STUDY OF THE SIGNAL PROPERTIES OF MUSIC GENRES

A Degree Thesis Submitted to the Faculty of the

Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona

Universitat Politècnica de Catalunya by

Andreu Boadas Rabassedas

Carried out at the University of West Bohemia at the department of Electical Engineering.

In partial fulfilment of the requirements for the degree in

AUDIOVISUAL SYSTEMS ENGINEERING

Advisor: Martin Sýkora

Plzen, June 2018

Abstract

In this project different features of the music signals are studied and evaluated, for a better understanding of its behaviour from the technical point of view, to attribute similar patrons to the different genres studied, and also be able to differentiate them.

In order to differentiate one genre from each other and find some similarities between the audio tracks of the same genre, some specific features are extracted: statistical descriptors, dynamic and psycho-acoustical features, and frequency component features. With the information extracted from these features, a comparison with some aspects from the theoretical definitions of music genres is done to corroborate their specific characteristics. In the last part of the project, the rhythm features are extracted and classified by a support vector machine, to evaluate the obtained results.

The results obtained show that some music genres, like classical music or jazz, are easy to differentiate from the other genres. Moreover, there are other genres, like metal, disco or hip hop, that some of their features are mostly equal in all data set, being easy to establish common patterns on them.

Resum

En aquest projecte s'estudien i avaluen diferents característiques dels senyals de música, per tal d'entendre millor el seu comportament, i poder atribuir patrons similars als diferents gèneres musicals, a més de poder diferenciar-los entre ells.

A partir de l'extracció d'alguns descriptors estadístics, l'estudi de les característiques dinàmiques i psico-acústiques, i l'anàlisi freqüencial de la base de dades, s'intenten buscar certs patrons en el comportament de les característiques mencionades, que diferenciïn els 10 gèneres musicals. Per altra banda, s'intenten buscar similituds amb les definicions teòriques que se'ls hi atribueixen. A la part final del projecte, s'extreuen les característiques rítmiques, per finalment acabar classificant la base de dades amb màquines de suport vectorial, per així avaluar els resultats obtinguts.

Els resultats obtinguts mostren com certs gèneres musicals, com la música clàssica o el jazz, es diferencien en molts aspectes de la resta de gèneres musicals i de manera molt clara. A més, hi ha altres gèneres, com és el cas del metall o la musica disco, els quals algunes de les seves característiques tenen un comportament pràcticament igual en tota la base de dades i fa que sigui senzill establir un patró entre ells.

Resumen

En este proyecto se estudian y evalúan diferentes características de las señales musicales, por entender mejor su comportamiento des de un punto de vista técnico, y poder atribuir patrones similares a los diferentes géneros musicales, además de poder diferenciarlos entre ellos.

A partir de la extracción de algunos descriptores estadísticos, el estudio de características dinámicas y psicoacústicas, y el análisis en frecuencia de la base de datos, se intentan encontrar ciertos patrones en su comportamiento, que diferencien los 10 géneros musicales. Por otro lado, se intentan buscar similitudes con las definiciones teóricas que se les atribuyen. En la parte final del proyecto, se extraen las características rítmicas, para finalmente terminar clasificando la base de datos con máquinas de soporte vectorial, y poder así evaluar los resultados obtenidos.

Los resultados obtenidos muestran cómo ciertos géneros musicales, cómo la música clásica o el jazz, se diferencian en muchos aspectos del resto de géneros, y de manera muy clara. Además, hay otros géneros, cómo el metal o la música disco, los cuales algunas de sus características tienen un comportamiento prácticamente igual en toda la base de datos y hace que sea fácil establecer un patrón entre ellos.

Acknowledgements

First of all, I would like to thank on one hand my project supervisor, Marin Sýkora, for his help in the analysis of the results. I would also like to thank my home advisor, Antonio Bonafonte, his help at the beginning of the project, and the goodwill during all the semester. Finally, I would like to thank Harry Sagel his help in the grammar and vocabulary corrections in some parts of the project.

Apart, I would finally like to thank the authors of MIRToolbox software to make easier my project and for its potential toolbox.

Revision history and approval record

Revision	Date	Purpose
0	25/04/2018	Document creation
1	09/05/2018	Document revision
2	23/05/2018	Document revision
3	04/06/2018	Document revision
4	15/06/2018	Document approval

DOCUMENT DISTRIBUTION LIST

Name	e-mail
Andreu Boadas Rabassedas	andreu.boadas@gmail.com
Martin Sýkora	msykora@ket.zcu.cz
Antonio Bonafonte Cávez	antonio.bonafonte@upc.edu

Written by:		Reviewed and approved by:	
Date	25/03/2018	Date	15/06/2018
Name	Andreu Boadas Rabassedas	Name	Martin Sýkora
Position	Project Author	Position	Project Supervisor

Table of contents

List of Figures

List of Tables:

1. Introduction

In this chapter the objectives, requirements and methods are presented in order to introduce the main goals of the current project.

1.1. Statement of purpose

Nowadays, music is a fundamental part of our lives. There exist so many types of music, from all parts of the world and many different styles and genres. In the last few years, music production and artists have increased substantially, and consequently so many new genres and subgenres have appeared. Everybody perceives the music in a different way, and there exists kinds of music for every personal taste. But, what are the features which characterize every genre, and makes different one genre from each other?

We have to take into account that every time is more difficult to classify songs and establish a label of genre for them, and for that reason we have to be very careful when we label a song, because not always is as trivial as it seems. Music genres are considerate classes which contain songs that share a similar rhythm or tempo, usually played with the same instruments, and sometimes although share melodic patterns. But in a few years have appeared a lot of subgenres that are very difficult to differentiate each other. For this reason, this project is focused on the 'classic established genres' as Rock, Pop, Jazz, Classical Music, etc.

The aim of this project is the study of the different dynamic and psychoacoustic features extraction methods in music signals and the comparison of 10 music genres from the technical point of view, taking as a baseline the theoretical definitions of them. A frequency component analysis is also done to evaluate genres in the same direction. Finally, rhythm pattern features are extracted to check the influence on the rhythm in music genres, and a support vector machine is implemented to evaluate these results.

1.2. Requirements and specifications

The main project requirements are to implement a feature extraction program which allows to analyse the music signals, focused on the differences between genres. After that, evaluate the results comparing them to the theoretical definitions, and implement a classifier to validate the results of some of the extracted features.

The project specifications are:

- A perceptive analysis of music genres and its theoretical definitions must be studied.
- Psychoacoustic and features about dynamics, on one hand, and frequency component features, must be extracted from the data-set.

- A rhythm features extraction must be implemented, and evaluated by a support vector machine.
- To use the GTZAN Genre Collection Dataset in order to analyse the different features.
- To use Matlab as the main programming language.

1.3. Methods and procedures

The project is carried out at the Faculty of Electrical Engineering at the University of West Bohemia in Pilsen, Czech Republic, during the Erasmus mobility program, in agreement with the home university UPC.

The dataset used to extract features, test and classify is the "GTZAN Genre Collection" database. It contains 10 music genres, each one represented on 100 audio tracks of 30 seconds each one. This 1000 audio-tracks database was created in 2002 for the specific 'genre classification' task, and has been used many times for other researchers. The dataset will be split in train (70%) and test (30%) subsets to evaluate and classify, at the last step of the project, the rhythm pattern features.

This project has been developed using MATLAB software (R2017a version) and also some functions of the MIRToolbox (1.7 version), a MATLAB free toolbox [25] used to facilitate some work.

The main purpose of this project is, as explained on section 1.1., a better understanding of the music behaviour, and the identification of some common patterns. The project is divided in 3 main parts. The first one is the analysis of some low level features, as some statistical descriptors (maximum, minimum and mean), crest factor and dynamic range, and other psychoacoustic features. The second part is the frequency component analysis of the audio tracks, in particular based on the bark scale spectrograms. These two parts are objectively analysed, comparing the results to the theoretical definitions of music genres, and trying to relate how humans perceive music genres with the technical definition. The third and last feature extraction part, is the rhythm patterns feature extraction. Unlike the first two parts, the only way to check the results in this part is by developing a feature classifier. So in that direction, a support vector machine is developed to train the data-set, classify the test data, and evaluate how the rhythm patterns feature extraction works. Finally, some low level features are added to the SVM with the intention of improving the classification.

1.4. Work plan

The work plan consists in 7 work packages announced in the section below, and a more detailed explanation can be found in appendix A.

1.4.1. Work Packages

WP1: Project Proposal and Work Plan

WP2: Information research and documentation

WP3: Software development

WP4: Critical review

WP5: Test and results

WP6: Final report

WP7: TFG presentation

1.4.2. Gantt Diagram

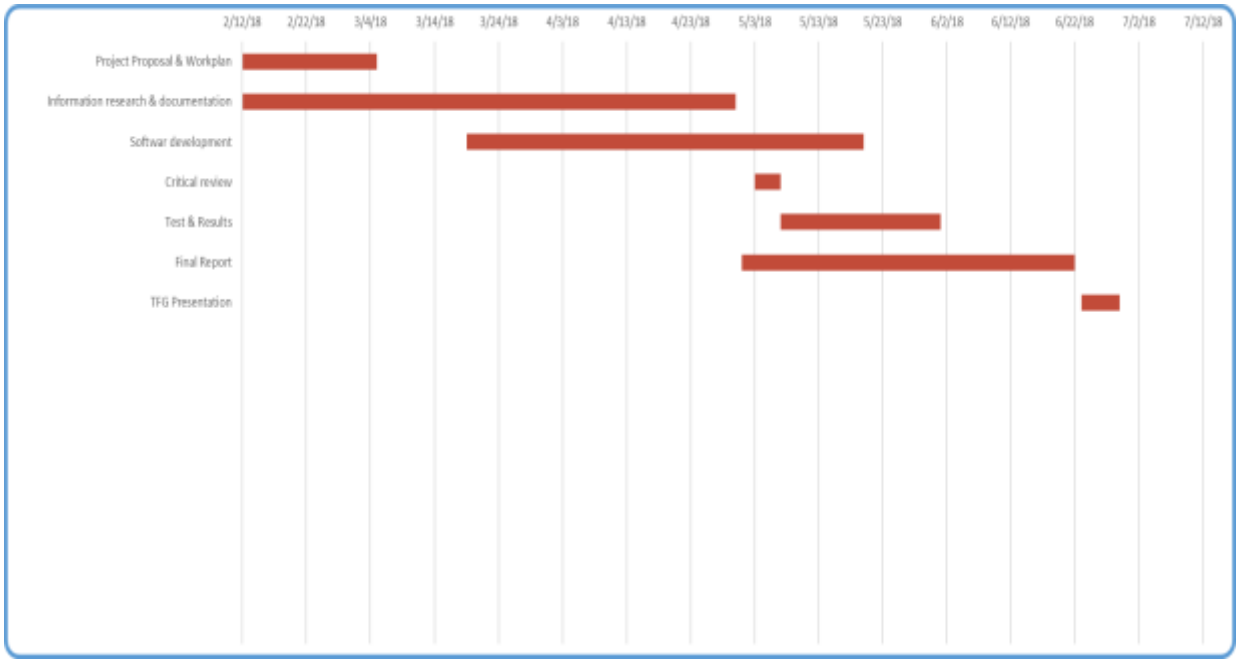

Figure 1: Gantt diagram of the work plan

1.5. Incidents and modifications

There have been many incidents during the project that have modified some parts of the work packages and the duration of them.

The work package that has suffered the main changes has been the software development (WP3). As at the beginning the main objectives weren't exactly specified, the procedure was a kind of trial-error based development process: from the study of the

state of the art, some possible features to extract were developed, and after the analyse of the results was checked if there was some useful information to take into account and continue on this way, or they had to be discarded and look for another starting point. This situation took to extend the WP3, and finish it more or less at the same time that the results (WP5).

Apart, at the beginning, the final report (WP6) was supposed to be delivered in the beginning of June, and finally it was delivered at the end of June, allowing to have 3 more weeks to extend WP3 and WP5.

2. State of the art of the technology used or applied in this thesis:

The study of the music analysis from the technical point of view and the automatic recognition of music has been since 30 years ago an important researching field of many engineers and scientists. Since the categorical classification of musical genres is one of the most important ways that people uses to search and discover new kind of music, this specific field has become one of the main purposes for researchers.

Music Information Retrieval (MIR) is the interdisciplinary science that studies and has developed many tasks focused on the retrieval of music information. This research field involves other study fields as musicology, psychoacoustics or signal processing and machine learning among other ones. Its main objective is the deep study of the music on the engineering point of view to improve the actual knowledge, and develop many useful applications, as music recognition and classification, music transcription or generation, etc. [43][44]

To achieve a good initial knowledge of all this field, it's important a deep research among all the wide bibliography and the previous researches about music feature extraction, focused on genre recognition methods.

2.1. Music Genre Recognition Overview

In this chapter, a brief overview of previous works is explained, all focused on music genre classification. But it has only been a starting point for the knowledge of the current state of the art, because as explained on the introduction, the main objective of this research was understanding the behaviour of genres in some specific features, instead of achieving a good classification, which is the main objective of the most of the previous works.

Music users are able to recognize and differentiate genres easily. In [4] Perrot and Gjerdigen demonstrate that with only 250ms of a song, people can recognize genres correctly in almost 50% of the songs, and with 3000ms the accuracy achieves the 70%. But when talking about automatic recognition, it becomes a more difficult task.

At the beginning of 90's, MIR started carrying about the treatment of audio signals and music psychoacoustic feature extraction. In [2] Tzanezakis and Cook introduced this study field as a pattern recognition task, focused on timbre texture and beat-related (explained in section 2.3) features. Moreover, they built GTZAN dataset, that has been widely used in similar works, and it's also used in the current project. Since then, many works appeared and the MIR research community started to present new studies and new upgrades evaluating different kinds of music features. The growth of the digital music

available on the internet was one of the motivations to develop new works related to the automatic recognition of music, and some of them are briefly commented at the following paragraph.

The study of the rhythm and tempo of the songs has been one of the main ways to evaluate music genres, because it's one of the most differential characteristics between genres. Apart, in [9] Logan is the first one to introduce Mel Frequency Cepstrum Coefficients (MFCC) into the music characterization. Until that moment, MFCC were only used on speech recognition, where it was proven that they were useful. But the mentioned study proved the importance of using MFCC also for music recognition, and in 2004, Pampalk's system [12] won the MIREX competition (one of the most recognized MIR organizations) using these features. From that moment on, MFCC have been commonly used for this kind of researches, following rhythm features. Moreover, during the last years the main features extracted for the genre recognition task have focused on the spectrogram analysis, threating them as images and analysing them with some image processing tools, like local binary patterns, and combining them with bag of features techniques for a better classification accuracy [3][6][8][45][1].

2.2. Low-level Audio Features

As the objective of the works commented in section 2.1 was the music genre classification, this state of the art research differed a little bit from the initial purpose of the project. In order to achieve features which could be useful for the objective comparison of music genres, other features were studied and taken into account. In this chapter the features studied and used for the future project development are exposed.

2.2.1. Statistical descriptors

The statistics of the spectral distribution over time can be used in order to represent the "musical surface" for pattern recognition purposes. The knowledge of the frequency distribution during a piece of audio allow us to understand the energy distribution of the signal and its behaviour. For that reason, the statistical descriptors based on the spectrograms are commonly used to extract useful features to characterize audio tracks.

2.2.2. Bark Scale Spectrograms

Spectrograms are a very useful tool to represent the spectrum of frequencies as they vary with time [20]. To create this visual representation of the audio signal, it's needed to compute the squared magnitude of the short-time Fourier transform.

[1] This project wasn't focused on this methods, and for that reason the review of the state of the art in that direction is exposed briefly.

The bark scale was first introduced in 1961 by Zwicker [13]. It is a psychoacoustic scale, what means that tries to reflect the perception of human hearing, rescaling the linear frequencies to bark scale frequencies. The human auditory system has a logarithmic response, and the bark scale tries to simulate the critical bands created by the cochlea. From 0 to 500 Hz this scale has a linear bandwidth (100 Hz) and up to 500 Hz the bandwidth increase in a logarithmic way.

The bark scale ranges from 1 to 24 correspond to the first 24 critical bands of hearing [21]:

Bark	F_1	F_2	Bark	F_1	F_2	Bark	F_1	F_2
1	0	100	**9**	920	1080	**17**	3150	3700
2	100	200	**10**	1080	1270	**18**	3700	4400
3	200	300	**11**	1270	1480	**19**	4400	5300
4	300	400	**12**	1480	1720	**20**	5300	6400
5	400	510	**13**	1720	2000	**21**	6400	7700
6	510	630	**14**	2000	2320	**22**	7700	9500
7	630	770	**15**	2320	2700	**23**	9500	12000
8	770	920	**16**	2700	3150	**24**	12000	15500

Table 1: Bark scale band frequency ranges

Bark scale spectrogram are widely used in audio signal processing. The use of the Bark scale instead of the linear frequency scale is used to emphasize the most important frequencies of human hearing and be able to extract more similar features related to how we perceive the music.

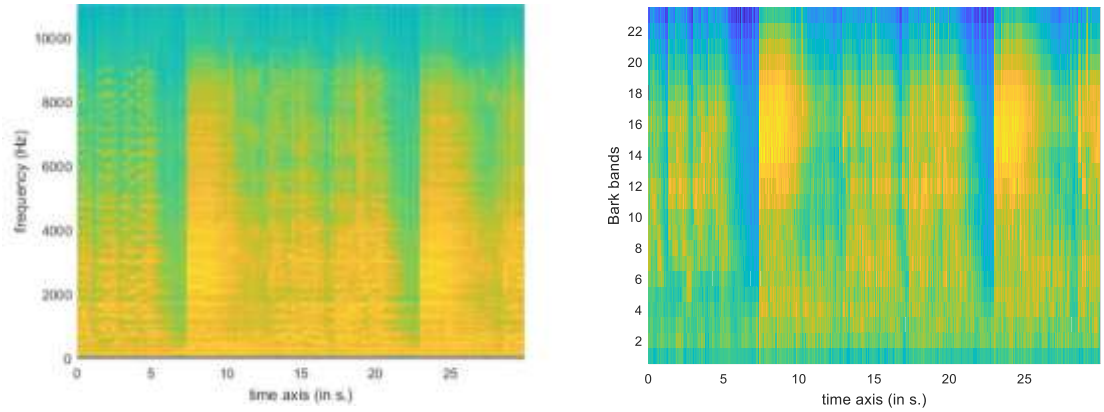

Figure 1: Normal Spectrogram (left) vs. Bark Scale Spectrogram (right)

2.2.3. Dyamics characteristics

The study of the music dynamics allows to understand better the loudness behaviour of the audio tracks, and the evolution of it during the song can give some important information about sound intensity and musical characteristics. To evaluate the dynamics, it has to be taken into account some definitions:

Root Mean Square is the average loudness of the signal, the midpoint between the maximum and the minimum of the signal. It is defined as the square root of the mean square [22].

$$RMS = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}}$$

The Crest Factor indicates how extreme the peaks are in the waveform. It is usually between 10dB (noisy environment) and 20dB. It's defined as the ratio between the strongest peak level and the average loudness (RMS) [23].

$$CF = 20 * log \frac{|x_{peak}|}{RMS}$$

Finally, the dynamic range is used to determine the ratio of the amplitude of the loudest signal to the noise floor. It's defined as the ratio between the largest and smallest values than a certain quantity can assume. With the analysis of these characteristics in short segments of the audio tracks, it's possible to get an idea of the loudness behaviour and get an idea of the main volume changes during the time.

2.2.4. Psycho-acoustical Descriptors

Psycho-acoustical descriptors have been widely used on music genre recognition, to know information about the spectral shape of the songs. In most of the works they are used as supporting features, following the main features as rhythm patterns or MFCC.

2.2.4.1. Centroid and Roll-off

The centroid is the balancing point of the spectrum. In other words, indicates the "center of mass" of the spectrum (the frequency where the energy below that frequency is equal to the energy above). It's a measure of the shape of the spectrum and it's associated with the spectral brightness [1][2]. It can be calculated as:

$$C = \frac{\sum_{n=1}^{N} Mt[n]*n}{\sum_{n=1}^{N} Mt[n]}$$

Where M[f] is the magnitude of the DFT at frame t and frequency bin f.

In the same direction, the spectrum roll-off is also a measure of the spectral shape. It is a generalization of the spectral centroid (the centroid is the roll off for r=50%), defined as the frequency R corresponding to the r% of the magnitude distribution. Usually is used a value of r=80-90%.

$$\sum_{n=1}^{R} Mt[n] = r * \sum_{n=1}^{N} Mt[n]$$

2.2.4.2. Low Energy

The low energy feature is computed around a certain number of windows through the signal. It is defined as the percentage of windows that have less energy than the "average" energy of all windows [1]. This feature is useful to detect music with large periods of silence (the more the track has silent parts, the higher will be the lowest energy values).

2.3. Rhythm Features

The extraction of rhythm features to characterize the music signal has been a common study field in MIR. As it's known, one of the most common ways to recognize a genre of a song is with its rhythm or tempo. As understanding rhythm as the "speed" of the song, transformed to beats per minute (BPM), it's a good feature for differentiate one genre from another one. There has been many proposals to extract features of the rhythm of the music, but the ones which this state of the art revision focus on are the ones proposed on [2] and [11].

In [2], Tzanezakis proposed the extraction of a rhythm feature histogram based on the wavelet transform. To develop this method, first of all a preprocessing is applied to divide the signal into 6 seconds temporal windows. In each window, some individual steps described below are applied, and then the mean of all the 6-seconds windows is computed, to achieve the final result.

The first step is applying the DWT to decompose the signal in 10 octave bands. Wavelets are developed to solve resolution problems of the STFT, as it gives a high temporal resolution and a low frequency resolution at high frequencies, and a low temporal resolution and a high frequency resolution at low frequencies.

Then, is extracted the envelope on each band: a perceptually important feature of musical instruments sounds and speech, correlated to the percussiveness of musical instruments sound [5]. To achieve it, the steps to be followed are:

- Full Wave Rectification
- Low Pass Filter
- Downsample
- Mean removal

After the envelope extraction, the final step is to sum all the envelopes of each band and compute the autocorrelation. Dominant peaks on the autocorrelation function correspond to the lags where the signal has the strongest self-similarity. Periodicities of the fifth strongest peaks are calculated and added in a "beat histogram".

