre of an unknown audio files. In fact, we can verify this with the single tests prediction results:

```
Correct results 4.
Incorrect results 1.

Music genre of selected sound pink_panther.wav is jazz.
Music genre of selected sound smoke_on_the_water.wav is rock.
Music genre of selected sound new_rules.wav is pop.
Music genre of selected sound redemption_song.wav is reggae.
```

The five imported songs are:

1. The Pink Panther Theme
2. Smoke on the water by Deep Purple
3. New rules by Dua Lipa
4. Redemption song by Bob Marley
5. Paranoid by Black Sabbath

4 out of 5 were classified in the right genre as we can see. Paranoid by Black Sabbath was predicted with the incorrect genre label so it’s not printed among the right answers.

References:

[1]https://blog.paperspace.com/introduction-to-audio-analysis-and-synthesis/

[2] https://towardsdatascience.com/understanding-audio-data-fourier-transform-fft-spectrogram-and-speech-recognition-a4072d228520

[3] https://en.wikipedia.org/wiki/Short-time_Fourier_transform#:~:text=The%20Short%2Dtime%20Fourier%20transform,as%20it%20changes%20over%20time.

[4] https://wiki.aalto.fi/display/ITSP/Zero-crossing+rate

[5] https://analyticsindiamag.com/a-tutorial-on-spectral-feature-extraction-for-audio-analytics/

[6] https://www.ee.columbia.edu/~dpwe/resources/matlab/chroma-ansyn/

[7] https://www.sciencedirect.com/topics/engineering/artificial-neural-network-model#:~:text=ANN%20models%20are%20the%20extreme,input%2C%20hidden%2C%20and%20output.