Music Genre Classification Using Deep Learning

Orhan Örs
Department of Computer Engineering
Ege University
İzmir,Turkey
orsorhan1@gmail.com

Abstract:

In this paper we will present basic audio processing features and how to use them with CNN deep learning architecture to predict audio genre. The aim of this work is to predict the genre of song by using CNN deep learning technique which is commonly used to image processing. For this purpose, feature extraction is done by using signal processing techniques, then CNN algorithm is applied with MFCC feature which helps to find tonal feature of music.

I. INTRODUCTION

A music genre is a conventional category that identifies some pieces of music as belonging to a shared tradition or set of conventions[1]. Music can be divided into different genres in many different ways, such as into popular music and art music, or religious music and secular music.

II. DOMAIN REQUIREMENTS

A music consists of a lot of audio signal that gives information about the general structure of the audio. Common presentation of an audio signal is a waveform that displays amplitude or level changes over time.

Figure 1. Audio waveform

Audio waveform(Fig. 1) does not include precious information about general structure of music. Audio is processed to get detailed information about music.

Details of the audio will be discussed in the following subsections.

A) Fourier Transform

The *Fourier transform* is an operation that transforms data from the time domain into the frequency domain.

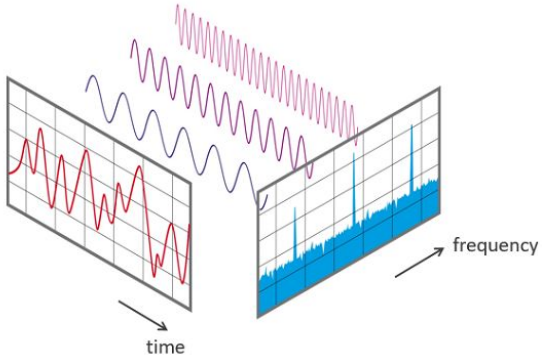

Figure 2. Audio Fourier Transformation

Result of Fast Fourier Transformation(FFT) shows Magnitude-Frequency values of audio.

A.1)Short Time Fourier Transform

The Short-time Fourier transform (STFT), 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[2].

STFT computes n.Fourier Transform (n: frame size) and preserves information about time, frequency and magnitude.

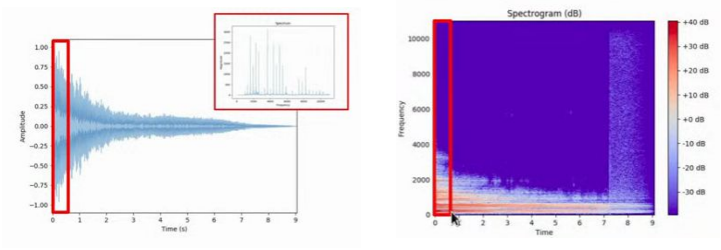

Figure 3. Representation of STFT for first frame of audio, from right to left: Audio waveform, frequency-magnitude graph, spectogram

B) MFCC

MFCC stands for Mel frequency cepstral coefficients.

Our ear has cochlea which basically has more filters at low frequency and very few filters at higher frequency. This can be mimicked using Mel filters. So the idea of MFCC is to convert time domain signals into frequency domain signal by mimicking cochlea function using Mel filters.

(Fig. 4) MFCCs are commonly derived as follows:[3]

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

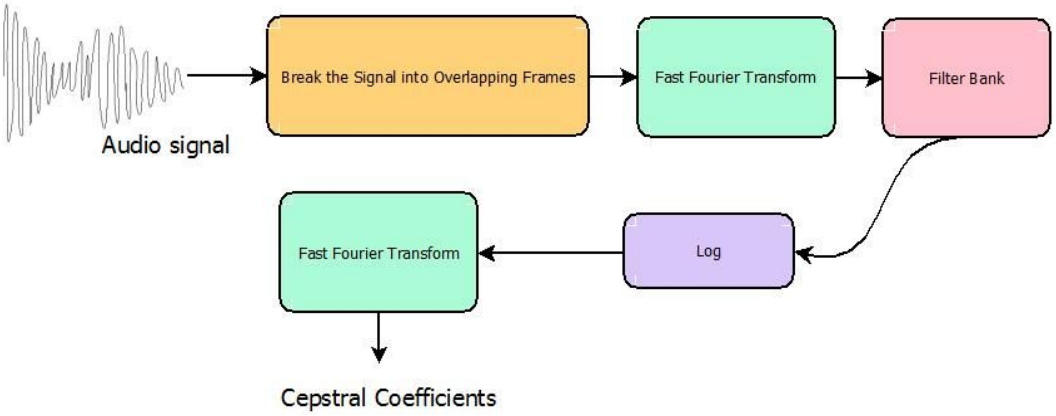

Figure 4. MFCC Process

III. DATASET

This dataset was used for 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.[4]

The dataset contents are described below:

- 10 genres
- Each genre represented by 100 tracks
- Each track 30 seconds long
- 1000 audio tracks

and all tracks are 22050Hz Mono 16-bit audio files in .wav format.

