University of Porto
Faculdade de Engenharia

Assignment A2:

Low-Level features and timbre characterization

Alena Tesařová (up201911219)
Stylianos Tsagkarakis (up2019112311)

May 13, 2020

Contents

1 Introduction

The goal of this assignment is to understand the low-level features of different types of signals (percussive, sustained, low-pitch etc.) and analyse their distribution over a collection of sounds, which are samples of isolated notes from musical instruments. For this assignment we have been using python (Librosa, Essentia) and MIR plugin in Sonic visualizer.

In Task 1 we chose 2 descriptors from each domain. The first is **Zero crossing rate** and the second is **Spectral centroid**. In Task 2 we plotted all the desire descriptors for all sounds and we studied the characteristic for each sound. In Task 3 there are 2D plots visualizing the values of 2 descriptors for different samples and here we tried to find common characteristic for percussive/non-percussive sounds, sustained/non sustained, low-pitch/high pitch, and instrument. Finally in Task 4 there are examples of application and in task 5 formalizing/setting the rule of different groups of sounds.

2 Task 1

Please pick 2 descriptors by group (one from time-domain and another from frequency-domain), depart from the formula and explain the expected values for a sinusoid and white noise. Calculate these values and comment on your results. If they're not implemented in your software library, please find that implementation in another library (e.g. MPEG7 Matlab) and use it.

2.1 Zero crossing rate

Zero crossing rate of each frame is calculated as:

$$z_i = \frac{1}{2N} \sum_{n=0}^{N-1} |sgn(x_i[n]) - sgn(x_i[n-1])| \tag{1}$$

- $x_i[n] = 0, 1, \ldots N-1$ be the sample of the i^{th} frame

Explanation for sinusoid

To explain better the expected values for sinusoid, let's look closer at the formula. Signum function returns either -1 for negative x, 0 for 0 or +1 for positive x. For the sinus function we will get the the same amount of 0,1,-1 (because it is a periodical function). The difference between 2 neighbour values will be either 2,1 or 0. At most cases we get 0 and then depending on the frequency we get some values of 2. Lower frequency is, lower amount of amount of value 2. Finally, the mean value will be very close to zero. If the frequency is low, it will be closer to 0 than for higher frequencies.

$$z_i = \frac{1}{2N} \sum_{n=0}^{N-1} |sgn(x_i[n]) - sgn(x_i[n-1])| =$$

$$= \frac{1}{2N}(m * |1-1| + m * |-1-(-1)| + k * |-1-1| + k * |-1-1|) =$$

$$= \frac{1}{2N}(2k * 2) = \frac{2k}{N} = \frac{2k}{2 * m + 2 * k} = \frac{k}{m+k}$$

- k is the number of samples of transitions between positive and negative values of the signal
- m is the number of samples on positive or negative axis y of signal
- it is clear that:

$$m > k$$

I did some calculation to prove this statement. For frequency 100 Hz, I was getting the zero crossing rate around 0.009 and for frequency 1000 Hz, the values were around 0.09. The results can be also seen in the code attached to this file.

Explanation for white noise

White noise is a signal made of uncorrelated samples, such as the numbers produced by a random generator. It contains all frequencies in equal proportion.

It is not periodic but the distribution of numbers is regular. It means that by calculation the difference between values, we will get approximately the same amount of 1,2 and 0 so the mean value will be around 1. Finally, we can not forget the division by 2 therefore the final value for zero crossing rate of white noise will be around 0.5.

$$z_i = \frac{1}{2N}\sum_{n=0}^{N-1}|sgn(x_i[n]) - sgn(x_i[n-1])| =$$

$$= \frac{1}{2N}(m*|1-1| + m*|-1-(-1)| + m*|-1-1| + m*|-1-1|) =$$

$$= \frac{1}{2N}(2m*2) = \frac{2m}{N} = \frac{2*m}{4*m} = \frac{1}{2}$$

The check this result, we plotted the zero crossing rate of white sound and got the value around 0.45.

2.2 Spectral centroid

Spectral centroid is the barycenter of the spectrum. A higher value of SC corresponds to more energy of the signal being concentrated within higher frequencies. The computation of the i^{th} frame is following:

$$Z_i = \frac{\sum_{k=0}^{k=N-1} k * X_i(k)}{\sum_{k=0}^{k=N-1} X_i(k)}$$

Let Xi(k) be the DFT coefficients of the sequence.