After compute the mean of all 6-second window values, from the final "beat histogram", periodicities and amplitudes of the peaks are extracted as rhythm features. Looking at the weight of its periodicities, it's possible to get an idea of the "strongness" of the beat of the music. This method was one of the first to be developed, and it's useful to discriminate "rock music" to complex world music, or distinguish classical music because of the not accentuated beat.

Another method to extract rhythm patterns and its histogram was proposed on MIREX 2005 by Lidy and Rauber [11]. As the method previously explained, before the feature

extraction, the audio track is segmented into pieces of 6 seconds, and in this case, the first and the last segments are skipped because of the usual rhythm difference between the first or the last 6 seconds in comparison of one segment from the middle of the song. After the segmentation, the steps to get the rhythm patterns are the following:

- Compute the Short Time Fourier Transform (STFT) using a 23 ms hamming window with 50% of overlap.
- Sum up the frequency bands to the Bark scale bands (24 critical bands according to the human hearing system).
- Successively, the data is transformed to the logarithmic scale, to the Phon scale according to Zwicker and Fastl [13] equal-loudness curves, and afterwards into the unit Sone, to reflect the specific loudness sensation. At this point, some statistical descriptors are computed: maximum, minimum, mean, median, variance, skewness and kurtosis.

At this point, each bark scale band is seen as an amplitude modulation through the time. A Fourier transform is applied to the bark scale transformation to achieve a time-invariant representation of the signal. The result is a representation of the magnitude of modulation per modulation frequency for each critical band. The algorithm only takes frequencies up to 10Hz, due to the sensation of roughness starts at 15Hz and the notion of rhythm ends. Moreover, according to the human auditory system, values around 4Hz are accentuated.

Finally, to create the “rhythm histogram”, 60 bins are extracted from the modulation frequencies (from 0 to 10Hz at a resolution of 0.17Hz) of each band, and summed up to create the segment histogram. At the end, for a given audio track, the “rhythm histogram” will be the mean of all the 6 second segment histograms.

2.4. Classification models

To evaluate how the features explained above work, a classification model based on machine learning is needed. There exists an amount of classification models, but the most widely used for MIR researchers are the support vector machines. The following explanation is based on [6].

SVM are supervised machine learning models introduced by Vapnik et al. [46], and are the system of choice in a large number of applications due to their robust performance with respect to sparse and noisy data.

Given a vector space which accomplishes $\Phi: S \rightarrow R^n$, and a set of positive and negative examples mapped in vectors, SVMs classify them according to a separating hyperplane $H(\vec{x}) = \vec{w} * \vec{x} + b = 0$, where $\vec{x} = \Phi(s)$, and $\vec{w} \in R^n$ and $b \in R$, learned by applying the Structural Risk Minimization principle [46].

According to the kernel theory, the maximal margin hyperplane is $w = \sum_{i=1}^{m} \alpha_i y_i \overrightarrow{x_i}$, where y_i is equal to 1 for a positive example and -1 for a negative example, and $\alpha_i \geq 0$. $\Phi(s_i) = \overrightarrow{x_i}$ $\forall_i \in \{1, \ldots, l\}$ are the training instances, and the product $K(s_i, s) = \langle \Phi(s_i) \cdot \Phi(s) \rangle$ is the kernel function. For example, the linear kernel uses the scalar product as a kernel function $K(s_i, s) = \overrightarrow{x_i} * \vec{x}$. Another commonly used kernel is the polynomial one $K(s_i, s) = (c + \overrightarrow{x_i} * \vec{x})^d$, where c is a constant and d the degree of the polynomial. Kernel functions are particularly important when data are not linearly separable, and they give powerful tools to learn class differences.

2.5. Music genre definitions

In order to understand better the music genres which are studied– blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae and rock – and have a theoretical definition of its behaviour to be able to compare the features analysed, a research of how the music genres are seen for the music experts has been done. But despite exist a general definition of each of the music genres, it has to be said that nowadays it's very difficult to define by words which type of music are we listening, and sometimes these definitions cannot be as accurate as we would like to be.

There are wide and very specific definitions for all the genres, but this review it's only focused on the main characteristics and differences to be able in the next chapters to correlate this information with the results extracted. The definitions have been summed up taking references from [41] [48], besides of the specific bibliography for each genre.

2.5.1. Blues

Blues music was originated by the Afro-American communities in the USA during the end of the 19th century. It is a cyclic musical form, what means that it's based on a repeating progression of chords (called "call and response" scheme) [29]. It does not use to have changes in its rhythm structure, and the main instruments are always drums, guitar, bass and voice, what makes it to have a flat loudness during all the song. Guitar solos are also common in blues music.

2.5.2. Classical

Classical music is one of the most ancient western music that exist. When talking about classical music, we have to take into account that there are many different types of it (ancient music, medieval, renaissance, baroque, romantic, modernist, etc.) because of the anthropological evolution of the society across the centuries, and there are also different types of aggrupation – from duos, trios or chamber music, to symphonic

orchestras. For this reason, it's very difficult to define exactly the classical music behaviour.

One of the main differences between classical and the other genres is the kind of instruments which it's played, as it's almost the only one that does not need to be electrically amplified and does not have a drum baseline. Moreover, because of the amount of instruments "solos" and different passages across a classical piece, the loudness changes during the whole piece are very significant, specially on the symphonic orchestras.

2.5.3. Country

Country music appeared at the beginning of the 20th century in the USA and Canada, mixing folkloric Irish music with Afro-American music as blues or gospel [31]. As it is a blues music ramification, it also has a base constant rhythm, but it uses to have a bigger amount of different instruments apart of the baseline ones, as violins, harmonic, accordion, keyboard, etc. and this makes it, according to Table 2, a genre with higher frequencies and a wider main frequency range.

2.5.4. Disco

Disco music appeared in the earlies 1970's in the America's urban nightlife. Disco sound often has several components: it's common to listen hi-hat patterns, syncopated electric bass line, and it use to have rhythm changes during its songs [32]. There are also several instruments playing at the same time, creating a rich background sound, and also with high frequency responses, as electric guitars, horns and keyboards, making this music genre concentrate the main frequencies on the high range.

2.5.5. Hip-hop

As disco music, hip-hop appeared in the 70's in the USA in contrast of the most popular music genre of the scene: rock music. It consists of a stylized rhythmic music, commonly followed by rap music. It usually has strong constant beats with sticking and repetitive melodies, sometimes extracted from other kinds of music, and mashed up with a hip-hop base [33]. This strong baseline makes it a music genre with low frequency based content sound.

2.5.6. Jazz

Jazz music appeared at the end of the 19th century, at the same time and with the same roots as blues music. It's one of the most difficult genres to define, due to the

complexity and variability of its rhythms and melodies. As it happens with classical music, jazz is a “music genre family” with a lot of variants and different kinds of sub-genres, and this makes it very difficult to stablish a common pattern between songs [34].

Jazz songs use to have a baseline with drums and bass guitar, and it rarely changes its tempo during the whole song. The main structure is always the same: there is a main melody, and it is combined during the song with “solos”, usually from keyboards, guitars or metal wind instruments (as explained in blues music, this behaviour is usually called “call and response” scheme). So it’s loudness uses to change significantly due to the difference when whole the band is playing, or only one instrument is performing a “solo”.

2.5.7. Metal

Metal music is a kind of rock music that developed in the 60’s, largely in the United Kingdom. It is characterized by highly amplified distortion, extended guitar solos and strong and fast and constant drum beats, so the energy component in whole songs is huge [35]. Moreover, because of the importance of the electric guitar in its pieces, it use to have high frequency components, acording to the frequency range of this instrument (see Table 2).

2.5.8. Pop

Pop music was originated in the USA and UK during the middle of the 20th century. During the years the meaning of pop music has been constantly changing. Pop music comes from the abbreviation of “popular music”, so at the beginning pop music was all the music that “was most in line with the tastes and interests of the urban middle class” (The New Grove Dictionary Of Music and Musicians).

Nowadays, it could be defined as a kind of music with a consistent and noticeable rhythmic element, a mainstream style and a simple and traditional structure, with catchy melodies and chorus [36].

In pop songs (and it also happen on rock songs) there exist what’s called the loudness war, which refers on the trend of increasing audio levels digitally. This started to happen with the introduction of digital signal processing, capable to increase the loudness through dynamic range compression, and it happened mostly in this two music genres because of its popularity and also because of being the most common music in radio broadcasting **¡Error! No se encuentra el origen de la referencia.**.

2.5.9. Reggae

Reggae is a music genre originated in Jamaica in the middle of the 20th century. It is a mash up of rhythm and blues, calypso, jazz and traditional African and Latin music. One

of the most recognizable elements are the offbeat rhythms, usually played by the guitar or keyboards, creating a syncopated rhythm. Reggae baseline uses to be simple (drums, guitar, bass and keyboards) sometimes followed by wind instruments as saxophones or trombones [38].

2.5.10. Rock

Rock music appeared in the 50's in the USA, and it's one of the most popular music genre in the world. Rock music is traditionally centred on the amplified electric guitar, so the range of main frequencies is usually high. The guitar is always followed by strong rhythmic drums and a constant bass guitar, creating a high energy baseline.

As it's one of the most popular genre, it's variants are also very wide, and it's difficult to establish a common song pattern. Also the main rhythm through one song is usually difficult to know, because there use to exist many different stages and phases during a rock piece, and sometimes it's difficult to identify as a rock song to the similarities on other genres, as blues, jazz or pop.

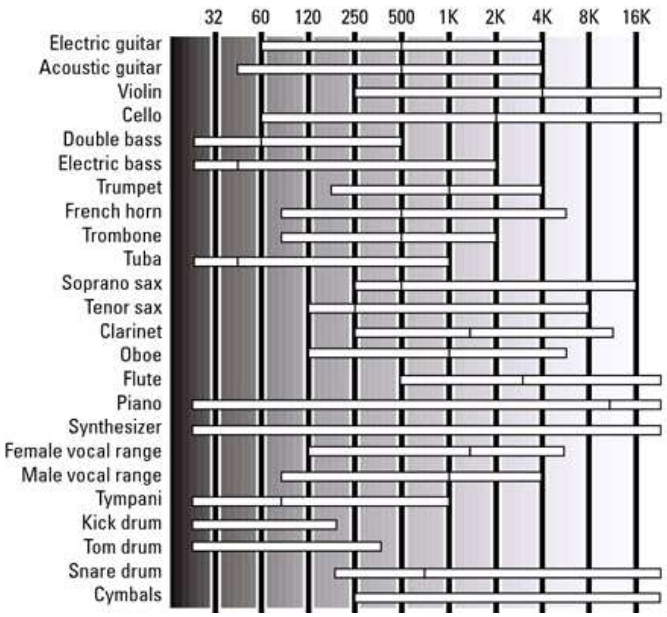

Table 2: Frequency range of musical instruments [42]

3. Project development:

3.1. Data-set

To evaluate the signal properties of the music signal, the GTZAN Genre Collection was used, developed by G. Tzanetakis and P. Cook for the work "Musical genre classification of audio signals" [2], one of the first works of music classification. After that, this dataset has been widely used in a lot of studies related. In [18] a review of the most common datasets is exposed. Based on it and taking into account the project needs, it was finally decided to use the GTZAN data set, due to its free online availability, the clear separation between genres, and its common use in similar tasks. One of the main problems of this dataset exposed in [18] is its small size and that it has some repeated files, but due to the project objectives, these weren't huge problems in comparison of the advantages.

The GTZAN Genre Collection dataset consists of 1000 audio files in 'wav' format each 30 seconds long, 22050 Hz sample rate Mono and 16-bit resolution [19]. They are divided into 10 genres (blues, classical, country, disco, hip hop, jazz, metal, pop, reggae and rock), and each genre is represented by 100 tracks.

Once the data-set was obtained, the first step was developing a software to split this data automatically in train and test, to classify the data-set and be able to evaluate the results. The data-set was separated in 70% of training data (70 tracks each genre, 700 in total) and 30% of testing data.

3.2. Software Toolbox

In order to extract some features and graphs, it was used a free Matlab software developed in 2007 by O. Lartillot and P. Toiviainen [25]. This software, "MIRToolbox", is available to use for academic researches under the terms of GNU General Public License [26]. "MIRToolbox" allow to use functions to extract timbre, tonality or form features, as well as functions for statistical analysis. It is a very powerful tool, but in this work it was shortly used in some specific cases.

3.3. Methodology and Project Development

3.3.1. Low Level Features

The first step to understand better music genres behaviour, was evaluate the statistical and psycho-acoustical descriptors in two different ways to look if there exists considerable differences and try to distinguish some common patterns.

In order to have baseline results to compare with, the features on the whole 30 seconds audio tracks were evaluated. These low level features were some statistical descriptors (maximum, minimum, median), and as explained on section 2.2, on one hand dynamic descriptors (Crest Factor, Dynamic Range and the average power of each song), and on the other the psycho-acoustical descriptors: Centroid, Roll off 85% and Low Energy.

To contrast results and view the behaviour of the signal along the time, each audio track was divided into 30 segments of 1 second (taking into account that the duration of the data-set audio tracks is 30 seconds each one, as explained in section 3.1), and the statistical, dynamics and psycho-acoustical descriptors were evaluated in each segment. Once these features were extracted for each segment, the mean and the standard deviation were computed across the segments to compare with the baseline values (Figure 2).

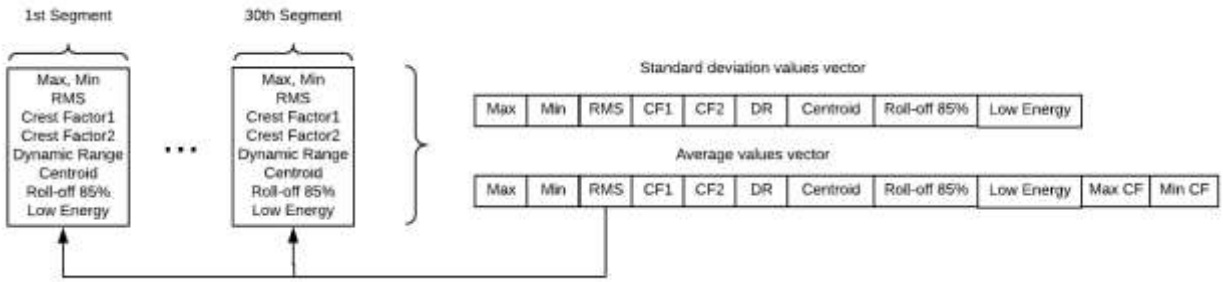

Figure 2: Low level features extraction scheme

This process was completed across the 700 audio tracks of the training data-set, and the result was kept in two matrixes for future calculations[2]. From the "standard deviation values matrix", the average for each genre was computed to check if there is any genre which its features during the 30 seconds period have a small data range (small standard deviation value), or if they differ a lot from its mean value. From the "average values matrix" it was computed the average and the standard deviation also for each genre, to prove if there is a common behaviour in a certain genre (average), and moreover check if all the features from a specific music genre are similar between them or they have huge different values (standard deviation). This explanation is supported with a scheme in Appendix B.1 for a better understanding of the process.

Apart, to compare the loudness behaviour of the audio tracks, the crest factor was specifically taken into account. In each segment, the crest factor was computed in two different ways, as shown in Figure 2: the first one as the result of dividing the maximum segment value with the segment RMS (Crest Factor 1), and the second one replacing the

[2] The matrixes containing the values are not added in this project due to its wide dimension and its lack of visible information.

segment RMS by the mean RMS obtained from whole 30 segments (Crest Factor 2). The aim of this comparison was checking out if there exist any significant differences of these two values in the pieces of each music genre. If the difference between these two values in some part of the audio track is huge, this means that this music piece has significant loudness changes. On the other hand, the similarity between these two values denotes that whole audio track has a constant and flat volume. Computing the standard deviation in each segment between both crest factor values was the way to test the mentioned differences in audio tracks.

3.3.2. Bark Scale Spectrogram Features

As it was explained in section 2.2.1., the bark scale is a psychoacoustic scale which allows us to represent the spectral energy in a frequency distribution of 24 scales according to the human hearing behaviour. In this project only 23 bark scale bands are used, because the first one (from 0 to 100 Hz) is dismissed. From the bark scale spectrogram, a “bark scale vector” was created containing the average energy value for each bark scale band. Following the “bark scale vector” methodology across the 700 audio tracks, the result was a “bark scale matrix” containing the mean energy on each bark scale band for every audio track of the training data-set.

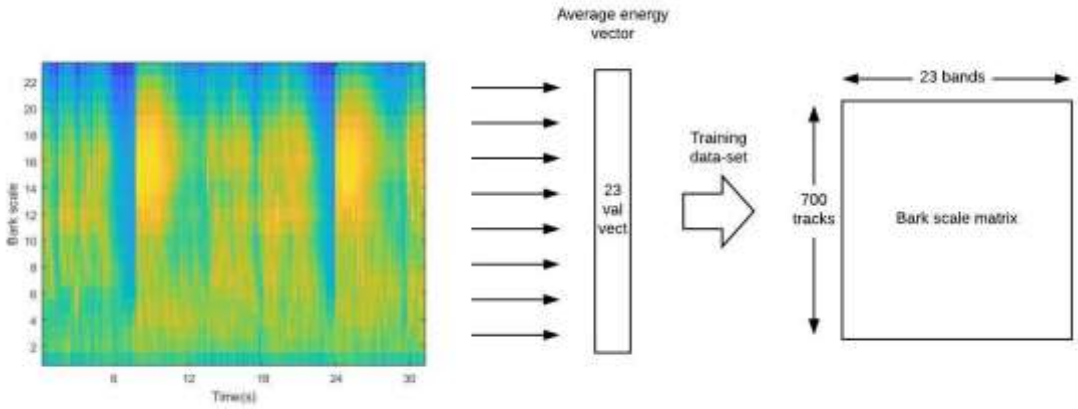

Figure 3: Bark scale feature extraction scheme

From this “bark scale matrix”, a certain number of features were extracted in attempting to find any similar behaviour in music genres on the frequency domain. The first features extracted were the 1st, the 2nd and the 3rd highest energy bark bands from each song, and their respective power values.

Besides the standard deviation was extracted throughout bands on each audio track, to prove if in some specific genre all the bands had similar average energy, which would indicate that this genre uses to have a smooth behaviour across frequency bands.

Finally, for each music genre it was calculated:

- The average power on each bark band
- The standard deviation on each bark band
- The average of the standard deviation throughout bands

See appendix B.2 for an extended scheme of the process to better understand the process.

The aim of this measurements was searching for common frequency based component behaviour in music genres, and also corroborate some aspects from the low level features extraction, comparing the energy of the main bark bands.

3.3.3. Rhythm Features

The last method used to identify common features between genres, was extracting the rhythm patterns, as it was explained in section 2.3. Following the idea developed by Lidy and Rauber [11], and the steps described in [14], a rhythm pattern algorithm and the posterior rhythm histogram was developed, modifying slightly the procedure.

A scheme showing the following steps is at appendix B.3 for a better understanding of the process. The rhythm patterns are a matrix representation of fluctuations of critical bands. The first pre-processing step was segment the audio signal in 6 seconds chunks (2^{17} samples as the sampling frequency is 22050 Hz). For each segment, the bark scale spectrogram was computed, using a Hanning window with 23 ms window size (512 samples). Instead of applying the psycho-acoustic transformation into Phon and Sone scale, a Terhardt outer ear model function was used to reflect the specific loudness sensation of the human auditory system. Note that this curve emphasises frequencies between 2 and 5 kHz and attenuates lower and higher frequencies.

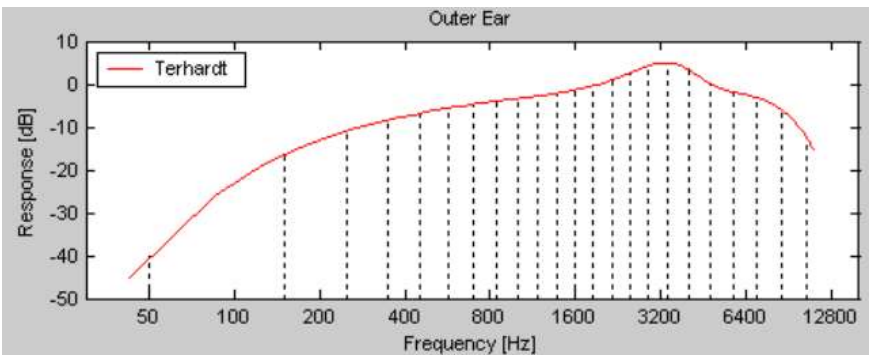

Figure 4: Terhardt outer ear model

In the second part of the algorithm, the energy of each band from the bark scale spectrogram was considered as a modulation of amplitude over time. The next step was applying another time a Fourier transform only taking the frequencies between 0 and 10 Hz (because up to 15 Hz the sensation of roughness starts to rise up significantly), and with a frequency resolution of 0.17 Hz. The result was a 59 bin spectrogram for each of the 23 bark scale bands (the first bin and the first bark scale band were dismissed).

At this point, a certain number of statistical descriptors were computed per critical band to capture additional timbre information. These descriptors are: mean, median, variance, skewness, kurtosis, maximum and minimum. Finally, to compute the rhythm histogram, each of the 59 bins of each band were summed up to create a "rhythmic energy" histogram with 59-bin modulation frequencies.

At the end, after computing the average values of all 6-second segments, a 1577 feature vector was obtained for each audio track: 1357 rhythm pattern features (59 bins in each 23 bark scale bands), 161 statistical descriptors features (7 statistical descriptors in each 23 bark scale band), and 59 rhythm histogram features.

3.3.4. Classification

In order to have numerical results to see how the rhythm pattern features worked, and also if the low level features improved the result, a simple support vector machine was developed to train and test the GTZAN data-set[3].

At this point, two configurations were tested using SVM with a linear, gaussian and polynomial kernel:

- Rhythm pattern features (1577) → 1577 feat.
- Rhythm pattern features (1577) + low level features (14) → 1591 feat.

[3] In this project the study of the optimal configuration for the support vector machine, or other machine learning techniques, has not been deeply taken into account because the classification wasn't the main objective.

4. Results

In this section the results obtained following the methodology specified in chapter 3 are exposed and discussed. A large number of graphs and tables were extracted and analysed, but in this chapter only the results of the features which has allowed a better understood of music genres are presented, and also differentiate them. The rest of the features will be commented but the results will be showed in the appendix C and D.

4.1. Evaluation of the Low Level Features

As explained in section 3.3.1., the following low level features were extracted from each one of the 700 audio tracks data-set: maximum, minimum, RMS, crest factor, dynamic range, centroid, roll-off, low energy and the average power of the song. To sum up the results obtained, and make them easier to analyze and evaluate, some graphs containing averages for each genre were extracted.

