**CSCI 6908**-Deep Learning Summer 2018

**Project**: Audio Tagging

Submission date: 07/Ago/2018

Student ID: B00735030 - Student Name: Miria Rafante Bernardino

Contents

[Abstract 3](#_Toc519672490)

[Introduction 4](#_Toc519672491)

[Literature review 4](#_Toc519672492)

[Data Preparation 4](#_Toc519672493)

[Methods 4](#_Toc519672494)

[Results 5](#_Toc519672495)

[Conclusion 6](#_Toc519672496)

[References 7](#_Toc519672497)

## Abstract

## Introduction

## Literature review

## Data Preparation

### Dataset

The dataset is composed by 9473 samples unequally distributed among 41 categories. According to DCASE [1], “the minimum number of audio samples per category in the train set is 94, and the maximum 300. The duration of the audio samples ranges from 300ms to 30s ...”. Considering the rate is 44.1 kHz, 44100 values per second, and that the recording time is different for each sample, the number of points per sample variety drastically. The greater sample has 1323000 values.

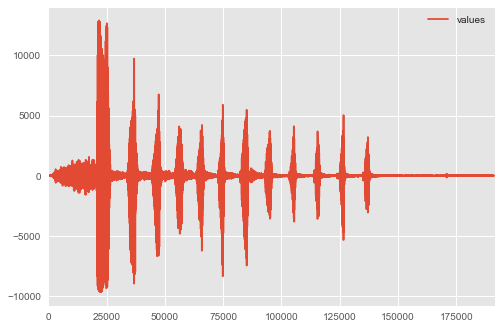


Figure 1 - Plotting a Hi-hat sound sample. The x-axis is time in ms.

Figure 1 and Figure 2 show two samples’ values (Hi-hat and a Saxophone) over time (ms).

Some of the samples’ label were manually verified and some were not. For treat this problem we are using only the verified samples, in other words, we are using only the 3710 samples that were verified manually.

### Sounds properties

Sounds are waves propagating over a physical matter [3]. The height (or amplitude) of the waves and the number of waves flowing by second (or frequency) are examples of the properties of a sound [2]. The higher the amplitude, the more energy the wave has, and intensity is the unity to measure the amount of energy a wave has in a given area [2]. Besides, tones, overtones, harmonics, speed of sound, timbre, loudness, etc., are other properties of the sounds and they variety according to its source [2][3].

Therefore, to identify the source of a sound, we need to study how the waves are changing over time and we also need a minimum period of time to be able to observe such changes [3], and we did that by describing the samples in function of some of its properties.

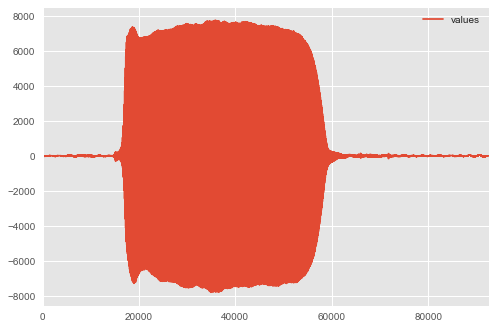


Figure 2 - Plotting a Saxophone sound sample. The x-axis is time in ms.

### Describing the data

We used the 4 functions shown in Figure 3 from librosa [4] package for feature extraction to describe the original data: tonnetz (computes the tonal centroid features (tonnetz)), spectral\_centroid (computes a 6D description of chords), spectral\_bandwidth and mfcc (computes the Mel-frequency cepstral coefficients (MFCCs)).