Explanation for sinusoid

To better understand this formula, let's take a sinus wave of 3 Hz. It is clear that in the spectrum the highest amplitude (magnitude) will be for 3 Hz and for other frequencies the amplitude will be very low (around 0). Then, we will get following:

$$Z_i = \frac{\sum_{k=0}^{k=N-1} k * X_i(k)}{\sum_{k=0}^{k=N-1} X_i(k)} = \frac{0*0 + 1*0 + 2*0 + 3*magnitude...(N-1)*0}{0+0+0+magnitude+....+0} = \frac{3*magnitude}{magnitude} = 3Hz$$

As we can see we got the given frequency of the sinusoid as a result of the formula of spectral centroid.

Explanation for white noise

In spectrum of the white noise there are all frequencies equally covered so the spectrum is flat and magnitude is the same for each frequency. In this case, we will get following result:

$$Z_i = \frac{\sum_{k=0}^{k=N-1} k * X_i(k)}{\sum_{k=0}^{k=N-1} X_i(k)} = \frac{0*aml + 1*aml + 2*aml + 3*aml...(N-1)*aml}{aml*N} =$$

$$= \frac{aml*(0+1+2+..+(N-1))}{aml*N} = \frac{0+1+2+..+(N-1)}{N} = \frac{(N)*(0+(N-1))}{2N} = \frac{N-1}{2}$$

The final result is $N - 1/2$ where N is the number of coefficients of the signal.

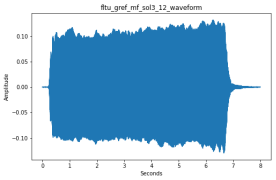
3 Task 2

Implement a function to obtain, for a given audio file, the mentioned set of instantaneous descriptors (1,2,6-9). To start, use similar analysis parameters: windowsize = 60 ms, hopsize=10 ms, no zero padding. Create plots to visualize the extracted instantaneous low-level descriptors and study their evolution for a small set of instrument samples (e.g. percussive, string, wind instrument). Play around with the STFT analysis parameters (windowsize, hopsize, etc.), and try to obtain the best compromise.

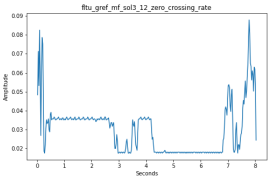
3.1 Solution

We have implemented a function *analyze_sound_instant_desc()* that can be found in code file that has one argument (audio file) and it returns set of instantaneous descriptors. We plotted the results for each sound given in InstrumentalSound's folder and we saved all the plots to a folder *plots* to be able to better distinguish each instrument.

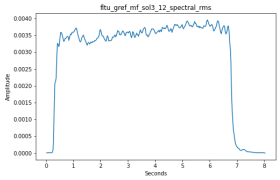
In Figure 1 there are plotted results for flute. In Figure 2 plots for violoncello and in Figure 3 results for percussive instrument.



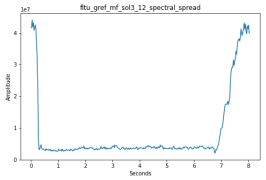
(a) Waveform



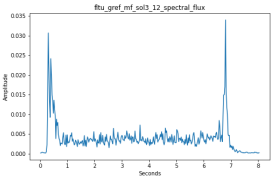
(b) Zero crossing rate



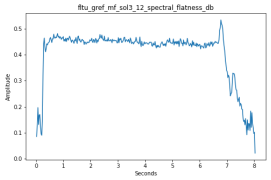
(c) Spectral RMS



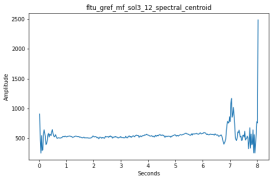
(d) Spectral spread



(e) Spectral flux



(f) Spectral flatness

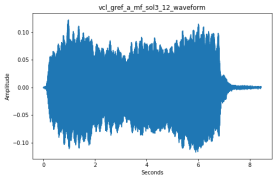


(g) Spectral centroid

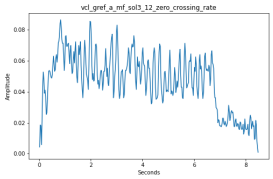
Figure 1: Descriptors for flute (woodwind instrument)

Zero crossing rate shows us the number of time the signal crosses the zero axes. It tends to have a small value for periodic sounds as we can see for the flute. Spectral centroid is quite stable across all non-silence parts, as it would be expected because this is a single note played on a flute with steady amplitude and steady frequency. Spectral centroid is also connected with spectral flux, flatness and spread that all use the spectral value. On the other hand, RMS shows the loudness of the signal so for example here we can see hat the loudness is stable and it decreases as the signal finishes.