In those mentioned graphs, the most important ones presented in Figure 5 and 6, and the others at appendix C.1, it can clearly be observed that there are two music genres that have visible different values in comparison to other genres: classical and jazz music. On one hand, the maximum and minimum values, and also the dynamic range are quite lower than the other genres values, and also the mean power values. According to the perceptual information about music genres explained at section 2.3, these results allow us to prove that this two music genres have low energy components due to the non-electric instruments amplification in the case of classical music, and also in both genres because of the smooth way to play the main instruments in its songs. But on the other hand, we have to take into account the standard deviation of the dynamic range values, which is the highest in both mentioned genres, which indicates that dynamic range values of each song are kind different between them. One of the reasons to have these results might be the huge amount of subgenres and different agrupations than exist on its genres, which makes the audio tracks be quite different inside the same genre, being difficult to stablish a common pattern.

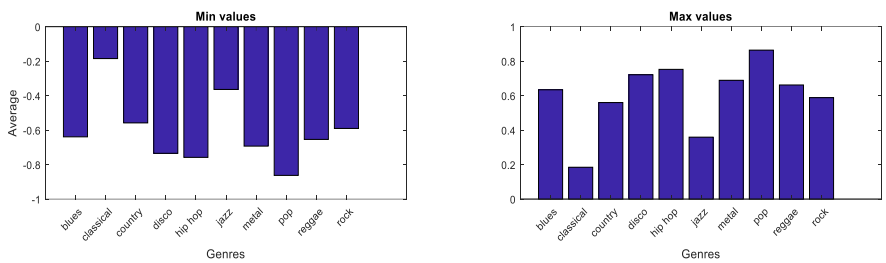

Figure 5:Average values of the maximum and the minimum of each music genre

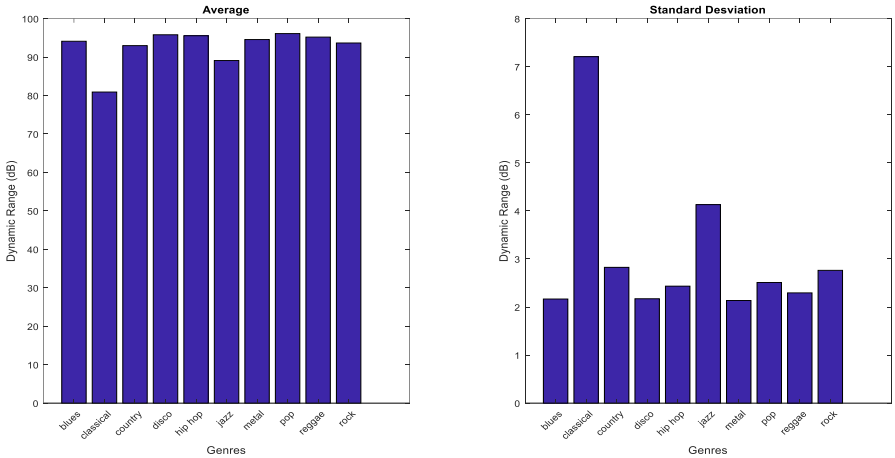

Figure 6: Average and Standard deviation of the dynamic range values

Another results to take into account are the centroid and roll-off values (Figure 7). Pop music has an average roll-off value of 7000 Hz approximately, so this is the frequency where it reaches the 85% of the power of whole song. This indicate that is the music genre which has the energy most distributed in the range 0-7000 Hz. Adding that is the music genre which also has the highest average power values, these results allow to think about the coincidence with the loudness problem explained in section 2.5.8.

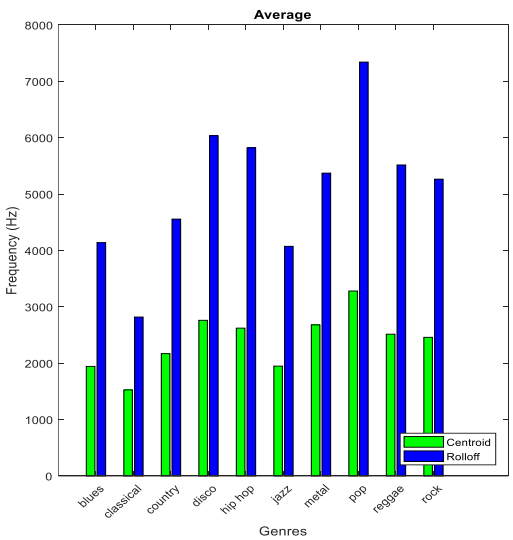

Figure 7: Average of the centroid (green) and roll-off (blue) values

To compare the crest factors values obtained in the two different ways explained in section 3.3.1, a certain number of graphs computing both crest factors were extracted to discuss and be able to see any possible difference. These graphs are shown at appendix C.2.

To have some numerical result and be able to extract some conclusion, the standard deviation between both crest factors values was computed on each audio track. After that, an average of these standard deviation values was computed for each music genre:

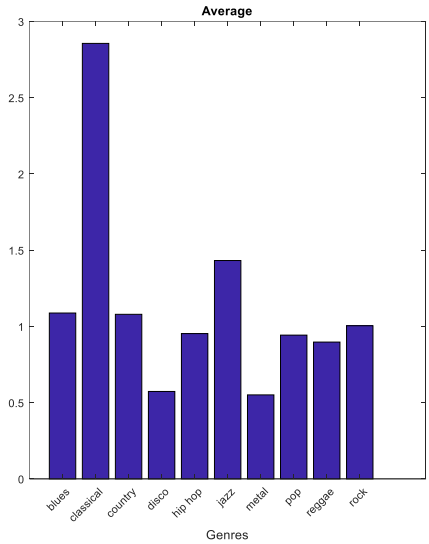

Figure 8: Average through genres of the standard deviation between the two different crest factors

The standard deviation is a measure used to know how concentrated the data is around the mean. The more concentrated, the smaller the standard deviation [40]. If both crest factor values are very similar between them (small standard deviation), this means the loudness of the song doesn't have strong changes, so it has a flat energy behaviour. Besides that, the more different they are, the more loudness changes appear during the song. Looking the results obtained at appendix C.2 and Figure 8, is partly reflected that in classical music the differences between the two crest factor values are substantially different, compared to the other music genres. This result also coincides with the classical music stablished definition, due to the recurrent volume changes along classical music pieces, and the recurrent instrument solos that both classical and jazz music also have.

It is also observed that disco and metal music have the lowest values. This coincides with the idea we have of this kind of music, because both genres have the same constant powerful baseline rhythm during all audio pieces, which makes the crest factor values be similar in whole segments.

4.2. Evaluation of the Bark Scale Spectrograms

This section exposes and discusses the results obtained from section 3.3.2. At appendix D.1 there are some bark scale spectrogram examples to take the first contact to this kind of spectrogram. Because of the application of the Terhardt outer ear correction (see Figure 4), the bark scale bands between 13 and 19 have more importance than the others.

The bark bands with highest energy of each audio track were extracted to prove some common behaviour in some music genres. The graphs containing these results are at appendix D.2. In Table 3 the three most common highest energy bark bands are shown, even with the number of tracks where the bark scale band coincides. For example, for the blues genre, there are 16 tracks which its highest energy band is 4, there are 15 tracks which its highest energy band is 17, and 14 tracks which its highest energy band is 16. Finally, in the last column the percentage of the 70 audio tracks which have one of these 3 bands as a "highest energy band" is presented. Analysing the graphs and the mentioned results, the first observation to take into account is the metal and also disco music behaviour. On one hand, looking at Table 3, almost all of the 70 tracks coincide with the highest energy bands (97.14% for metal and 90% for disco music). With this information, and also checking the "highest energy bands" graphs on appendix D.2, it's clearly observed that the most of the energy of metal and disco songs is compressed between 3-4 bark scale bands containing the frequency range from 2000 Hz to 4000 Hz (according to the Table 1). These results make sense if we compare them with the perceptual definition of metal music, which talks about the constant powerful rhythm and the importance of electric guitars on it, carrying the most of the high frequency components (see Table 2). Disco music has a similar behaviour in terms of frequency according to the definition. Finally, also hip hop music has recurrent main bark bands, as it's similar to disco music, although are not as recurrent as the other two genres.

	1st high energy band	*# of tracks*	*2nd high energy band*	*# of tracks*	*3rd high energy band*	*# of tracks*	*% of the 70 tracks*
Blues	4	16	17	15	16	14	64,28
Classical	6	17	4	11	5	6	48,57
Country	4	19	17	19	16	12	71,42
Disco	17	32	16	26	18	5	90
Hip hop	17	28	16	18	18	8	77,14
Jazz	4	17	16	12	17	9	54,29
Metal	17	34	16	25	18	9	97,14
Pop	17	30	16	12	4	9	72,86
Reggae	16	18	17	15	4	8	58,57
Rock	16	24	17	16	4	8	68,57

Table 3: 1st, 2nd and 3rd most common highest energy band with the number of audio tracks with the same bark band. (last column) Tan per cent of the 70 audio tracks which have the highest energy in these 3 bark bands.

The other genres don't have any common frequency patterns, but we can observe that classical music is the only genre which its "highest energy bands" are between 4 and 7 bark bands, instead of the bands grouping the frequencies between 2 and 5 kHz.

These results were corroborated by computing the standard deviation of every bark band, and then computing the average for each genre (the results are shown at appendix D.5). In the bar graphs mentioned, it's seen that metal and disco genres have small standard deviations in the most of the bark scale bands (followed by hip hop and pop), which means that every of the 70 metal or disco songs have a very similar frequency distribution. Another observation made was that the lowest bark bands in all genres have always smaller standard deviation values, except classical music. These results could indicate that, on one hand, the most of the genres have a similar energy component on low frequencies. On the other hand, as commented on other sections, classical music has a completely different behaviour compared to the other music genres, and we cannot take the frequency based components as a baseline to define classical music, due to as commented before, the different amount of subgenres it has.

The power of the highest energy band was also computed (at appendix D.4 are shown the results, and at Figure 9 the average for each genre), and the results corroborate what was observed in section 4.1: classical music has small power values compared to the rest of music genres, and also jazz and reggae music on a smaller scale. On the other hand, metal has the highest power values, followed by pop music.

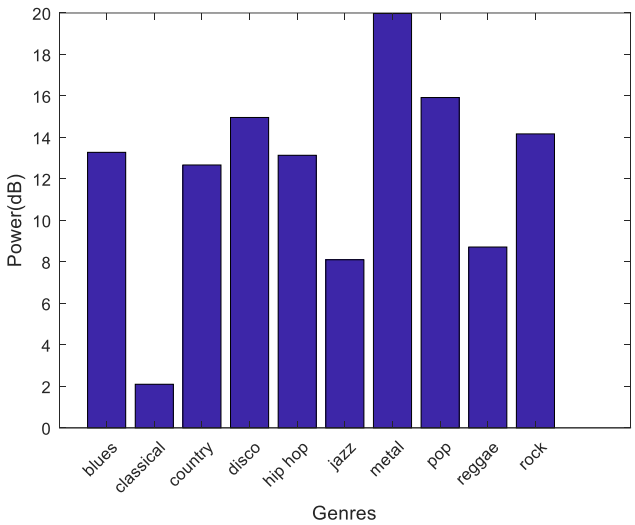

Figure 9: Genres average of the "highest energy band" power

A similar result was extracted to compare the energy of the audio tracks. The average power of each bark band was computed for all music genres. The results, exposed at appendix D.3 allow to prove the energy similarities between classical music and jazz (the most of the energy is negative because of the Terhardt outer ear model correction), or between metal and rock, which its energy is higher than the other music genres in all bark

scale bands. Also observing the pop genre values, we can see that it has almost similar power to all bark bands, supporting the roll-off values exposed at section 4.1, and also supporting the theory of the loudness war.

4.3. Rhythm features results

In this section, the results of the chapter 3.3.3. are shown, although are not analyzed because the features are later evaluated by a SVM classifier (section below). To see the difference in rhythm patterns shape between music genres, some rhythm patterns spectrogram were computed, and also its respective rhythm histogram.

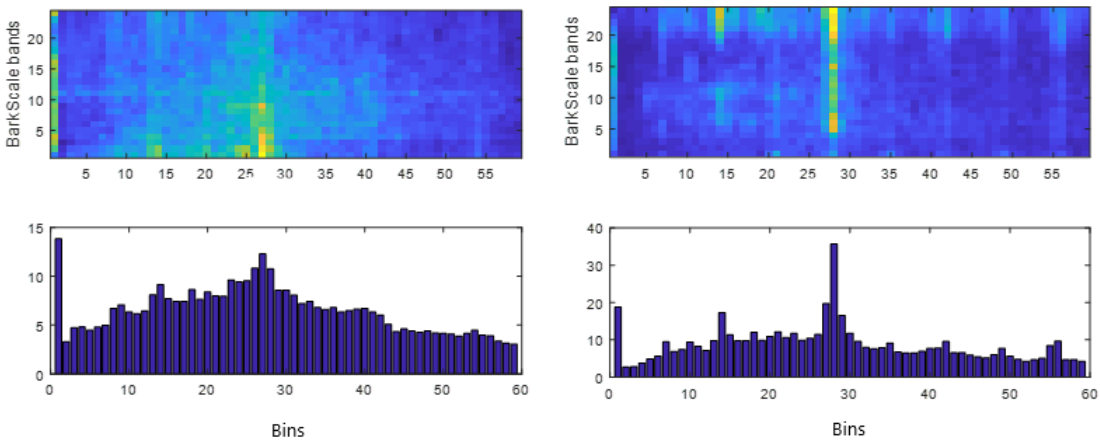

Figure 10: Rhythm patterns spectrogram (above) and rhythm histogram (below) of classical audio track (left) and metal audio track (right)

The most differentiate genres were chosen in order to easily see the main differences: classical and metal. Classical music rhythm patterns are more flat and never will have an extreme peak, due to the lack of rhythmic instruments. All the other genres, and specially rock and metal, have clear peaks indicating the song bpm approximately.

4.4. Evaluation and classification of the extracted features

As explained in section 3.3.4. a classifier was built to check how the rhythm pattern algorithm worked, and also analyse if adding other features extracted during the present work, the results improved or not. Two different configurations were evaluated:

- Configuration 1: Rhythm pattern features (1577) → 1577 feat.
- Configuration 2: Rhythm pattern features (1577) + low level features (14) → 1591 feat.

In order to evaluate how the classifier worked, some evaluation measures, explained at appendix E.2, were extracted: Accuracy, Precision, Recall and F_score.

		Train				Test			
	Kernel	**Acc.**	**Prec.**	**Rec.**	**F_s**	**Acc.**	**Prec.**	**Rec.**	**F_s**
Conf. 1	Linear	0.8814	0.8881	0.8814	0.8848	0.6333	0.6078	0.6333	0.6203
	Gaussian	0.9814	0.9824	0.9814	0.9819	0.6567	0.6389	0.6567	0.6477
	Polynomial	0.9986	0.9986	0.9986	0.9986	0.6600	0.6609	0.6600	0.6605
Conf. 2	Linear	0.8929	0.8973	0.8929	0.8951	0.6633	0.6390	0.6633	0.6510
	Gaussian	0.9814	0.9824	0.9814	0.9819	0.6533	0.6376	0.6533	0.6454
	Polynomial	1	1	1	1	0.6633	0.6682	0.6633	0.6658

Table 4: Evaluation measures

After evaluating both configurations for different kernels, the best results in both cases are achieved with a polynomial kernel.

As it was observed in other sections, the variables to classify (in this case music genres) have huge differences between them, so in order to analyze them individually, the accuracy for each music genre was computed in Table 5.

The first observation is that the low level features doesn't improve the results. Apart, as it can be observed, there are 3 genres which have high accuracy: classical music, metal and jazz. On the other hand, blues and rock music have very low results compared to the total accuracy, which decreases the total accuracy (without these two genres, the total accuracy would increase almost 10 points).

Observing the confusion matrix of each configuration (appendix E.1), it's seen that a huge amount of audio tracks are wrong classified as country or rock. This could happen because the most of the genres are rhythmically similar and taking into account only the rhythm behaviour is easy to classify them wrong.

Genre	Acc. 1	Acc. 2
Blues	0,4333	0,4667
Classical	0,9333	0,9000
Country	0,8000	0,7667
Disco	0,5333	0,5333
Hip hop	0,6333	0,6000
Jazz	0,8333	0,8667
Metal	0,9667	0,9667
Pop	0,6000	0,6000
Reggae	0,5667	0,6000
Rock	0,2667	0,3000

Table 5: Accuracy of each music genre

5. Budget

This project has been developed during 20 weeks, as considered in Gantt diagram. The resources used are available for free on the Internet: all bibliography, Matlab software (free for UPC students) and also MIRToolbox, as explained on section 3.2.

Accordingly, the budget only consists of the cost associated with the salaries of the persons who worked in this project: one junior engineer (author of the work), and one senior engineer (project supervisor).

Position	**Salaries**	**Dedication**	**Total**
Junior engineer	12 €/hour	40 h/week	9.600 €
Senior engineer	25 €/hour	1h/week	500 €
			10.100 €

6. Conclusions and future development:

Looking at the results, and also taking into account the perceptual definitions explained in section 2.5, some final conclusions are extracted, and also future work in this direction is proposed.

The first conclusion, which supports what everybody would expect before starting reading this work, is that classical music is the most different genre of the ten which has been analysed. The dynamic, statistical and also the frequency values are at a huge distance from the other values, and being the genre which has one of the highest accuracy rates in the classification, corroborates the theory. Apart, it is also observed looking at its high standard deviation values computed in all the features along the thesis, that there are not a lot of songs with similar values. This denotes that a lot of subgenres inside classical music exist, with different registers, number of instruments, etc. and that makes it difficult, or even non-sense, to analyse classical music in general, and it would be better to analyse specific classical subgenres.

In this direction, it was observed that jazz music has a similar behaviour to classical music. A priori, these two genres don't have anything in common, but it has been observed that, technically, they have many features in common. One of the reasons of this common behaviour, may be because of the simple baseline instruments structure (the most of the jazz songs have only drums, bass and a soft guitar), the quiet way to play the instruments during the pieces (no distortion, no hard percussive drums, etc) and for the recurrent presence of instrumental solos.

Another conclusion to extract is about metal music. The similar frequency and rhythm behaviour makes it one of the easiest genres to identify from the technical point of view, and also classify. But on the other hand, rock music, that theoretically is the closest genre to metal music, is the one with the worst standard deviation and accuracy values during the classification. This may happen because of the wide kinds of rock music variety, which doesn't allow to establish a common pattern and is confused by the classifier as some other genres. It is also true that sometimes music experts are not able to distinguish rock music from other genres because the line between them sometimes is very thin, so it's also normal that rock music is the most difficult to define and classify correctly.

The results of pop music also coincide with their definition. It's true that the database used is from 2002, but what's called "loudness war" started so many years before, and it can be seen in the results exposed. It has to be said that having a data-set which the audio tracks lasts 30 seconds make it difficult to be completely sure of the loudness behaviour of whole piece, as we only have a quarter of the song. So to corroborate these results it would be useful to analyse a database with complete songs.

About the classification, it was not possible to achieve good results compared to other studies, but as explained in the Introduction, this wasn't one of the main objectives of this project, it was only a way to check the rhythm pattern features.

First of all, it's clear that the low level features don't improve the classification. Moreover, the optimization of the SVM wasn't taken so much into account, and maybe study more configurations or also other machine learning techniques would have given better results. The rhythm patterns algorithm works very well with the most differentiate music genres, as metal, classical or jazz music, but more information is needed and features to characterize and at the end improve the classification on the other music genres, especially rock and blues music.

Taking this project as starting point, interesting future work could be the analysis of classical music subgenres: what are the differences between different kinds of orchestra (baroque, symphonic, philharmonic, etc.) or orchestras with different instruments (chord or wind instruments). This idea could also be valid in other genres, like rock or jazz, or also electronic music. Other future work, in the direction on MIR researchers, could be improving the rhythm features extraction, also improving the classifier with other features, like MFCC or some spectrogram image processing, or check the different machine learning techniques to achieve better accuracy results.

Bibliography:

[1] K. Kosina. Music Genre Recognition. *Diploma thesis, June 2002.*

[2] G. Tzanetakis, G. Essl and P. Cook. Automatic Music Genre Clasification Of Audio Signals. *2002*.

[3] R. de Lima, Y. Maldonado and L. Nanni. Music Genre Recognition Using Spectrograms with Harmonic-Percussive Sound Separation. *Computer Science Society, 35th International Conference of the Chilean, October 2016.*

[4] Perrot, D., and Gjerdigen, R.O. Scanning the dial: An exploration of factors in the identification of musical genres. *Journal of New Music Research, 37: 2, 93-100, 2010.*

[5] M. Caetano and X. Rodet. Improved estimation of the amplitude envelope of time-domain signals using true envelope cepstral smoothing. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2011*.

[6] Paolo Annesi, Roberto Basili Raffaele Gitto, Alessandro Moschitti. Audio Feature Engineering for Automatic Music Genre Classification. *Department of Computer Science, Systems an d Production, University of Roma, Italy, June 2007*.

[7] N. M. Norowi, S. Doraisamy, R. Wirza. Factors affecting automatic genre classification: an investigation incorporating non-western musical forms. *6th International Conference on Music Information Retrieval, ISMIR, January 2005.*

[8] L. Nanni, Y. M. G. Costa, A. Lumini, M. Y. Kim, S. R. Baek. Combining visual and acoustic features for music genre classification. *Expert Systems with Applications, Volume 45, Pages 108-117, March 2016*.

[9] B. Logan. Mel frequency cepstral coefficients for music modelling. *Cambridge Research Laboratory, 2000.*

[10] F. Gouyon, S. Dixon, E. Pampalk and G. Widmer. Evaluating rhythmic descriptors for musical genre classification. *AES 25th International Conference, London, June 2004.*

[11] T. Lidy and A. Rauber. Combined Fluctuation Features for Music Genre Classification. *MIREX 2005.*

[12] E. Pampalk. Speeding up music similarity. *"nd Annual Music Information Retrieval Evaluation eXchange, MIREX 2005.*

[13] H. Fastl and E. Zwicker. Psycho-acoustics, 1999.