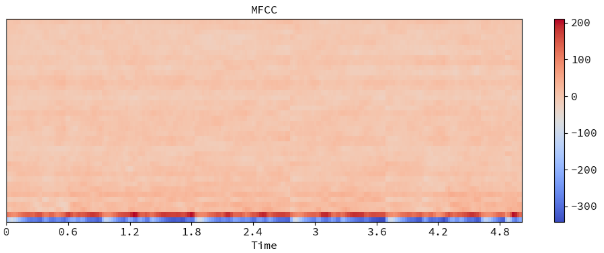
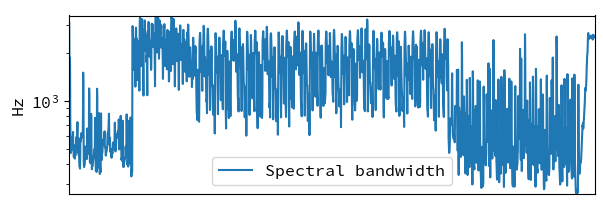
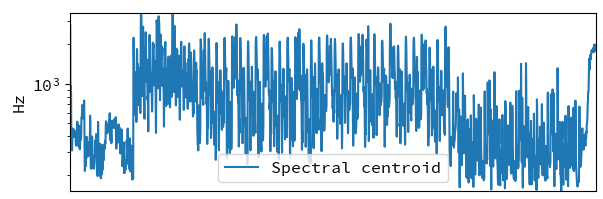
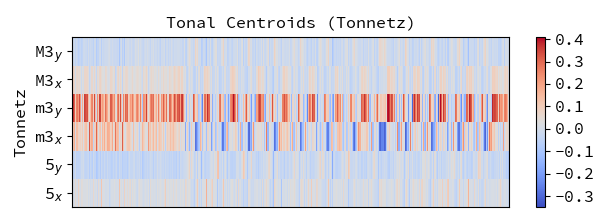
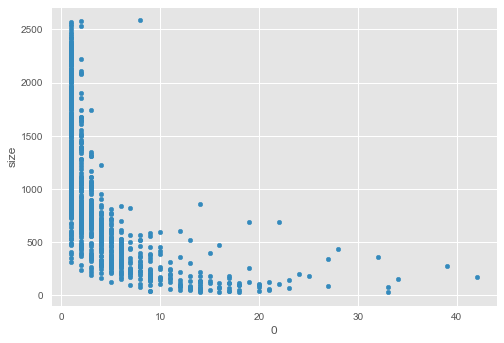


Figure 3 - Vizualization of some librosa's functions for feature extraction: (top-left) Tonnetz, (top-right) Spectral centroid, (bottom-left) MFCC and (bottom right) Spectral bandwidth. Source: https://librosa.github.io/librosa/index.html

Padding – paste the graphics here



## Methods

## Results

|  |  |  |  |
| --- | --- | --- | --- |
| Data description | MaxSize | RF – 3-fold | MLP- 3-fold |
| Tone | 1000 | Scores: [0.2971246, 0.31012146, 0.29681112]  Mean: 0.30135239277503162  Std: 0.0062019856754040002 | Scores: [0.0686901, 0.06720648, 0.06950123]  Mean: 0.068465933357223716  Std: 0.00095014194851230651 |
| Spectral Centroid | 1000 | Scores: [0.43210863, 0.43724696, 0.4480785]  Mean: 0.4391446950878993  Std: 0.006656336335106039 | Scores: [ 0.0686901, 0.06882591, 0.0678659]  Mean: 0.06846063676458795  Std: 0.00042417933477714254 |
| Spectral Bandwidth | 1000 | Scores: [ 0.44888179, 0.43076923, 0.43254293]  Mean: 0.4373979823782462  Std: 0.008152499127789991 | Scores: [ 0.0686901, 0.06720648, 0.06950123]  Mean: 0.06846593335722372  Std: 0.0009501419485123065 |
| MFCC | 1000 | Scores: [0.64616613, 0.65101215, 0.65494685]  Mean: 0.65070837731252074  Std: 0.0035911426034004267 | - |
|  |  |  |  |

|  |  |  |
| --- | --- | --- |
| Data description | MaxSize | RF – 10-fold |
| MFCC | 1000 | Scores: [0.67792208, 0.66318538, 0.67810026, 0.63852243, 0.7037037, 0.67204301, 0.67847411, 0.66111111, 0.64145658, 0.69714286]  Mean: 0. 67116615275898628  Std: 0. 019986117622205764 |

## Conclusion

## References

1. Dcase. "General-purpose Audio Tagging of Freesound Content with AudioSet Labels." General-purpose Audio Tagging of Freesound Content with AudioSet Labels - DCASE. Accessed July 19, 2018. http://dcase.community/challenge2018/task-general-purpose-audio-tagging.
2. "The Components of Sound." Conductors and Insulators. Accessed July 19, 2018. <https://www.nde-ed.org/EducationResources/HighSchool/Sound/components.htm>.
3. "Sound." Wikipedia. July 15, 2018. Accessed July 19, 2018. <https://en.wikipedia.org/wiki/Sound>.
4. "LibROSA." LibROSA - Librosa 0.6.0 Documentation. Accessed July 19, 2018. https://librosa.github.io/librosa/index.html.