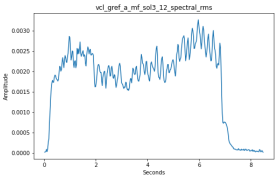
If we look at all the generated plots. We can notice some common difference for wind, string a percussive instruments. Zero crossing rate gives as in average higher values for percussive instruments than for string and wind instruments because they are louder and are not periodic. By looking at the string instruments descriptors we notices and unstable changing of spectral centroid as these instruments trembles and slightly changes their frequency over the time.



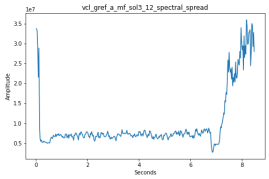
(a) Waveform



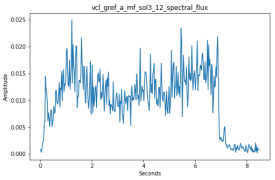
(b) Zero crossing rate



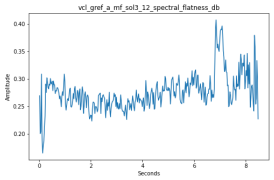
(c) Spectral RMS



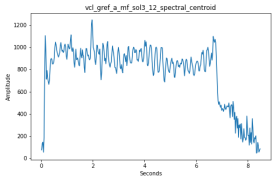
(d) Spectral spread



(e) Spectral flux

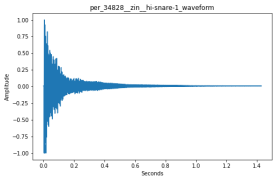


(f) Spectral flatness

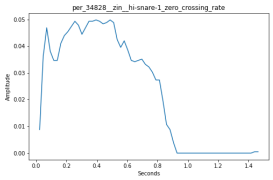


(g) Spectral centroid

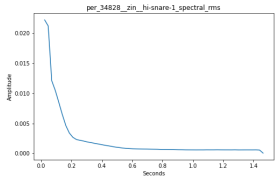
Figure 2: Descriptors for violoncello (string instrument)



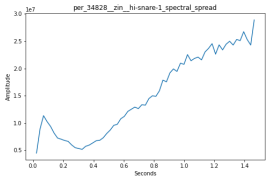
(a) Waveform



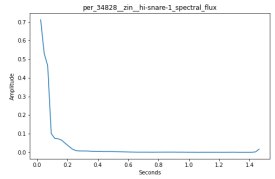
(b) Zero crossing rate



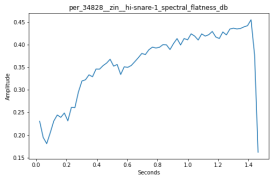
(c) Spectral RMS



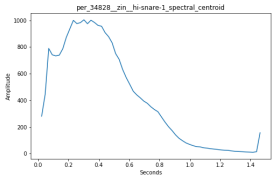
(d) Spectral spread



(e) Spectral flux



(f) Spectral flatness



(g) Spectral centroid

Figure 3: Descriptors for percussive instrument

4 Task 3

Implement a function to obtain, for a given audio file, the mentioned set of global descriptors (3,4,5), as well as statistics of the previous instantaneous descriptors (mean, standard deviation, min, max). Study the values of these descriptors for the previous instrumental samples and analyse how they represent the following aspects: percussive/non-percussive, sustained/non-sustained, low-pitch/high-pitch.