[14] M.Cord and P.Cunningham. Machine Learning Techniques for Multimedia. *Pages 251-262, 2008.*

[15] M. Biss. Rhythm Tips for Identifying Music Genres by Ear. *August 2015.*

[16] E.Guaus and E.Batlle. Visualization of metre and other rhythm features

[17] Wikipedia – Spectral Centroid. https://en.wikipedia.org/wiki/Spectral_centroid

[18] E. Fonseca, J. Pons , X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra. Freesound Datasets: a Platform for the Creation of Open Audio Datasets. *18th International Society for Music Information Retrieval Conference, 2017.*

[19] Music Analysis,Retrieval and Synthesis for Audio Signals. G. Tzanezakis. http://marsyasweb.appspot.com/download/data_sets/

[20] Wikipedia – Spectrogram. https://en.wikipedia.org/wiki/Spectrogram

[21] Wikipedia – Bark spectrogram. https://en.wikipedia.org/wiki/Bark_scale

[22] Wikipedia – Root Mean Square. https://en.wikipedia.org/wiki/Root_mean_square

[23] Wikipedia – Crest Factor. https://en.wikipedia.org/wiki/Crest_factor

[24] Wikipedia – Dynamic Range https://en.wikipedia.org/wiki/Dynamic_range

[25] O. Lartillot, P. Toiviainen. A Matlab Toolbox For Music Feature Extraction From Audio. *10th Int. Conferences on Digital Audio Effects (DAFx-07), Bordeaux, France, September 2007.*

[26] GNU General Public Licence. http://www.gnu.org/licenses/gpl-2.0.txt

[27] E. Terhardt. Calculation of virtual pitch. *Hearing Research, Volume 1, Issue 2, March 1979.*

[28] T. Jehan. Chapter 3: Music Listener. *Creating Music by Listening, PhD Thesis, Massachusetts Institute of Technology, September 2005.*

[29] Wikipedia – Blues. https://en.wikipedia.org/wiki/Blues

[30] Wikipedia – Classical Music. https://en.wikipedia.org/wiki/Classical_music

[31] Wikipedia – Country. https://en.wikipedia.org/wiki/Country_music

[32] Wikipedia – Disco. https://en.wikipedia.org/wiki/Disco

[33] Wikipedia – Hip hop. https://en.wikipedia.org/wiki/Hip_hop_music

[34] Wikipedia – Jazz. https://en.wikipedia.org/wiki/Jazz

[35] Wikipedia – Heavy metal. https://en.wikipedia.org/wiki/Heavy_metal_music

[36] Wikipedia – Pop music. https://en.wikipedia.org/wiki/Pop_music

[37] E. Vickerls. The Loudness War: Background, Speculation, and Recommendation. *November 2010.*

[38] Wikipedia – Reggae. https://en.wikipedia.org/wiki/Reggae

[39] Wikipedia – Rock music. https://en.wikipedia.org/wiki/Rock_music

[40] Wikipedia – Standard deviation. https://en.wikipedia.org/wiki/Standard_deviation

[41] Genre Definitions as used in the KOOP Music Library.. http://www.koop.org/library/genres-definitions

[42] Frequency Range of Musical Instruments. http://www.dummies.com/home-garden/car-repair/frequency-range-of-musical-instruments-for-car-audio/

[43] International Society for Music Information Retrieval (ISMIR) official website. http://www.ismir.net/

[44] Music Information Retrieval official website. https://musicinformationretrieval.com/

[45] Y. Costa, L. Oliveira, A. Koericb, F. Gouyon. Music genre recognition using spectrograms. *18th International Conference on Systems, Signals and Image Processing (IWSSIP), June 2011.*

[46] J. Aucouturier and F Pachet. Representing Musical Genre: A State of the Art. *Journal of New Music Research, 32:1, 83-93, August 2010.*

[47] A. Pons. Measuring the Evolution of Timbre in Billboard Hot 100. *Degree's Thesis in Audiovisual Systems Engeenering, UPC, 2017.*

[48] Oxford Music Online. http://www.oxfordmusiconline.com/

Appendices:

Appendix A: Additional information about the Work Plan

A.1. Work Packages

Project: Project Proposal and Work Plan	**WP ref**: 1	
Major constituent: Documentation	Sheet 1 of 7	
Short description: Establish the project basis and the main goals between my host university and UPC, and develop my work plan.	**Planned start date**: 12/02/2018 **Planned end date**: 05/03/2018 Start event: T1 End event: T3	
Internal task T1: Project description **Internal task T2**: Project development plan **Internal task T3**: Document review	**Deliverables**: AndreuBoadas_ ProjectProposal. Pdf	**Dates**: 05/02/2018

Project: Information research and documentation	**WP ref**: 2	
Major constituent: Documentation	Sheet 2 of 7	
Short description: Study of the state of the art of the project, and look for information and recent similar projects and ideas for improvements.	**Planned start date**: 12/02/2018 **Planned end date**: 30/04/2018 **Start event**: T1 **End event**: T4	
Internal task T1: Study of the state of the art of music properties **Internal task T2**: Study of the methods of possible analysis of the music signal. **Internal Task T3**: Study of well-known methods to feature extraction of the music signal. **Internal Task T4**: Choose the best methods for music signal than will be applied to the database	**Deliverables**: -	**Dates**: -

Project: Software development	**WP ref**: 3	
Major constituent: Software	Sheet 3 of 7	
Short description: Develop a Matlab program to be able to extract the chosen features from the audio sets of the database.	**Planned start date**: 19/03/2018 Planned end date: 21/05/2018 **Start event**: T1 **End event**: T3	
Internal task T1: Develop a program which extracts spectrogram and statistical and timbre descriptors **Internal task T2**: Develop the rhythm pattern descriptor **Internal task T3**: Develop the SVM classifier **Internal task T4**: Develop the final feature extractor program	**Deliverables:** -	**Dates:** -

Project: Critical review	**WP ref:** 4	
Major constituent: Documentation	Sheet 4 of 7	
Short description: Document that discusses the project until the date and also includes a review of the work plan.	**Planned start date**: 03/04/2018 **Planned end date**: 07/05/2018	
	Start event: T1 **End event:** T3	
Internal task T1: Writing about progress to date **Internal task T2**: Update the work plan **Internal task T3**: Document review and approval	**Deliverables**: AndreuBoadas_ CriticalReview.pdf	**Dates**: 07/05/2018

Project: Test and results	**WP ref**: 5	
Major constituent: Software	Sheet 5 of 7	
Short description: Test the program developed, compares the results obtained with the state-of-art, and extract the conclusions of the project.	**Planned start date**: 07/05/2018 **Planned end date**: 10/06/2018	
	Start event: T1 **End event**: T3	
Internal task T1: Software Test implementation **Internal task T2**: Comparison of the results **Internal task T3**: Result assessment	**Deliverables**: -	**Dates**: -

Project: Final report	**WP ref**: 6	
Major constituent: Documentation	Sheet 6 of 7	
Short description: Document that describes the entire project once it has been finished.	**Planned start date**: 01/05/2018 **Planned end date**: 22/06/2018	
	Start event: T1 **End event**: T2	
Internal task T1: Writing about the project, with the results and conclusions **Internal task T2**: Project review and approval	**Deliverables**: AndreuBoadas_ FinalReport.pdf	**Dates**: 22/06/2018

Project: TFG Presentation	**WP ref**: 7	
Major constituent: Documentation	Sheet 7 of 7	
Short description: Oral presentation that overviews the project	**Planned start date**: 23/06/2018 **Planned end date**: 27/06/2018	
	Start event: T1 **End event**: T2	
Internal task T1: Slides drafting **Internal task T2**: Prepare the presentation	**Deliverables**: AndreuBoadas_ TFGPresntation.pdf	**Dates**: 28/06/2018

A.2. Milestones

WP#	Task#	Short title	Milestone / deliverable	Date (week)
1	1	Project description	Documentation	1,2
1	2	Project development plan	Documentation	1,2,3
1	3	Document review and approval	AndreuBoadas_ProjectProposal.pdf	3
2	1	Study of the state-of-the-art	Documentation	1-4
2	2	Analysis of the music signal	Documentation	4,5
2	3	Feature extraction methods	Documentation	4,5
2	4	Chose the best methods	Documentation	6
3	1	Spectrogram & statistical descriptors	Software	6-10
3	2	Rhythm patterns descriptors	Software	9-11
3	3	Develop SVM classifier	Software	11,12
3	4	Develop final feature extraction program	Software	13-15
4	1	Writing about progress to date	Documentation	10,11
4	2	Update the work plan	Review	12
4	3	Document review and approval	AndreuBoadas_CriticalReview.pdf	13
5	1	Software test implementation	Testing software	15
5	2	Comparison of the results	Software/documentation	16
5	3	Result assessment	Result reports	16
6	1	Writing about the project and conclusions	Documentation	14-17
6	2	Project review and approval	AndreuBoadas_FinalReport.pdf	17
7	1	Slides drafting	Draft	18
7	2	Prepare the presentation	Documentation	18
7	3	Presentation review and approval	AndreuBoadas_TFGPresentation.pdf	19

Appendix B: Methodology schemes

B.1. Low Level Features Scheme

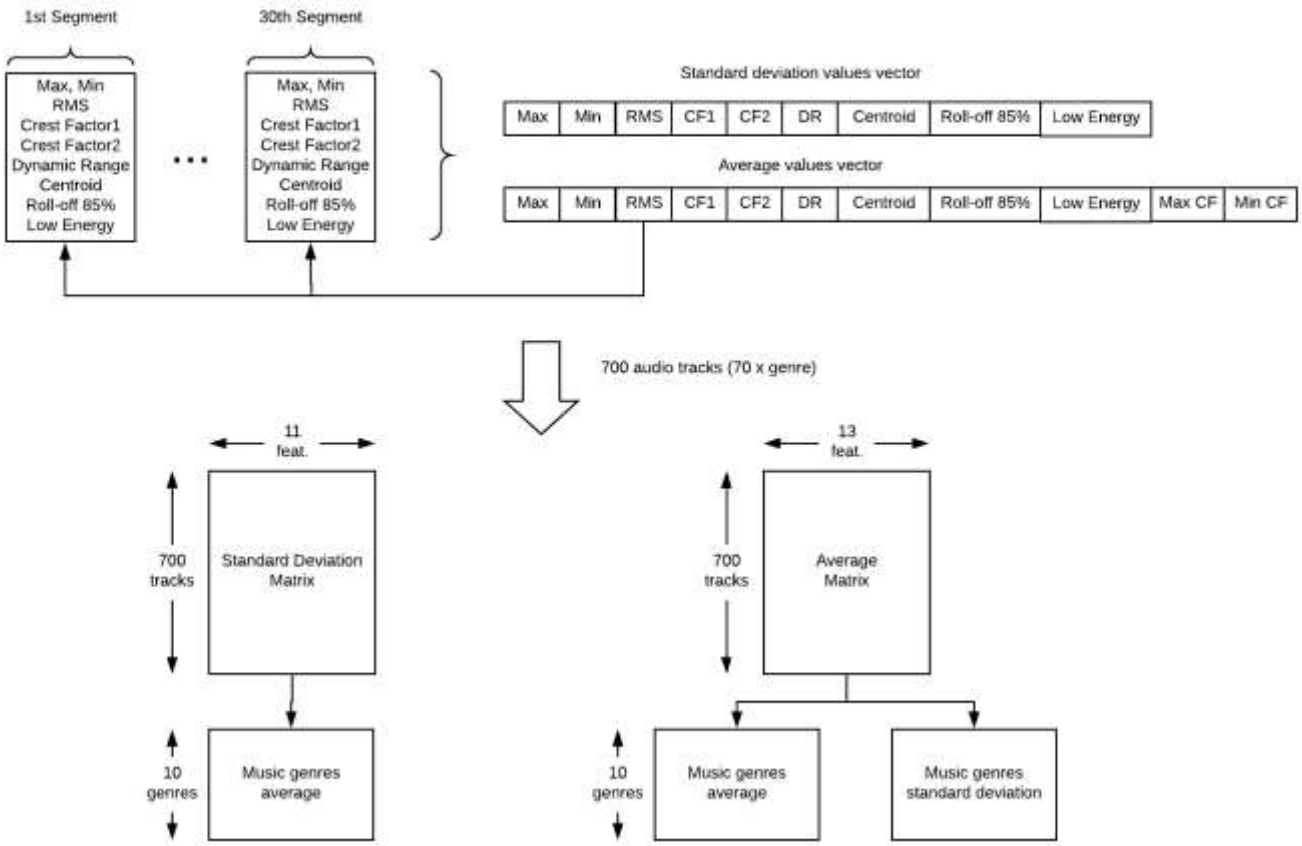

Figure 11: Low level features shceme (general)

B.2. Bark Scale Spectrogram Scheme

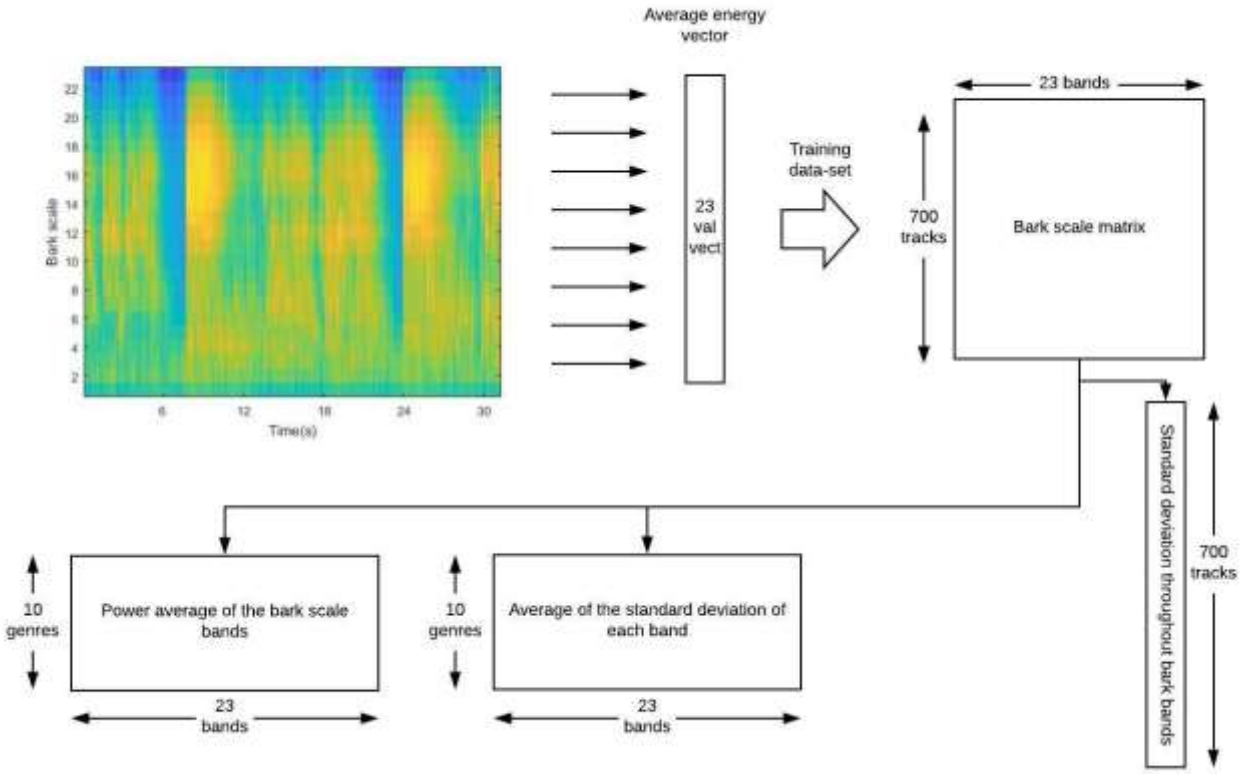

Figure 12: Bark scale spectrograms scheme (general)

B.3. Rhythm Features Scheme

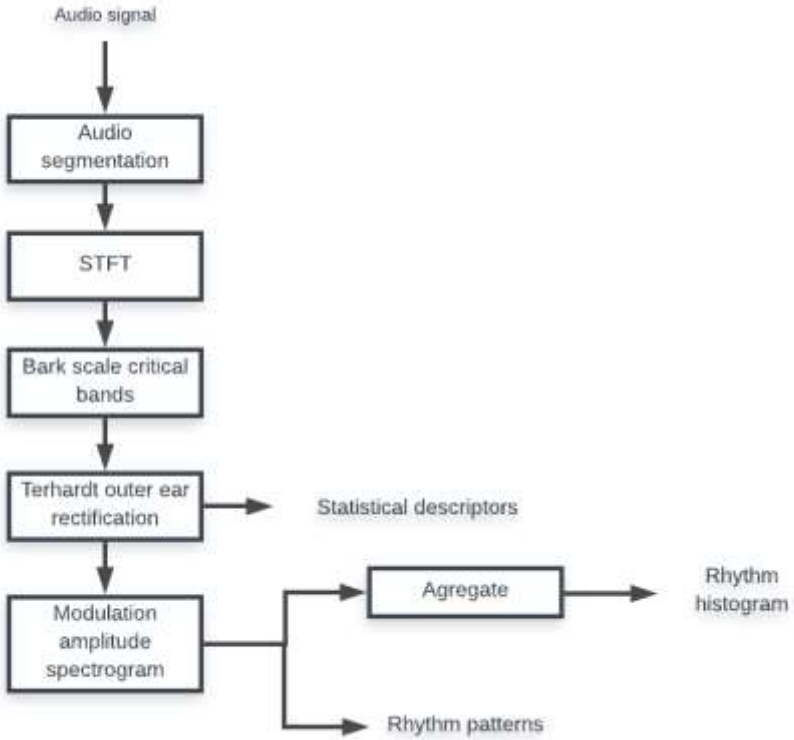

Figure 13: Rhythm features scheme

Appendix C: Low level features results

C.1: Average and standrad deviation graphs

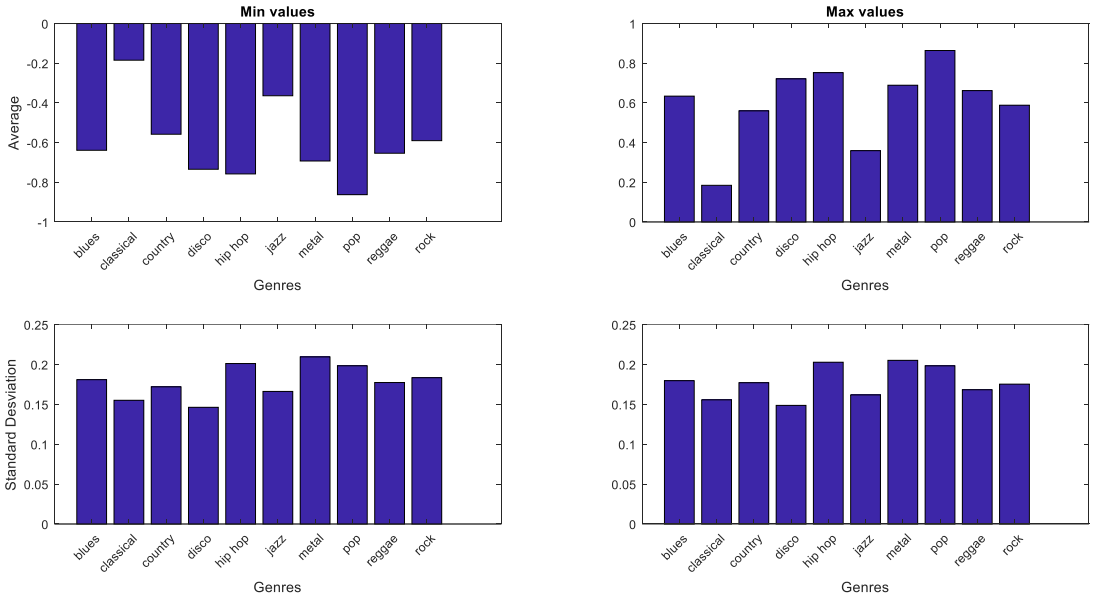

Figure 14: Average and Standard deviation values of the maximum and the minimum of each music genre

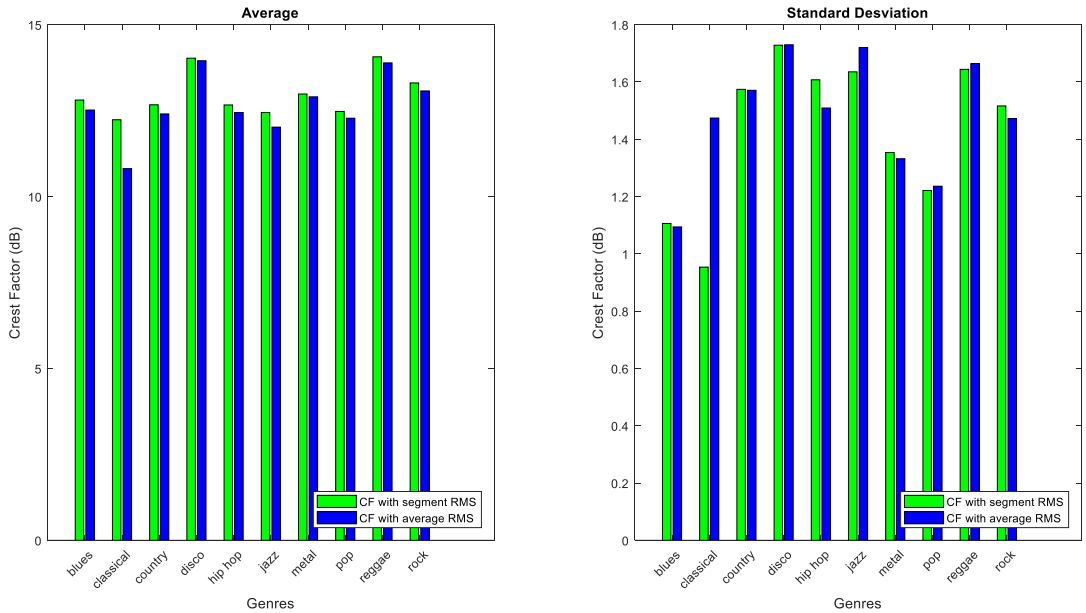

Figure 15: Average and Standard deviation of the crest factors computed in the two different ways