IV. IMPLEMENTATION

A) DATASET PREPARATION

```
data = {
    "genres": [],
    "labels": [],
    "mfcc": []
}
```

Figure 5. Dataset requirements

Dataset contains 10 genres, genre labels and MFCC features. All features are saved in json file.

Generally, MFCC coefficients are between 10-13. In this project parameters are specified as shown in below:

- Hop Length : 512
- Number of FFT : 2048
- Number of Segments: 20
- Sample Rate : 22050

$$Total\ Sample\ =\ Sample\ Rate * Duration\ of\ Track$$

$$Each\ segment\ sample\ =\ Total\ Sample * Sample\ Rate$$

Figure 4. Formula of segment size

Each intervals in a track consists of samples and group of samples named as segment. In dataset every segment act like a track.

B)NETWORK ARCHITECTURE

We build a CNN neural network (3 hidden layers) as the fundamental structure. We than train CNN for each network layer except the output layer as the weight initialization.

CNN is commonly used in image processing. Images consists 3 dimensional array (width,height and depth) to train with CNN. The key idea using CNN with audio is to reshape audio segments like a grayscale image. Reshaping steps of audio matrix described below:

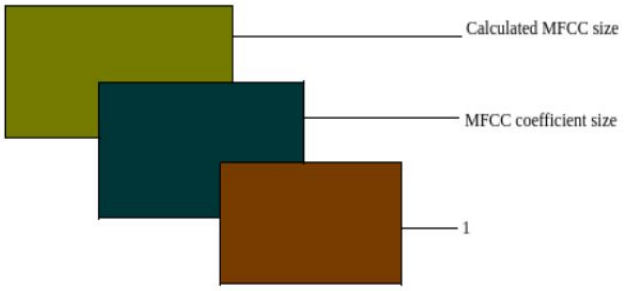

Figure 6. Array shapes used to train and prediction on CNN architecture

In the preprocessing step, we further divided the audio file 20 equal segments and extracted each of MFCC features. In general, we generated a 13 × 65 = 845 length of MFCC features stored in (65,13,1) matrix to represent a 30 second audio file for the later experiment.

The structure described above is similar to the grayscale image which represented with (width,height,1) matrix. In conclusion, we decided to use CNN architecture to train.

C)EXPERIMENT SETUP

We use 75% of mfcc features as training set, and the rest 25% as testing set. We then split training dataset into 80% training and %20 validation set.10 genres are equally weighted, each has 100 samples.

C.1) Avoid Overfitting

We used 3 convolution layer with activation function Rectified Linear Unit (ReLU), 2x2 strides and Batch Normalization. Then feed results with a neural network which has 64 hidden layers and then add dropout which is a regularization technique for neural network models.

V. RESULTS

As a result we achieved 85 % training set accuracy with %40 loss. Also test set achieved %75 accuracy. Accuracy and error changes described in Figure 7.

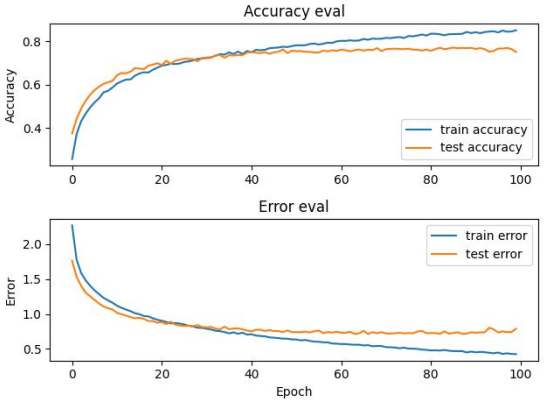

Figure 7. Performance metrics on result

VI. Future Works

Dataset that used in this project includes 10 different genres. As we know in music domain a lot sub-genres also exists. Our goal is extend dataset and testing different neural network approaches to increase accuracy and decrease error.

REFERENCES

[1] Samson, Jim. "Genre". In Grove Music Online. Oxford Music Online. Accessed March 4, 2012.

[2] Sejdić E.; Djurović I.; Jiang J. (2009). "Time-frequency feature representation using energy concentration: An overview of recent advances". *Digital Signal Processing*. 19 (1): 153–183

[3] Sahidullah, Md.; Saha, Goutam (May 2012). "Design, analysis and experimental evaluation of block based transformation in

[4] Marsyas, “GTZAN Genre Collection”, marsyas.info/downloads/datasets.html

[5] Machinelearningmastery, “Difference Between a Batch and an Epoch in a Neural Network” , https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/