4.1 Solution

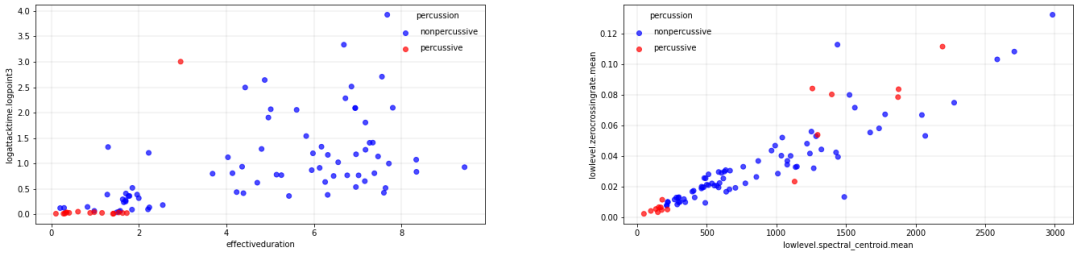
We implemented a function analyze_sound_global_desc() that analyzes all given audio files from a list and returns a dictionary of all global descriptors and low level statistics (this list is in *desired_stats*). This dictionary is saves as a *json* file to be further analysed.

We created a php script (called smul.php placed in /stats folder) that makes a csv file (excelfile.csv) from this json. The csv file contains addition information about the pitch of the instrument, sustainability, instrument group (wind, string, percussive, others) and kind of instrument (acc, flute, sax, etc.). This file with all statistics is then used for plotting the graphs that are shown in this section (function *scatterplot()*). To this report we just added some of them all the plots are located in folder *statistics/plot/*).

With all the plots and statistics we came up with following results.

4.2 Percussive

Zero crossing rate is a key feature to classify percussive sounds. Crossing rate for non percussive instruments tends to remain in low levels compared to other instruments [2] which we can confirm in Figure 4b. Also the spectral centroid tends to be low and it makes sense because the frequencies for percussive instruments are lower. In the Figure 4b we can also see that percussive sounds have lower effective duration and log attack time.

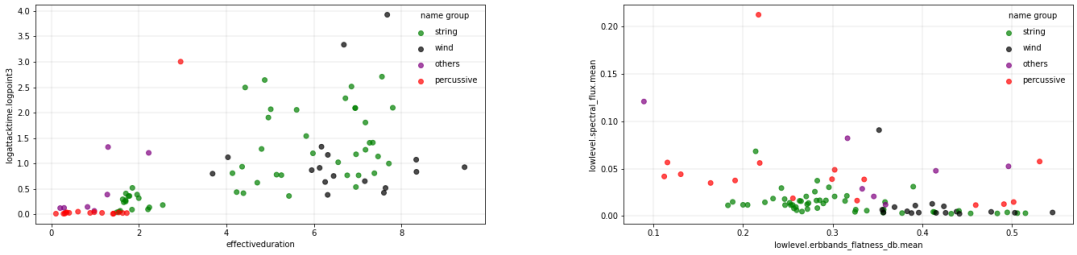


(a) Effective duration vs logattacktime (b) Spectral centroid vs zerocrossing rate

Figure 4: Classification between percussive / non percussive

4.3 Groups

In these example we put the instruments to 4 groups - string, wind, percussive and others. We found that the best descriptors for differentiating the groups were effective duration vs log attack time and Spectral flatness vs flux (because of similar behavior around spectral centroid of the sounds in the same group) as is shown in Figure 5b. Regarding effective duration vs log attack time as shown in Figure 5a we cam see that percussive sounds tend to have low effective duration as well as low stop attack time compared to other instruments.

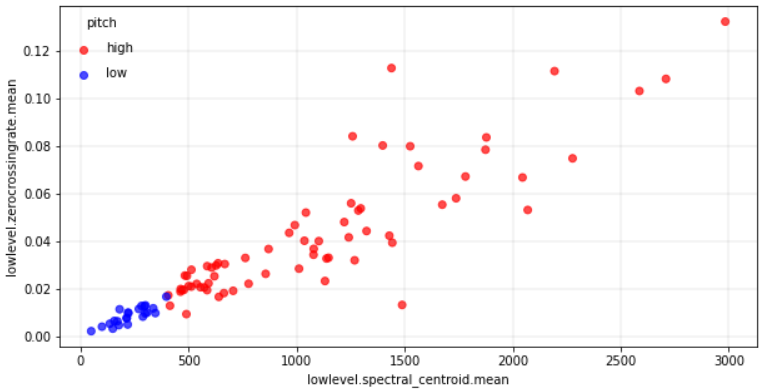


(a) Effective duration vs logattacktime (b) Spectral flatness vs flux

Figure 5: Classification between groups

4.4 Pitch

We used spectral centroid mean value to characterize pitch of each sound. We set a threshold of 400Hz. Values greater than that mean high pitch, while lower values indicate low pitch. The final results are in Figure 6a.