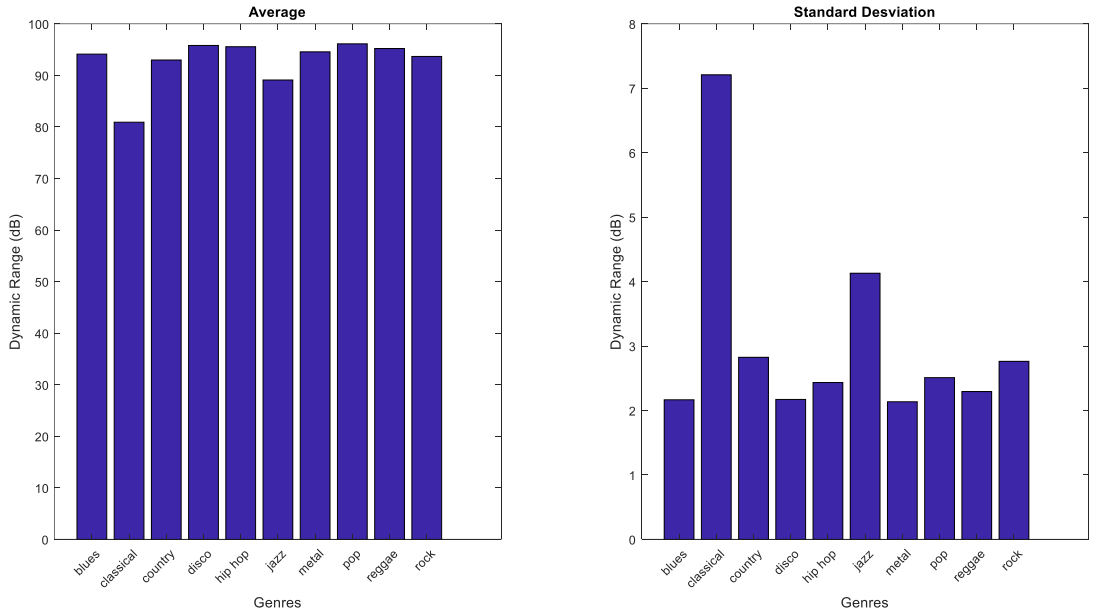

Figure 16: Average and Standard deviation of the dynamic range values

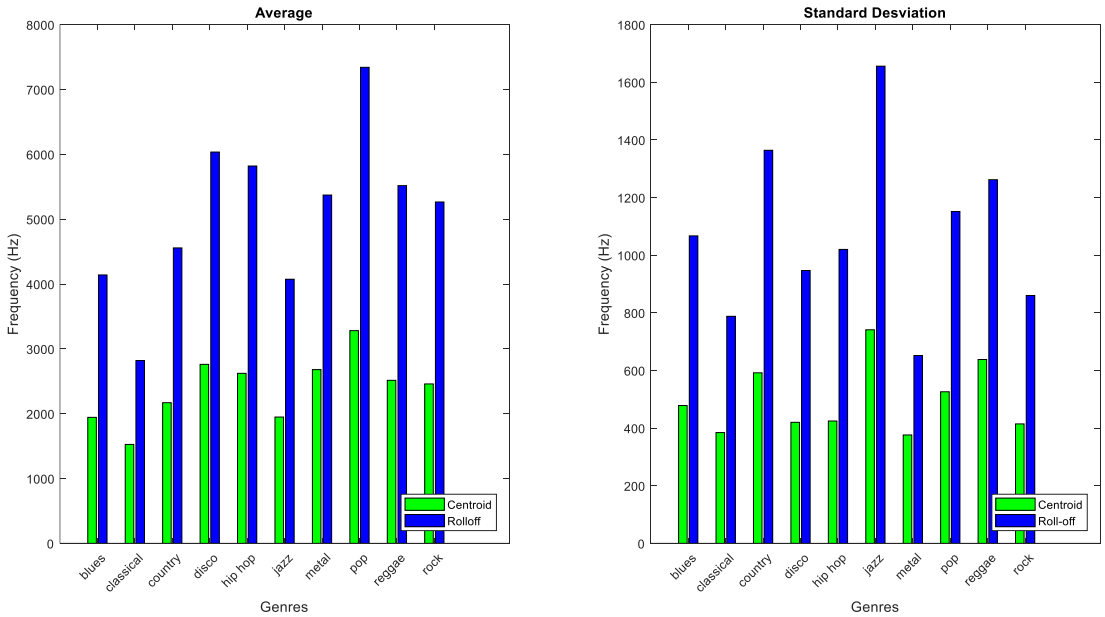

Figure 17: Average and Standard deviation of the centroid (green) and roll-off (blue) values

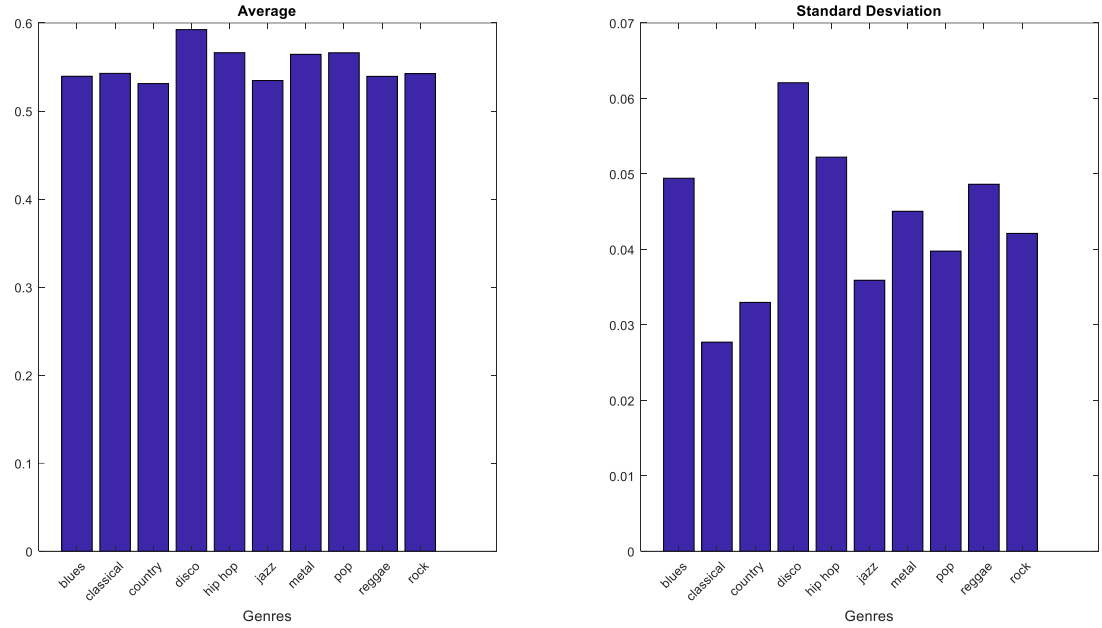

Figure 18: Average and Standard deviation of the low energy values

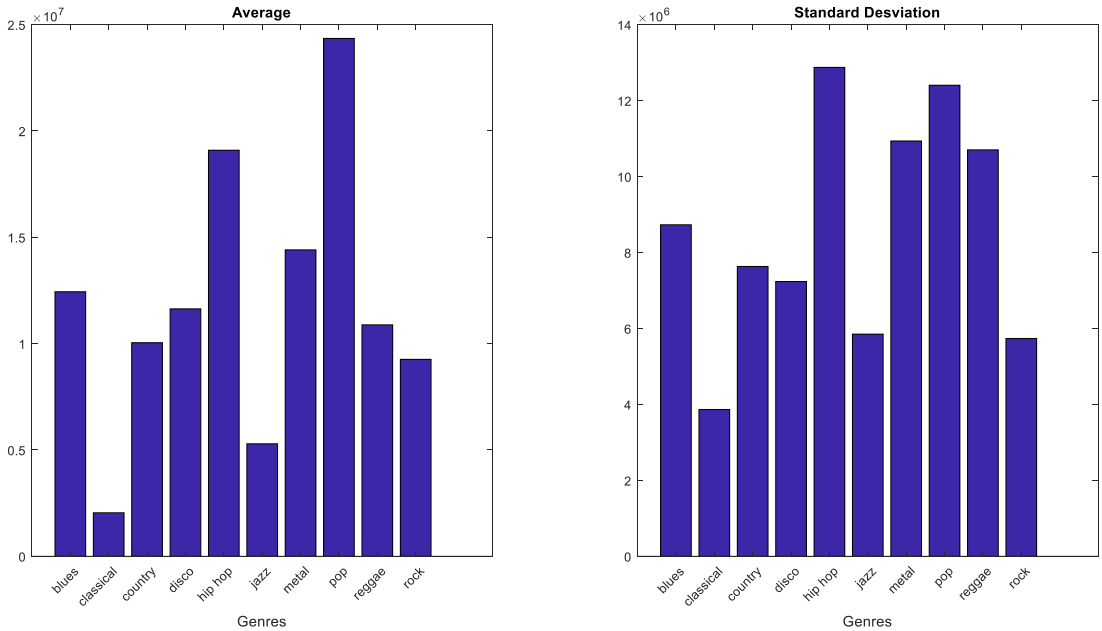

Figure 19: Average and Standard deviation of the mean power values

C.2 Examples of Crest factor graphs

Note: It was not possible to show the legend and the X and Y axis of the graphs because of the space of the window size.

Legend: · Crest Factor computed with the RMS mean of the segment (orange)

· Crest Factor computes with the RMS of each segment (blue)

Axis: · X axis: segments of the audio track

· Y axis: crest factor value (dB)

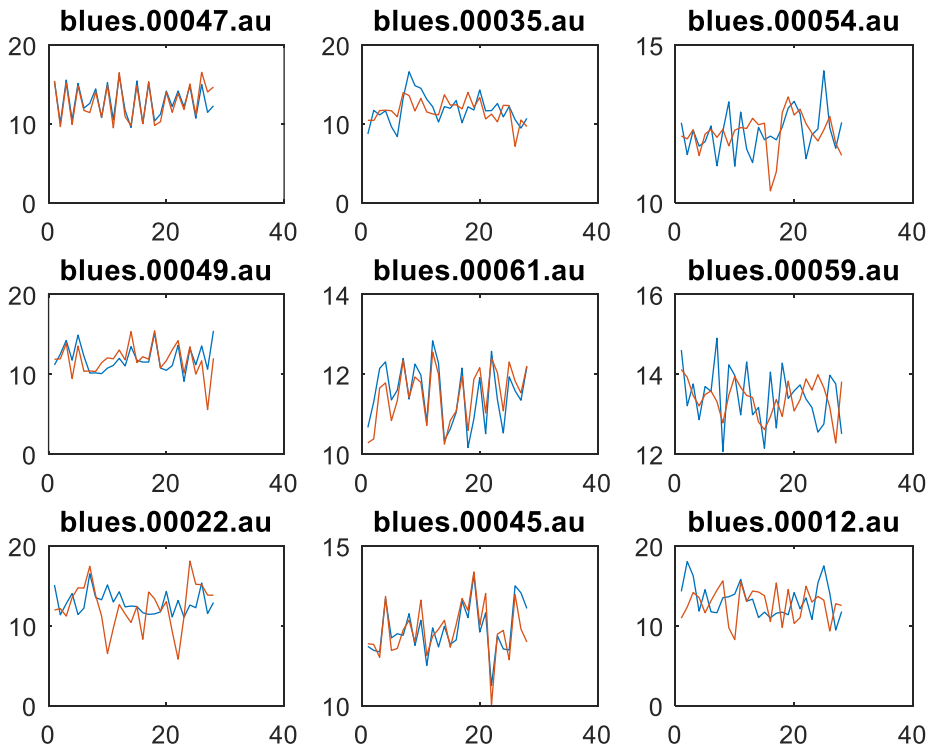

Figure 20: Comparison of the crest factors values in 9 blues audio tracks

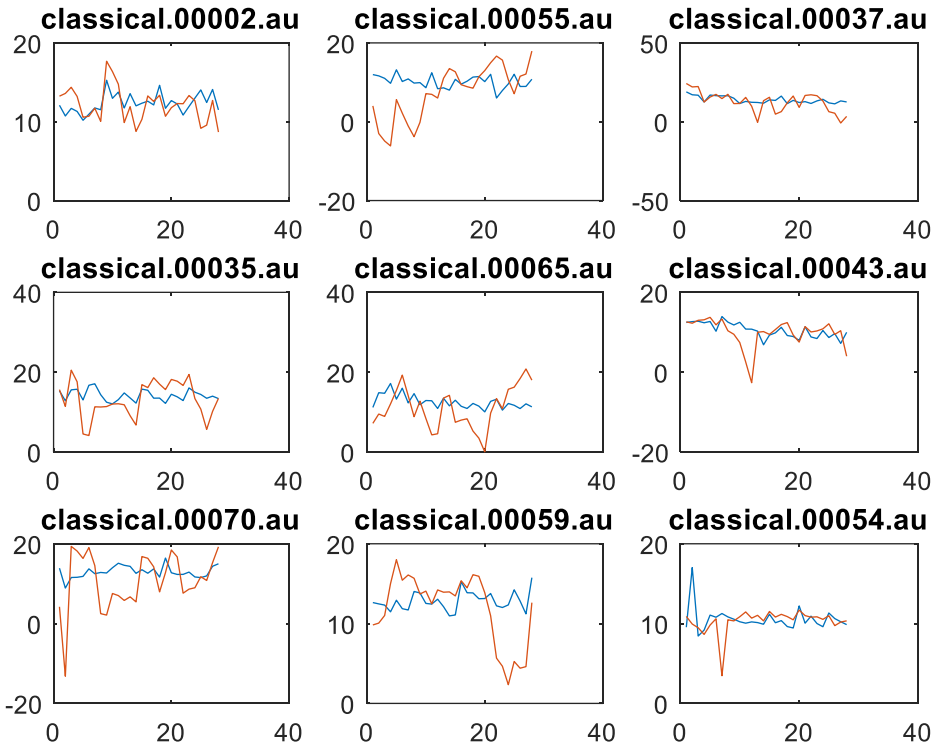

Figure 21: Comparison of the crest factors values in 9 classical audio tracks

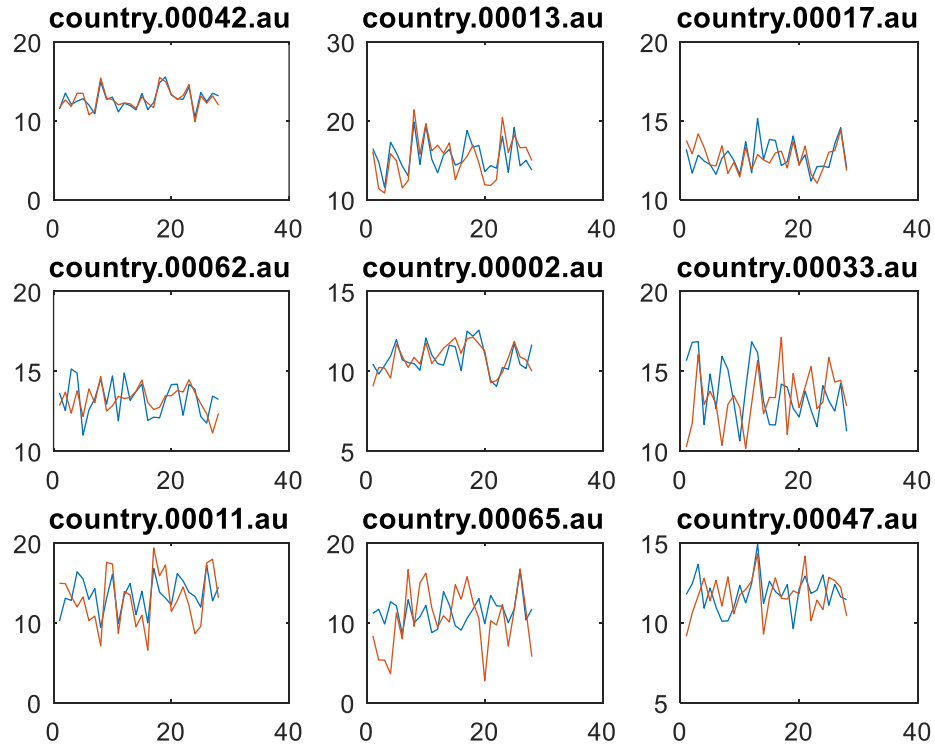

Figure 22: Comparison of the crest factors values in 9 country audio tracks

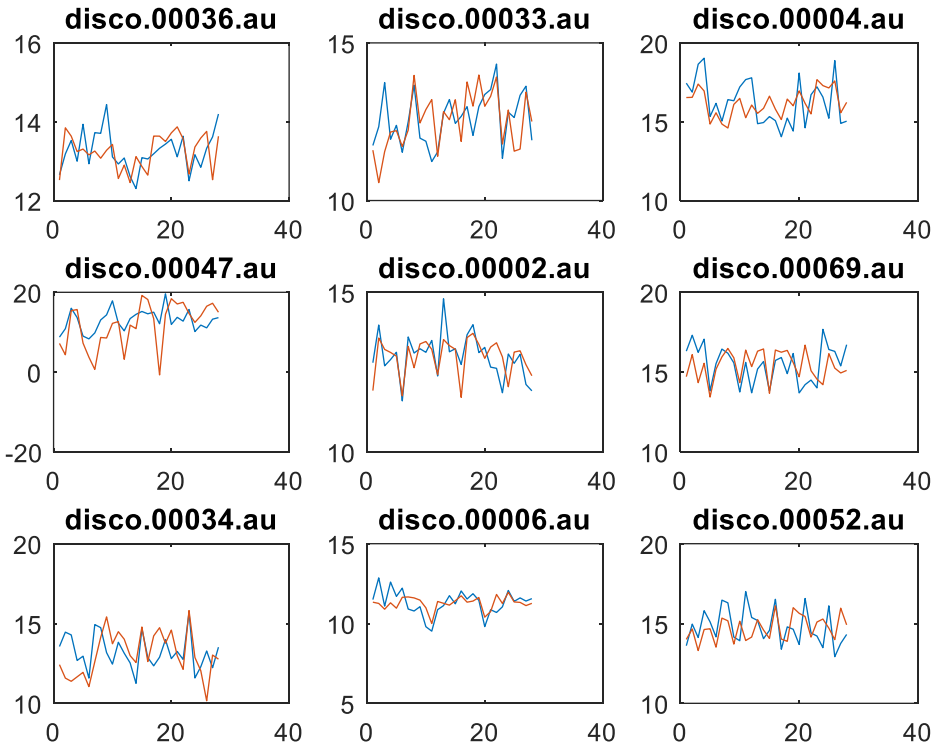

Figure 23: Comparison of the crest factors values in 9 disco audio tracks

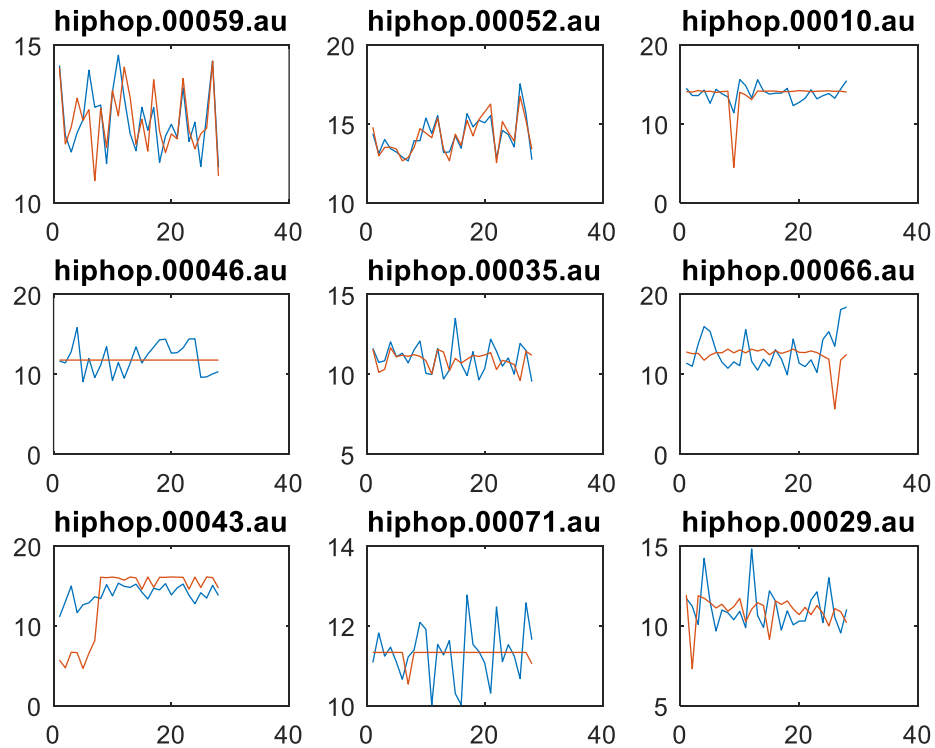

Figure 24: Comparison of the crest factors values in 9 hip hop audio tracks

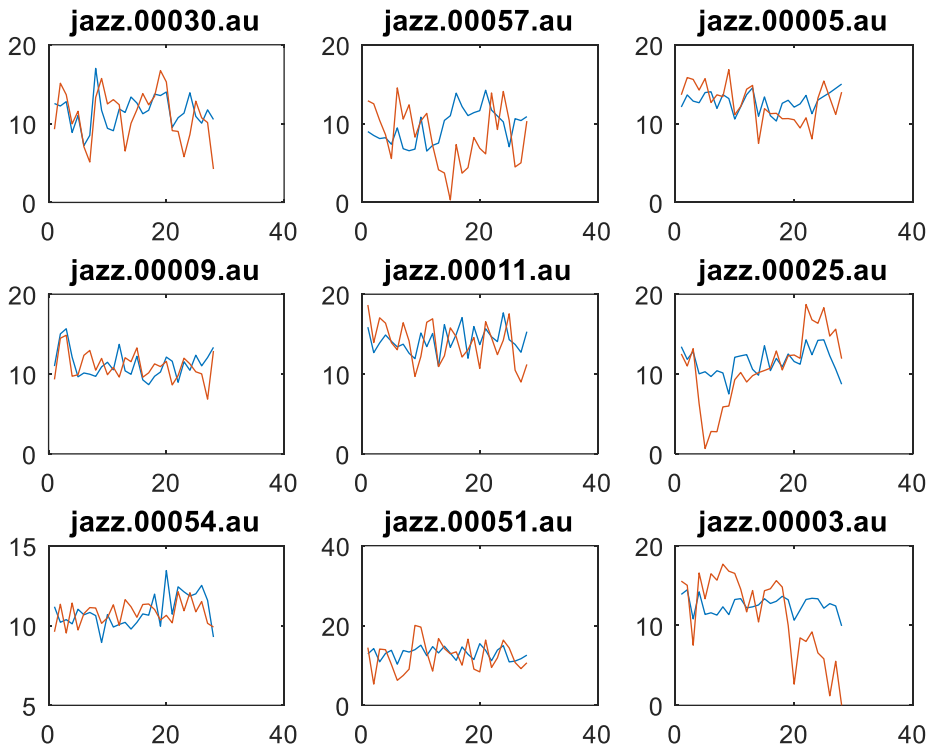

Figure 25: Comparison of the crest factors values in 9 jazz audio tracks

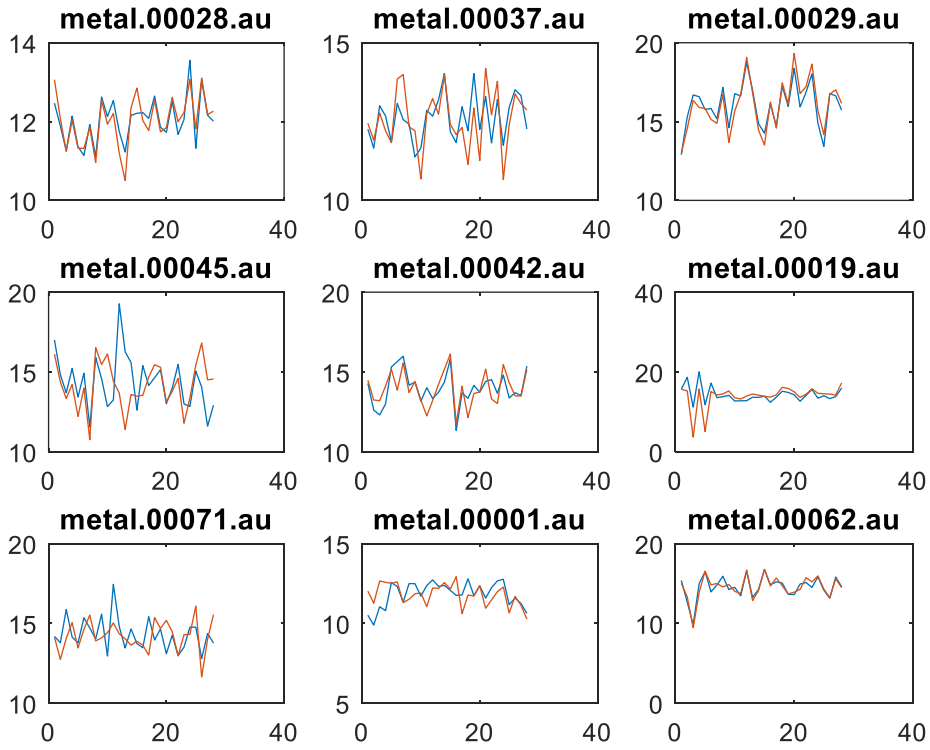

Figure 26: Comparison of the crest factors values in 9 metal audio tracks

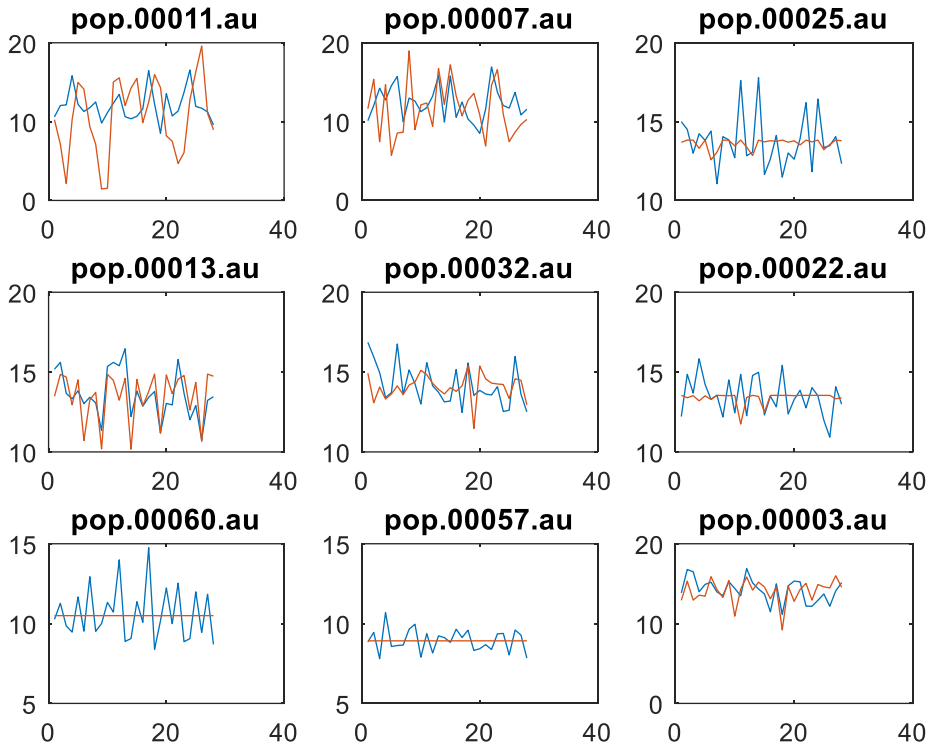

Figure 27: Comparison of the crest factors values in 9 pop audio tracks

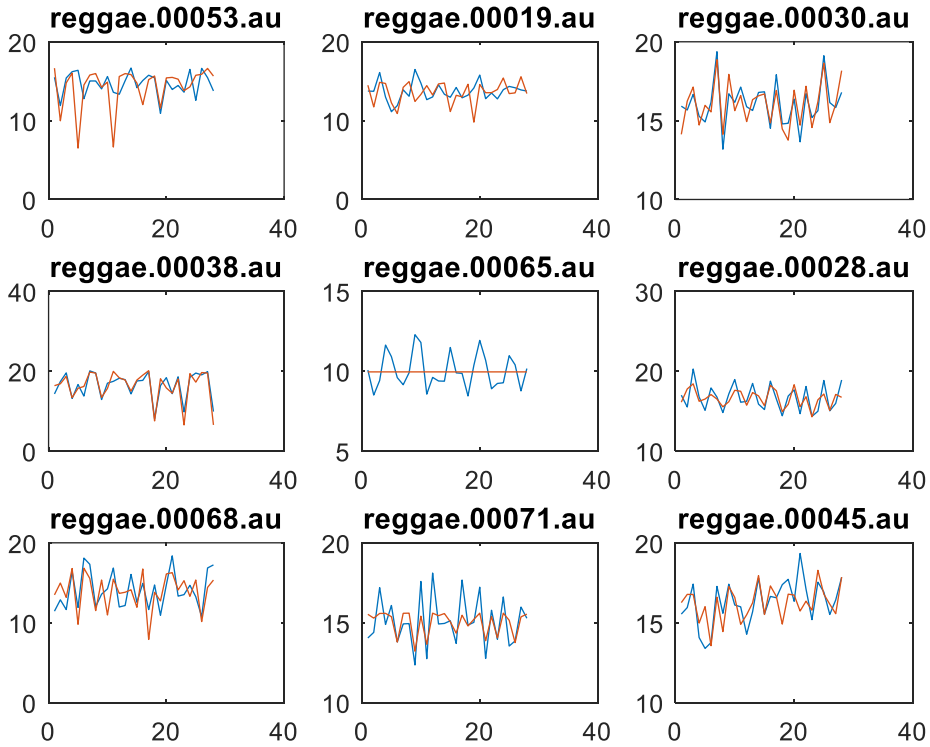

Figure 28: Comparison of the crest factors values in 9 reggae audio tracks

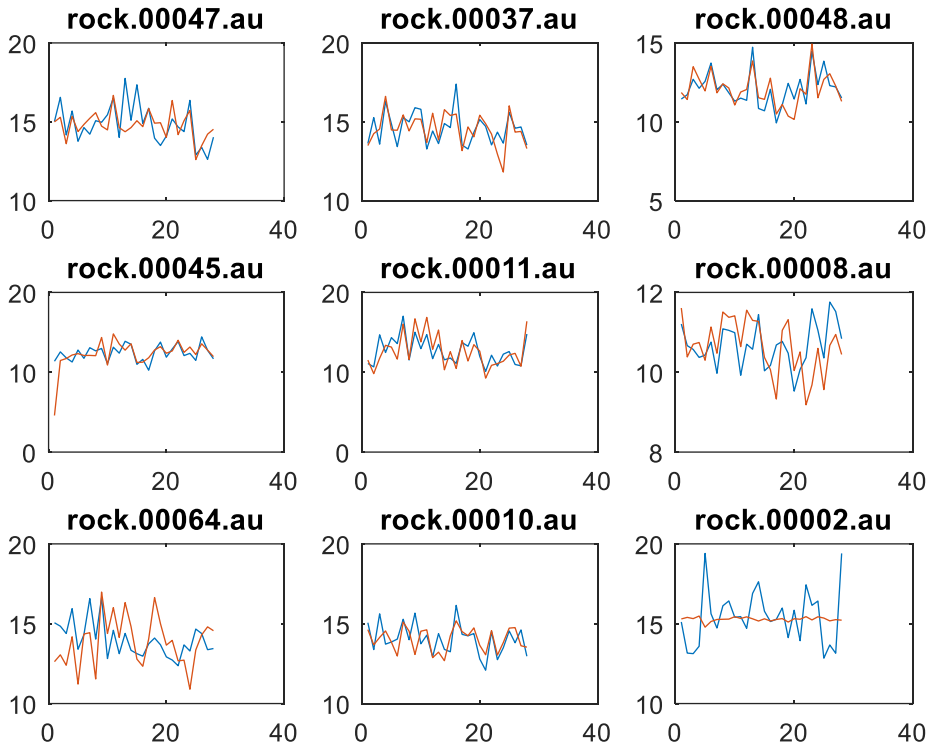

Figure 29: Comparison of the crest factors values in 9 rock audio tracks

Appendix D: Bark scale spectrogram results

D.1 Examples of bark scale spectrograms

Note: It was not possible to show the X and Y axis of the graphs because of the space of the window size.

· X axes: time (s)