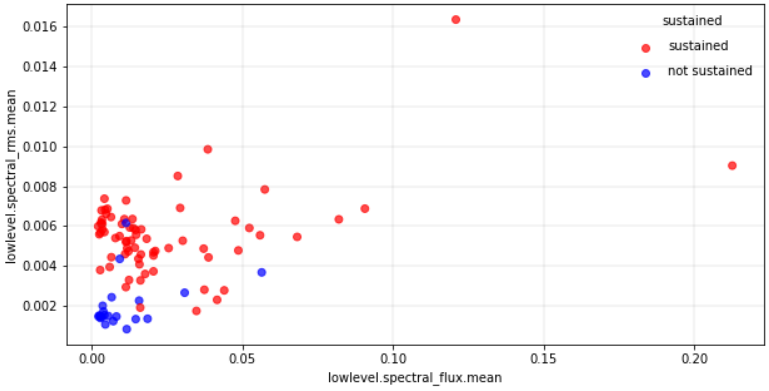


(a) Spectral centroid vs zerocrossing rate

Figure 6: Classification between high / low pitch

4.5 Sustained

We characterized a sound as sustainable if the effective duration was greater than 60% of the total duration of the sound. With this classification, we could better distinguish the groups in different pairs of descriptors. As we can see in Figure 7a non sustained sounds tend to have lower spectral RMS values as well as low spectral flux. On the other hand sustained sounds have higher spectral RMS mean. This is a reasonable conclusion since sustained sound means that it keep the RMS levels in high level in a long period of time.

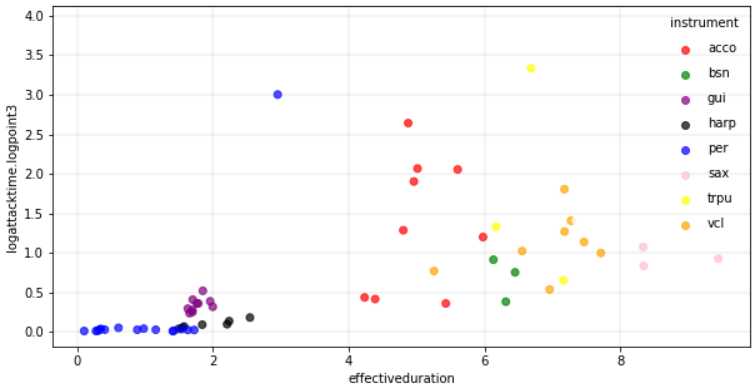


(a) Spectral RMS vs spectral centroid

Figure 7: Classification between sustained / not sustained

4.6 Instruments

Sounds with similar effective duration and same log attack time (stop attack value) help us to identify each instrument as shown in Figure 8a. Percussive instruments tend to have low effective duration as well as low attack stop time. Guitar's effective duration and attack stop time do not deviate much. Harp has slightly longer effective duration that percussive instruments but the same attack stop time. Accordion's attack stop time deviates a lot but effective duration remains the same, more than harp, less than trumpets. Bassoon has similar attack stop time and effective duration. The greatest deviation in attack stop time is presented in trumpets which tend to have slightly lower effective duration with saxophone. Saxophone has the greatest effective duration compared with other instruments. We can see that saxophone and trumpets are classified near and it makes sense since they are somewhat near in sound.



(a) Effective duration vs logattacktime

Figure 8: Classification between instruments

5 Task 4

Sound descriptors can be useful in many applications. The most important is the categorization of the sound using machine learning. Then, it can be used in instrument recognition that is about identifying the instruments involved or separating the music into one track per instrument. Last possible application would be in automatic music transcription that is able to convert the audio into symbolic notation (such as notes). [1]

6 Task 5

References

[1] *Music information retrieval [Online, update 1.4.2020]. URL* https://en.wikipedia.org/wiki/Music_information_retrieval

[2] *ON THE USE OF ZERO-CROSSING RATE FOR AN APPLICATION OF CLASSIFICATION OF PERCUSSIVE SOUNDS [Online]. URL* https://pdfs.semanticscholar.org/6509/14f8be2c96ab2f55faec54d3e3876c5b1b69.pdf