· Y axes: bark scale bands

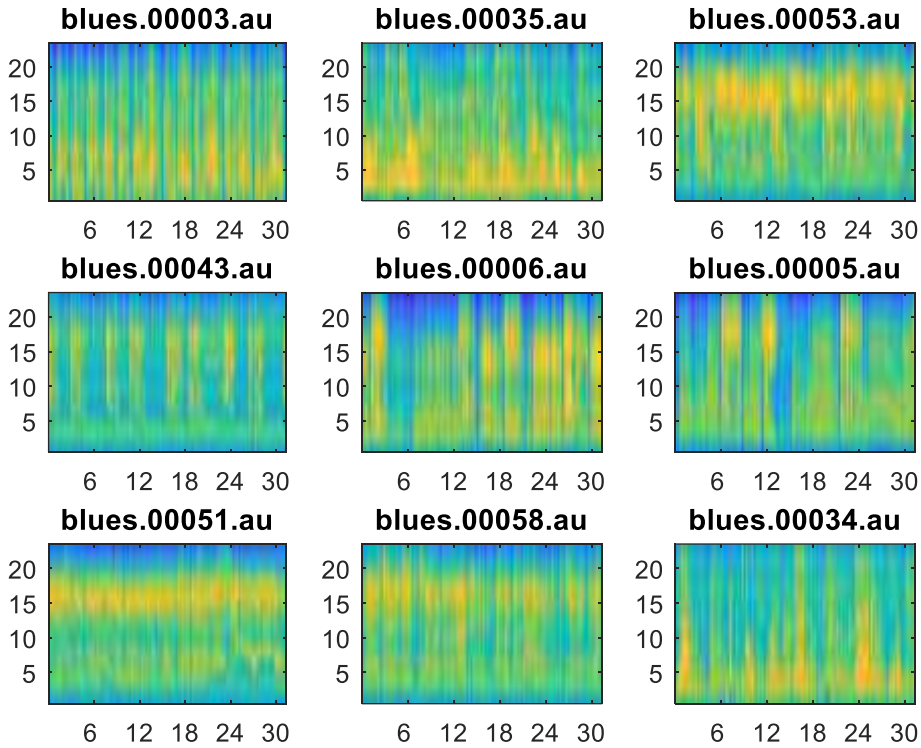

Figure 30: Bark scale spectrograms of 9 blues audio tracks

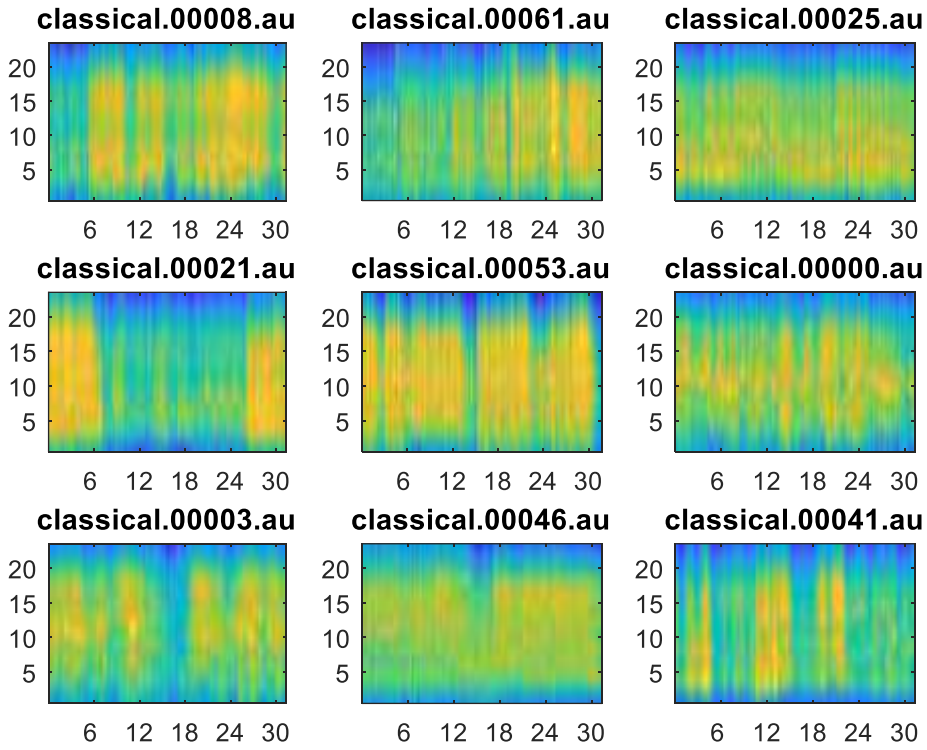

Figure 31: Bark scale spectrograms of 9 classical audio tracks

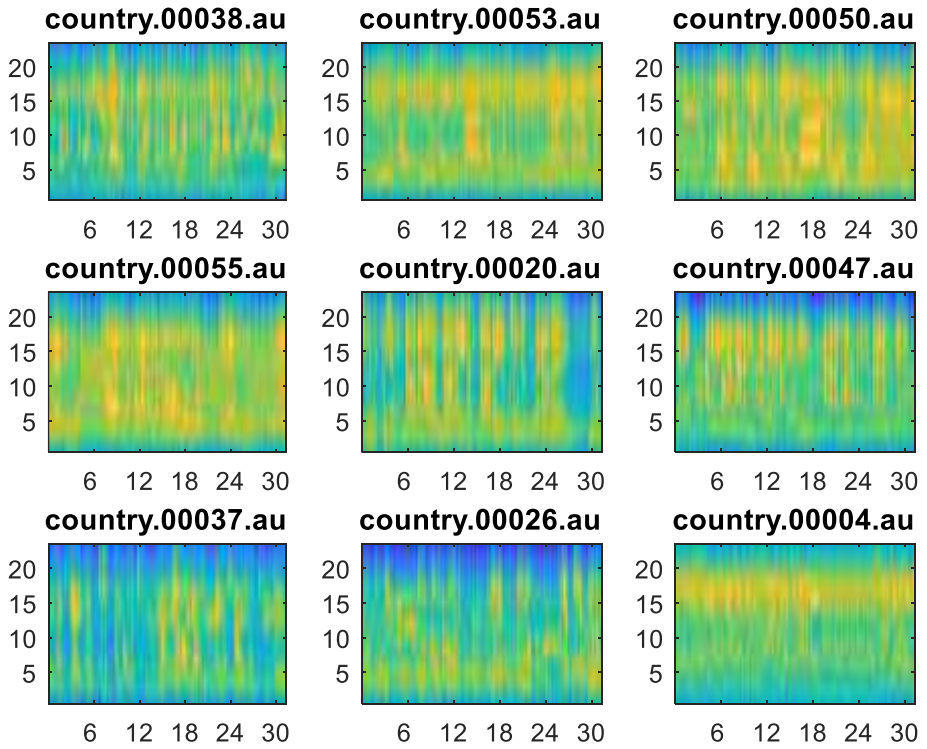

Figure 32: Bark scale spectrograms of 9 country audio tracks

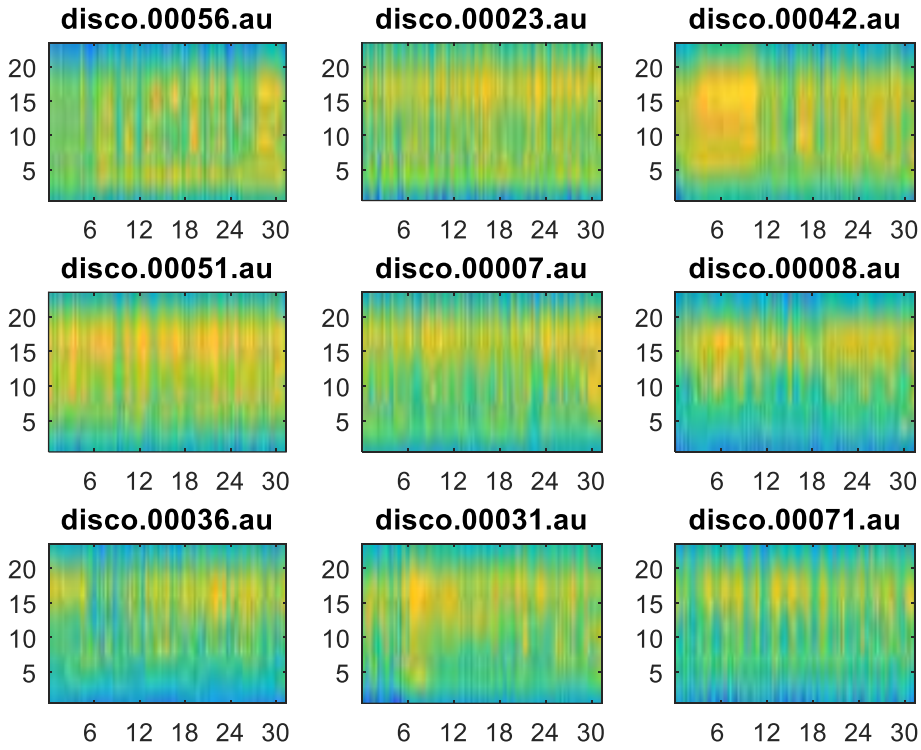

Figure 33: Bark scale spectrograms of 9 disco audio tracks

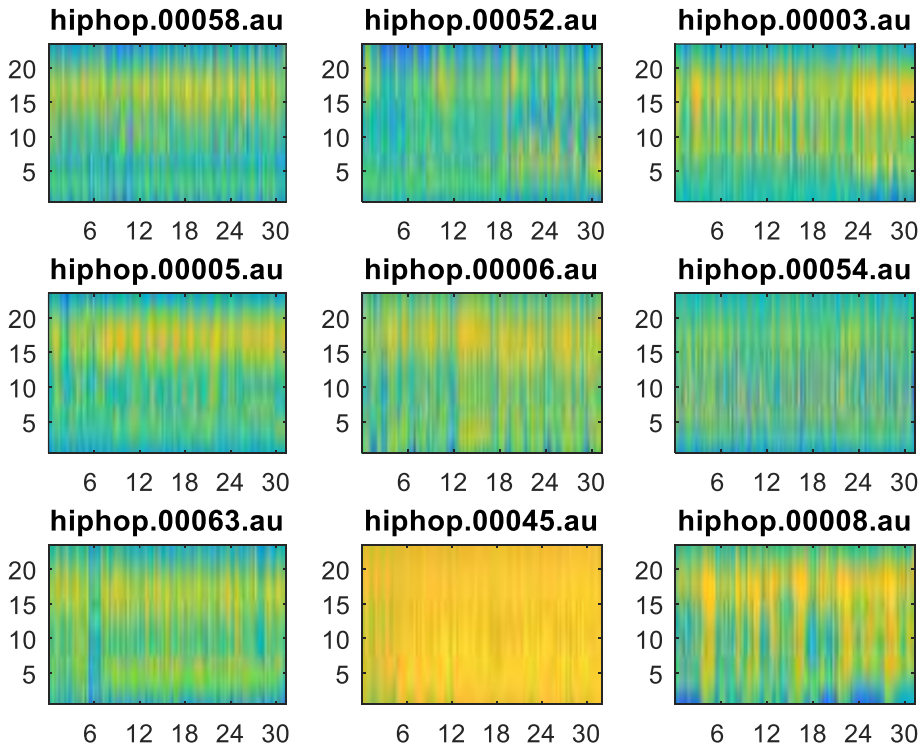

Figure 34: Bark scale spectrograms of 9 hip hop audio tracks

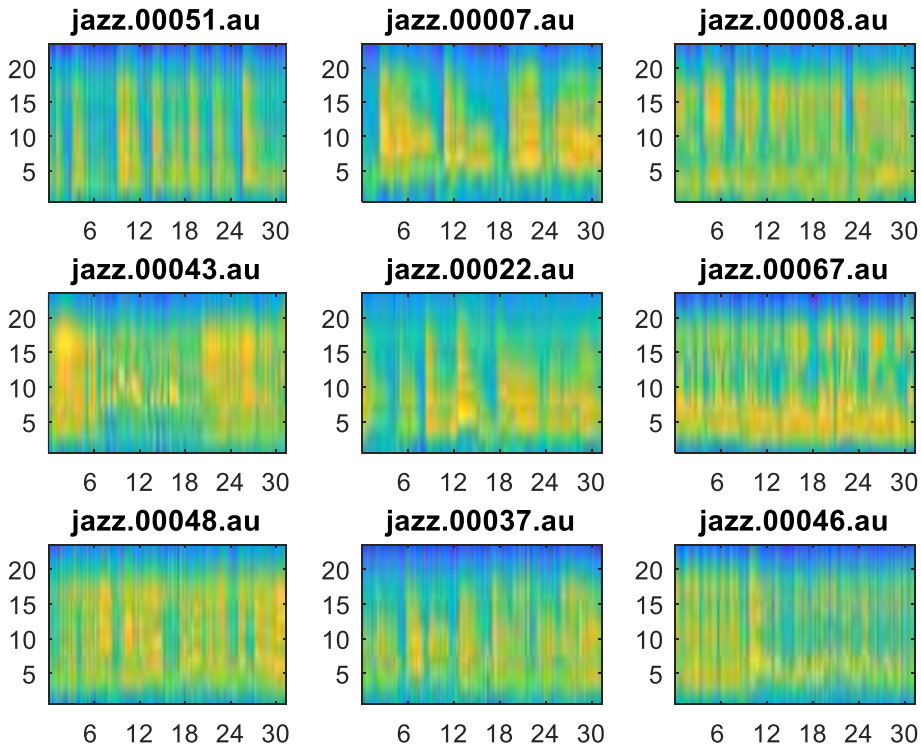

Figure 35: Bark scale spectrograms of 9 jazz audio tracks

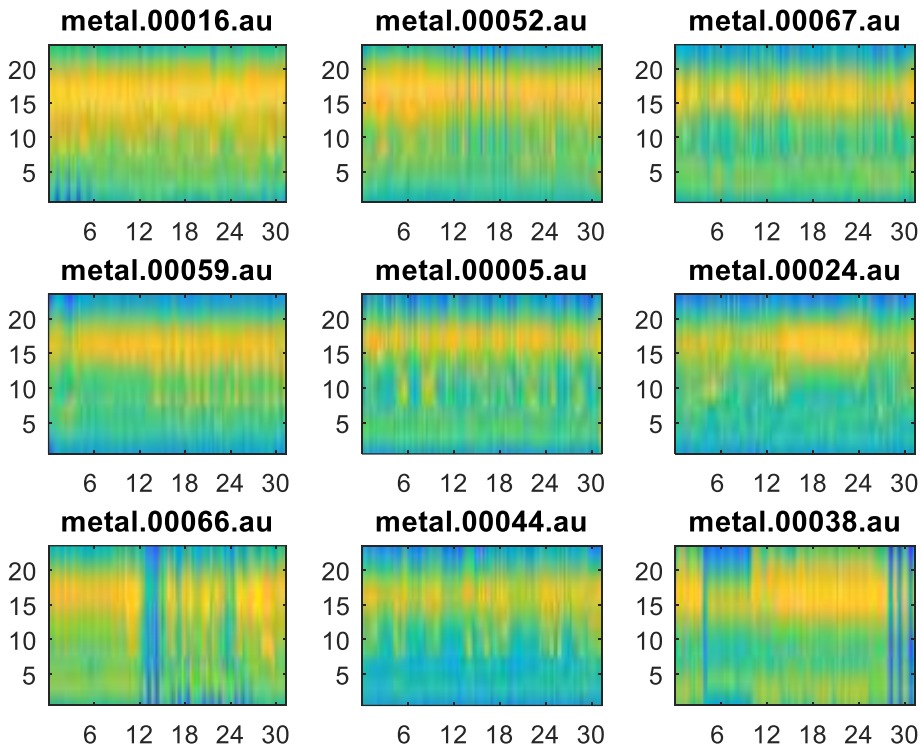

Figure 36: Bark scale spectrograms of 9 metal audio tracks

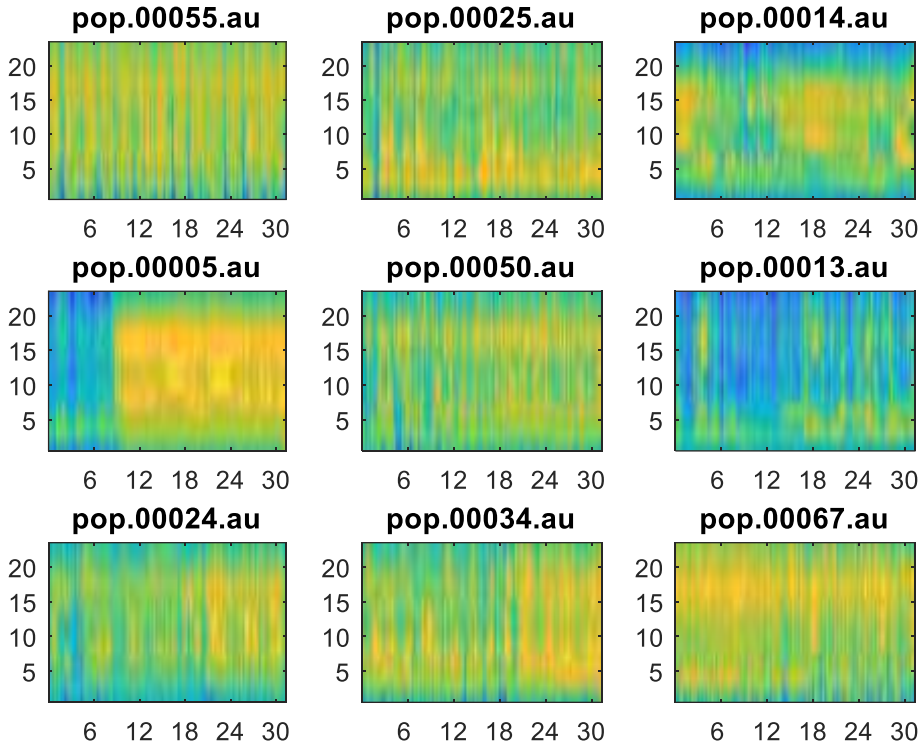

Figure 37: Bark scale spectrograms of 9 pop audio tracks

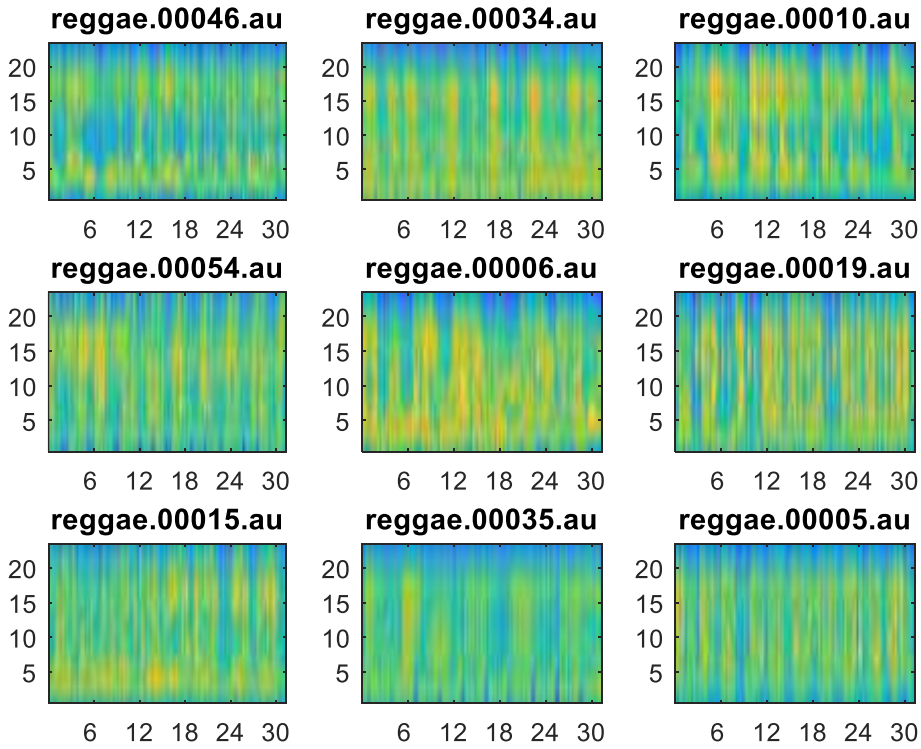

Figure 38: Bark scale spectrograms of 9 reggae audio tracks

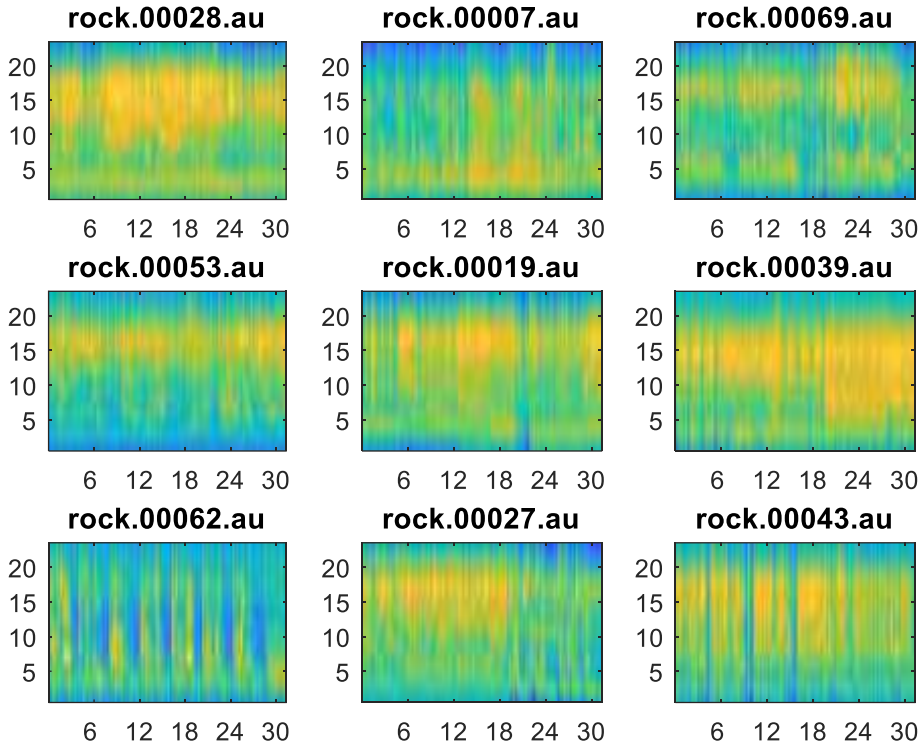

Figure 39: Bark scale spectrograms of 9 rock audio tracks

D.2 Examples of 1st, 2nd and 3rd bark bands with the highest energy

Note: It was not possible to show the X and Y axis of the graphs because of the space of the window size.

· X axes: number of audio tracks

· Y axes: bark scale band

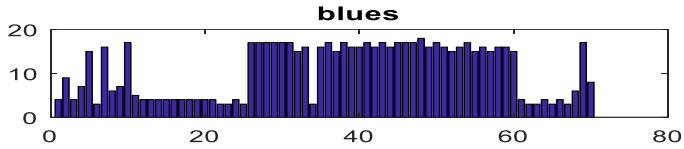

Figure 40: Bark band with the highest energy from the 70 blues audio tracks

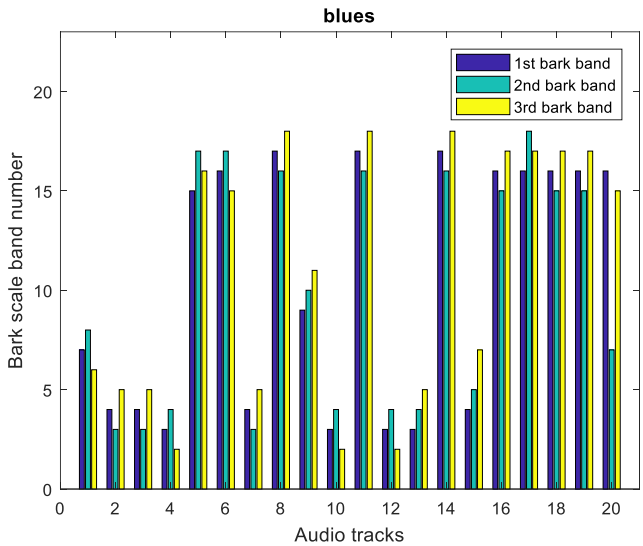

Figure 41:1st, 2nd and 3rd bark bands with the highest energy from 20 blues audio tracks

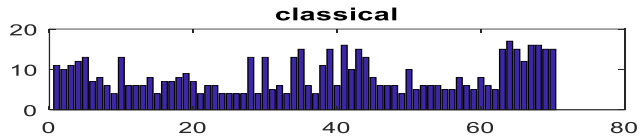

Figure 42: Bark band with the highest energy from the 70 classical audio tracks

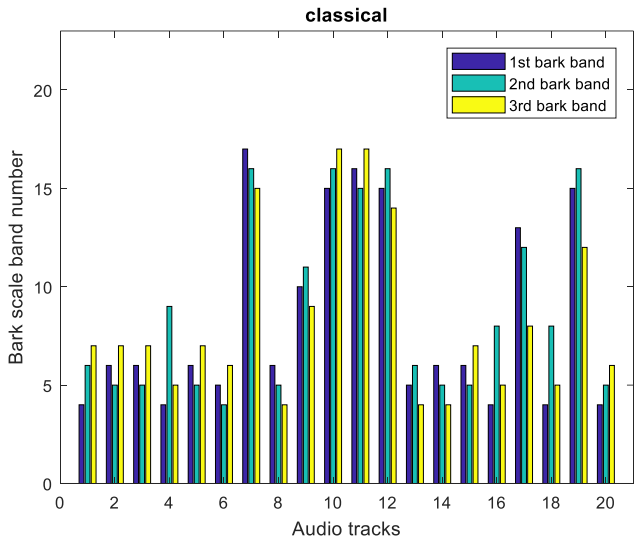

Figure 43:1st, 2nd and 3rd bark bands with the highest energy from 20 classical audio tracks

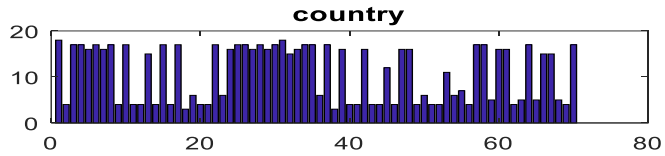

Figure 44: Bark band with the highest energy from the 70 country audio tracks

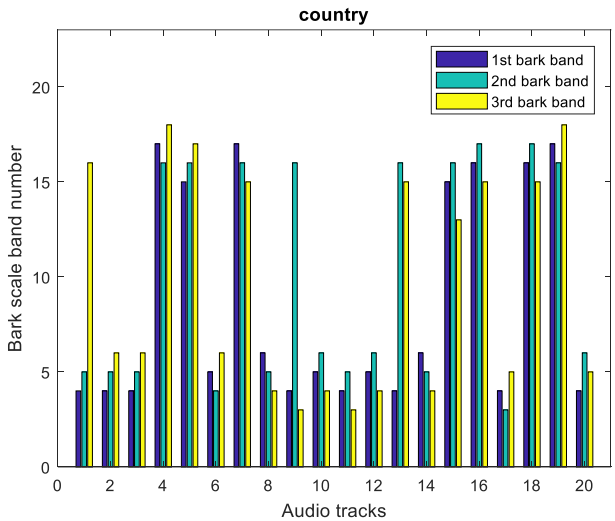

Figure 45: 1st, 2nd and 3rd bark bands with the highest energy from 20 country audio tracks

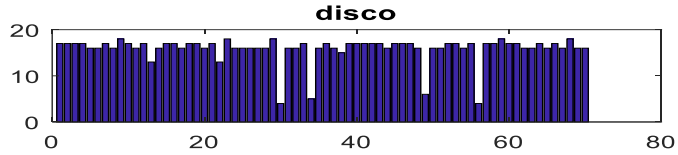

Figure 46: Bark band with the highest energy from the 70 disco audio tracks

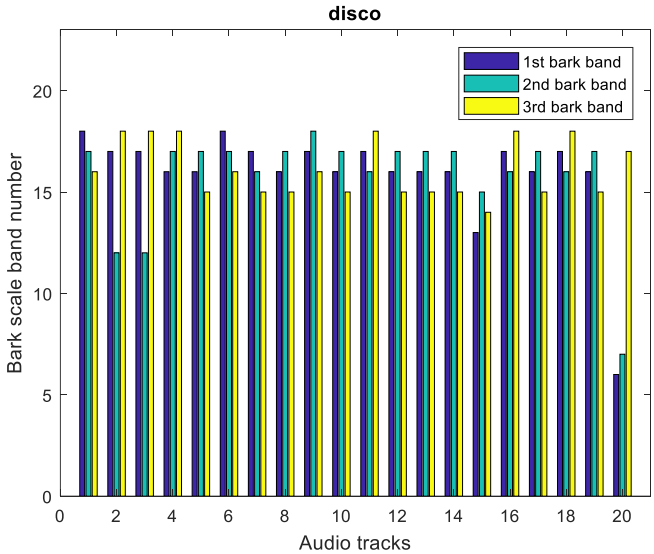

Figure 47: 1st, 2nd and 3rd bark bands with the highest energy from 20 disco audio tracks

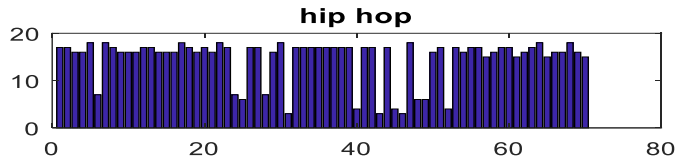

Figure 48: Bark band with the highest energy from the 70 hip hop audio tracks

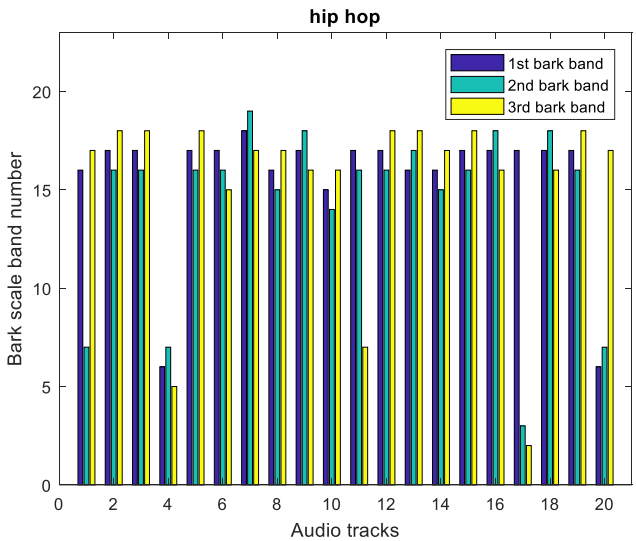

Figure 49: 1st, 2nd and 3rd bark bands with the highest energy from 20 hip hop audio tracks

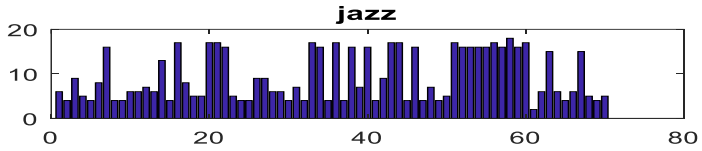

Figure 50: Bark band with the highest energy from the 70 jazz audio tracks

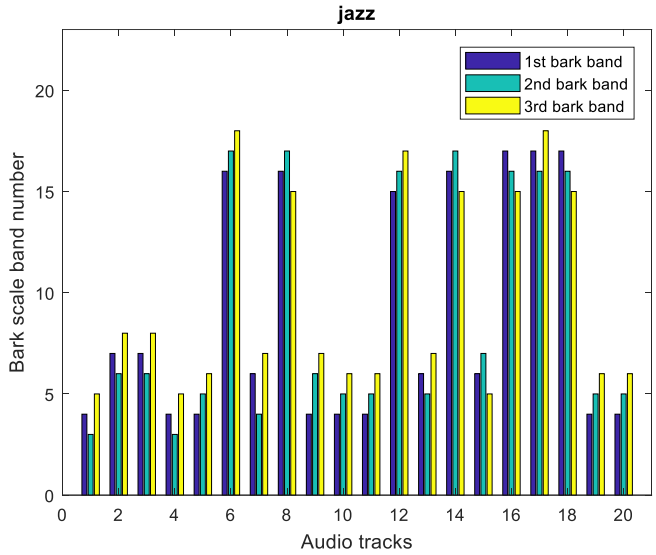

Figure 51: 1st, 2nd and 3rd bark bands with the highest energy from 20 jazz audio tracks

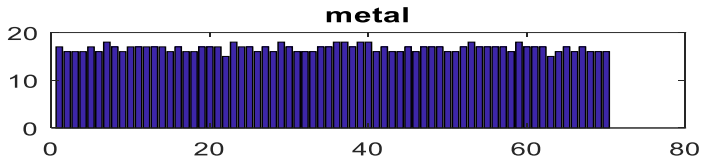

Figure 52: Bark band with the highest energy from the 70 metal audio tracks

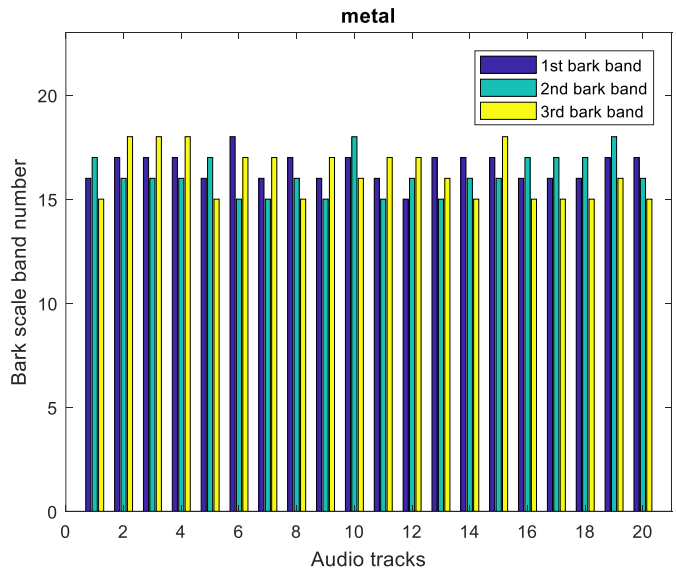

Figure 53: 1st, 2nd and 3rd bark bands with the highest energy from 20 metal audio tracks

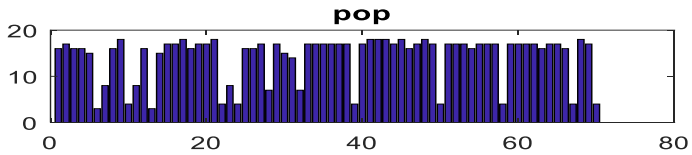

Figure 54: Bark band with the highest energy from the 70 pop audio tracks

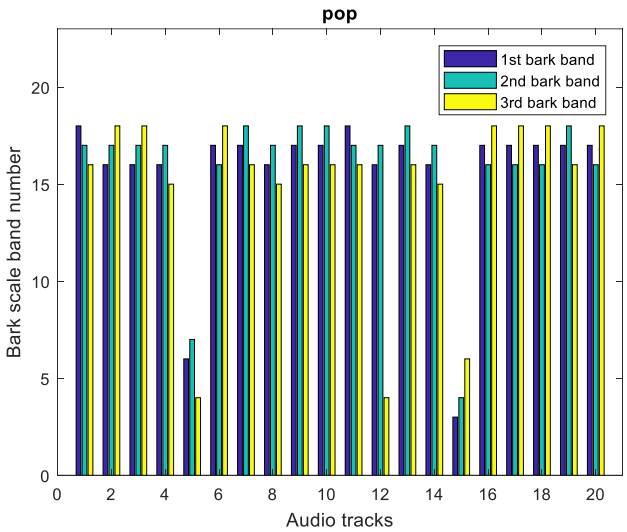

Figure 55: 1st, 2nd and 3rd bark bands with the highest energy from 20 pop audio tracks

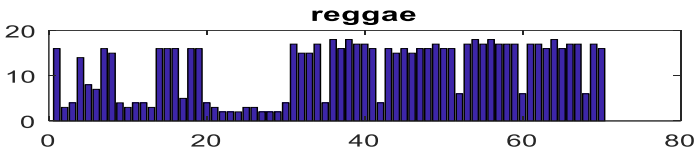

Figure 56: Bark band with the highest energy from the 70 reggae audio tracks

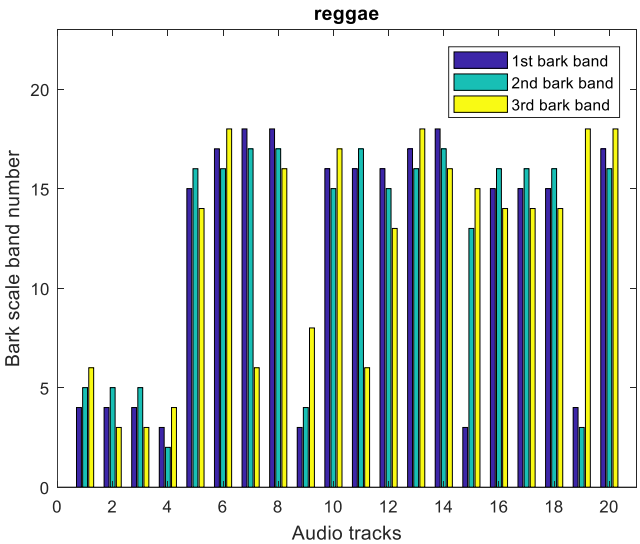

Figure 57: 1st, 2nd and 3rd bark bands with the highest energy from 20 reggae audio tracks

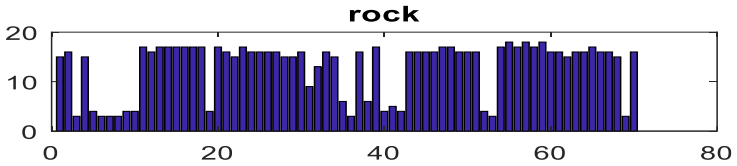

Figure 58: Bark band with the highest energy from the 70 rock audio tracks

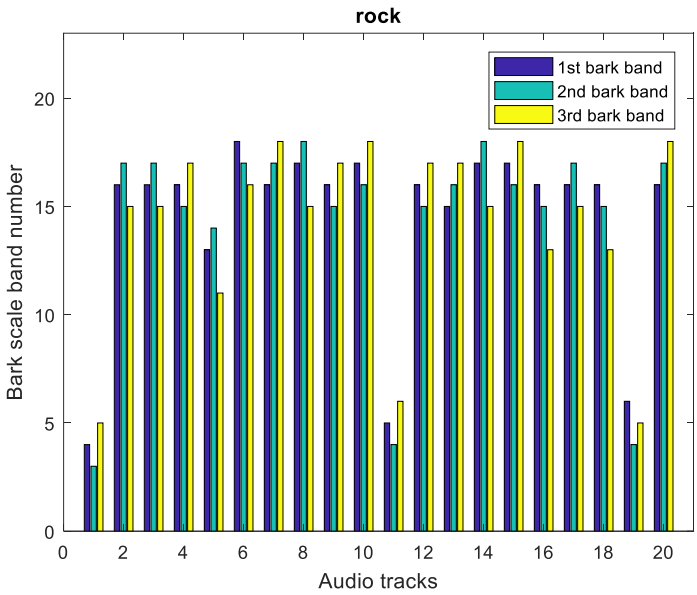

Figure 59: 1st, 2nd and 3rd bark bands with the highest energy from 20 rock audio tracks

D.3 Average power (dB) of each bark band

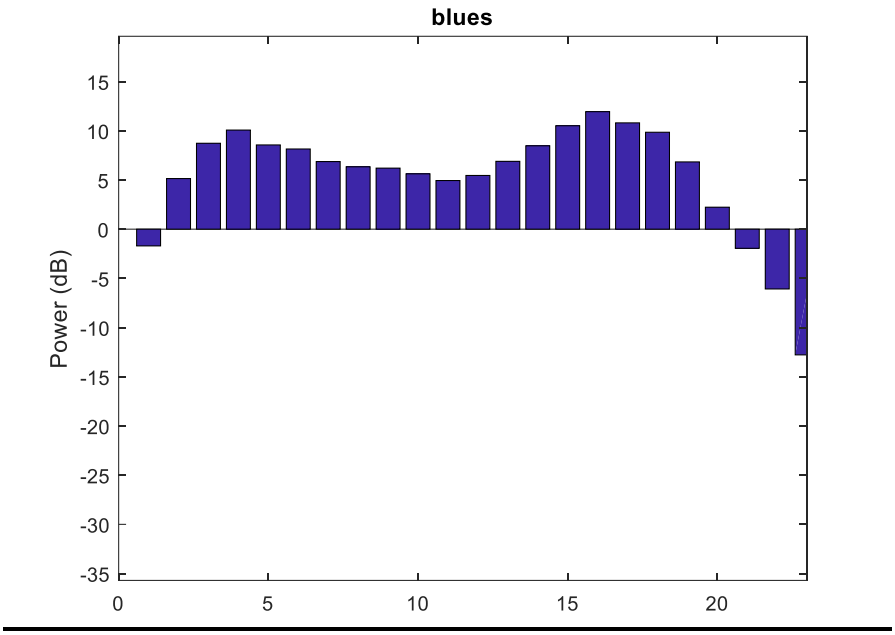

Figure 60: Average power of the 23 bark scale bands

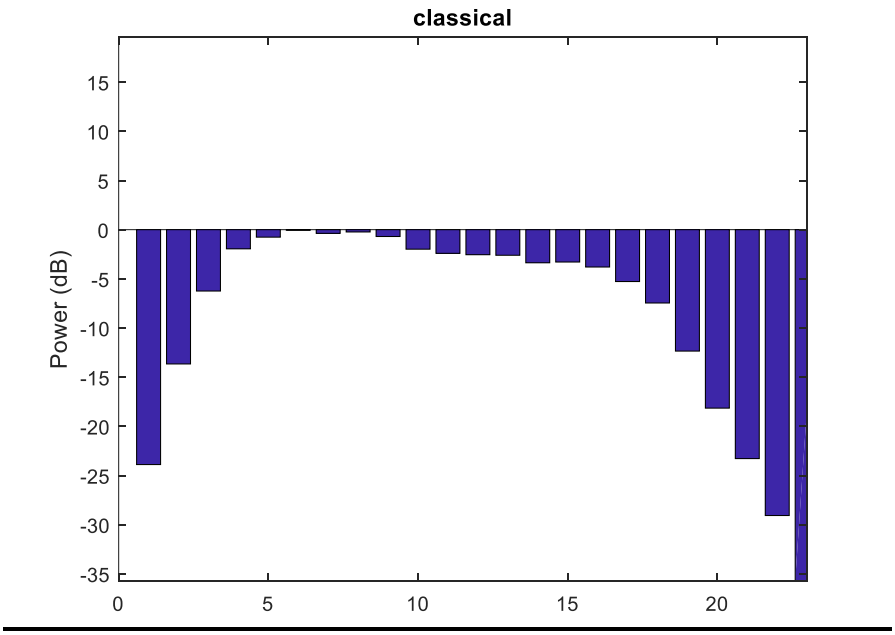

Figure 61: Average power of the 23 bark scale bands

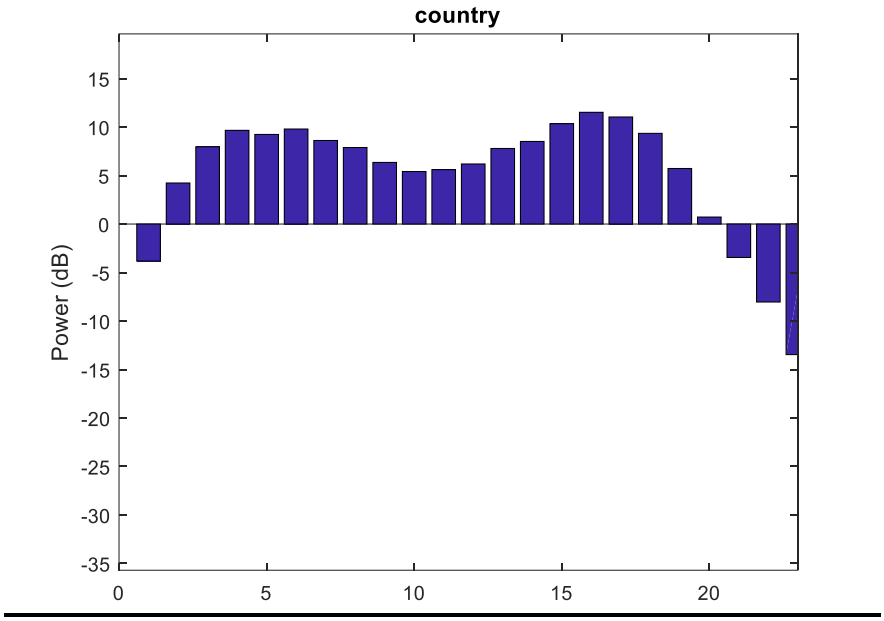

Figure 62: Average power of the 23 bark scale bands

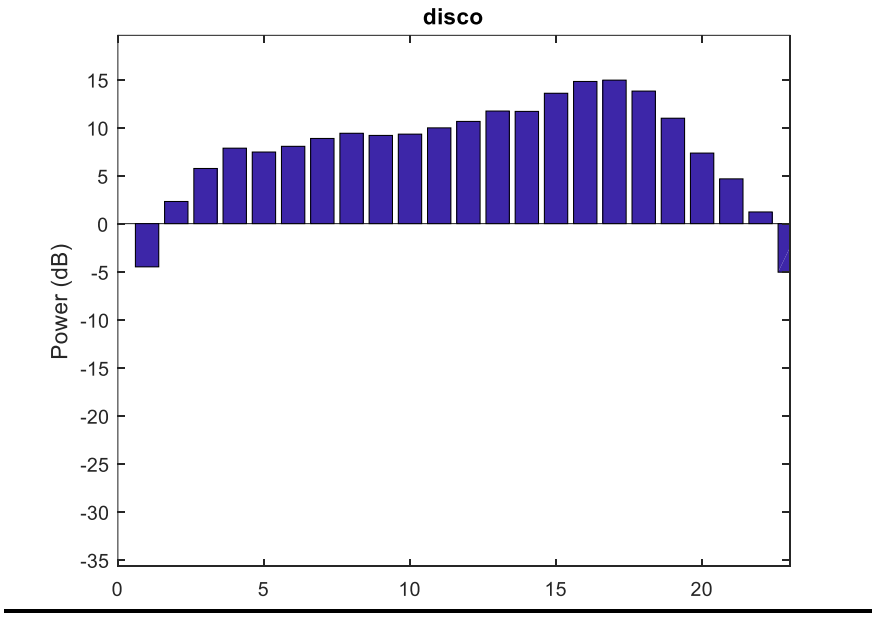

Figure 63: Average power of the 23 bark scale bands

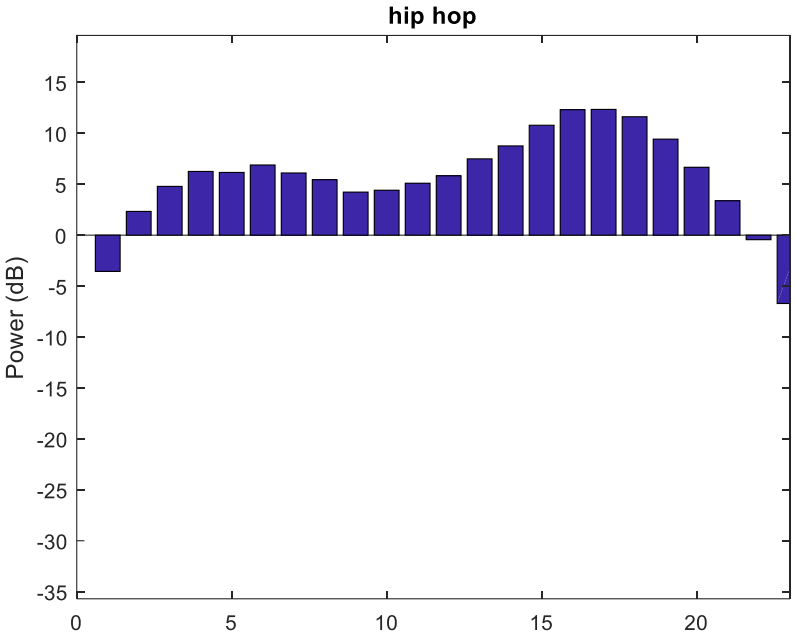

Figure 64: Average power of the 23 bark scale bands

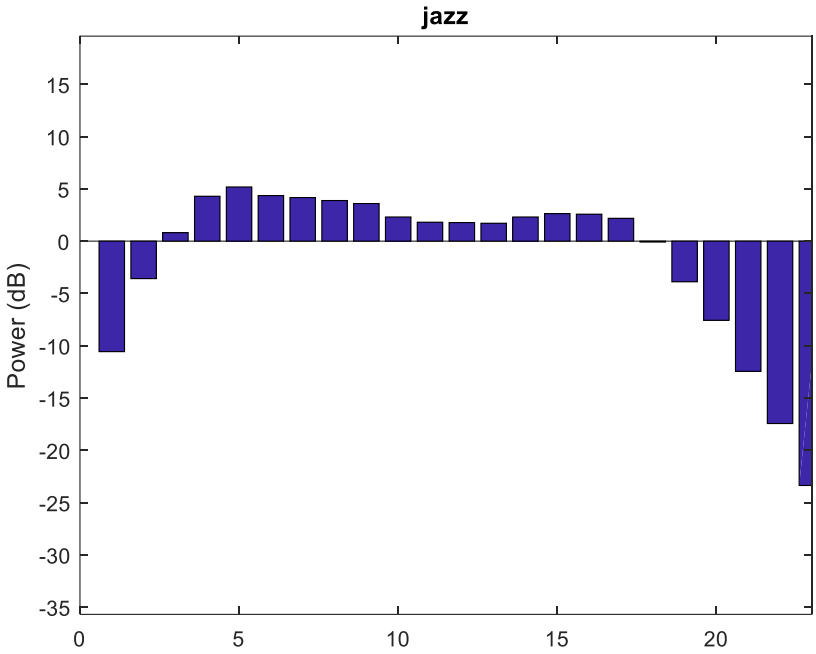

Figure 65: Average power of the 23 bark scale bands

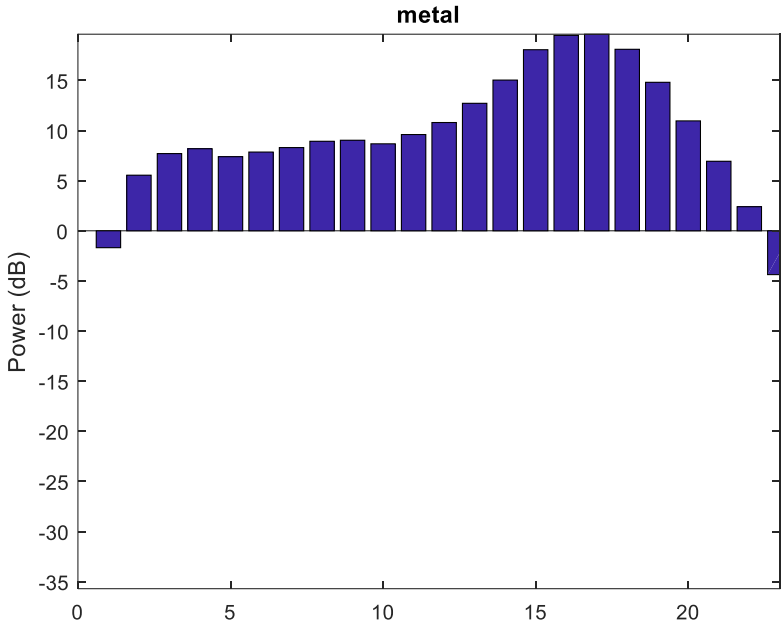

Figure 66: Average power of the 23 bark scale bands

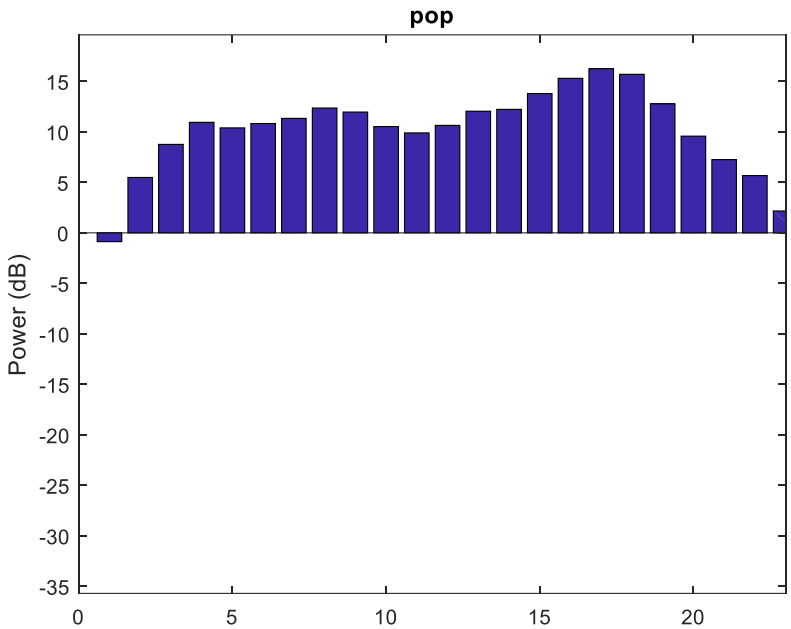

Figure 67: Average power of the 23 bark scale bands

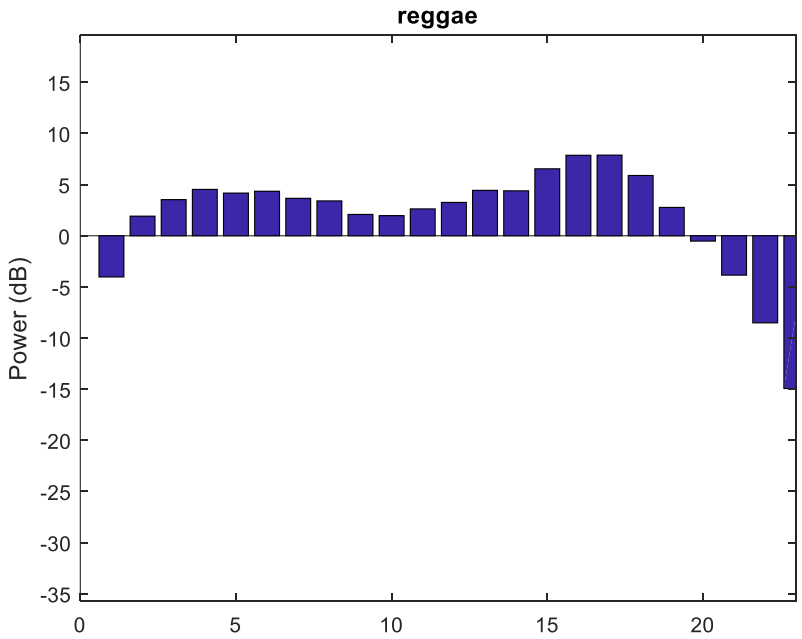

Figure 68: Average power of the 23 bark scale bands

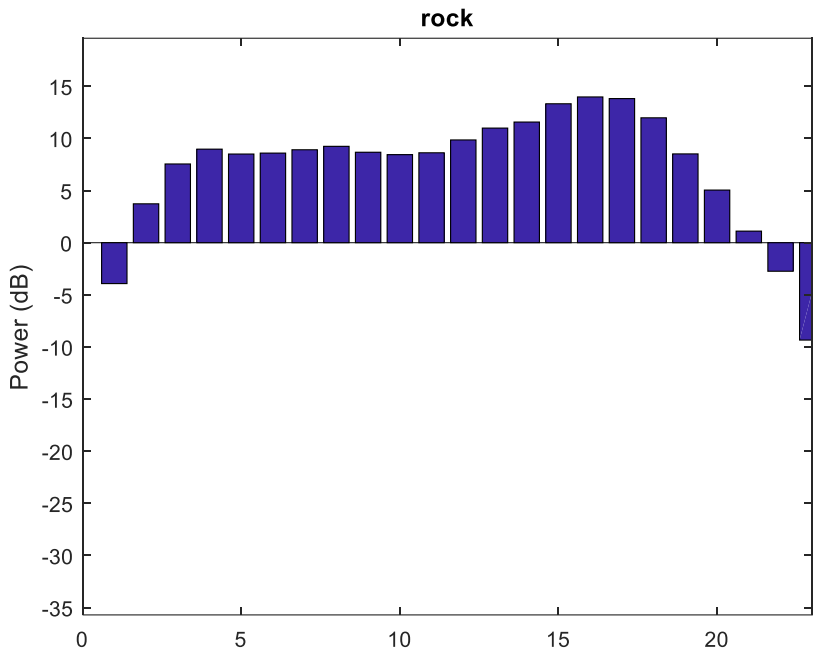

Figure 69: Average power of the 23 bark scale bands

D.4 Power of the highest energy bark band

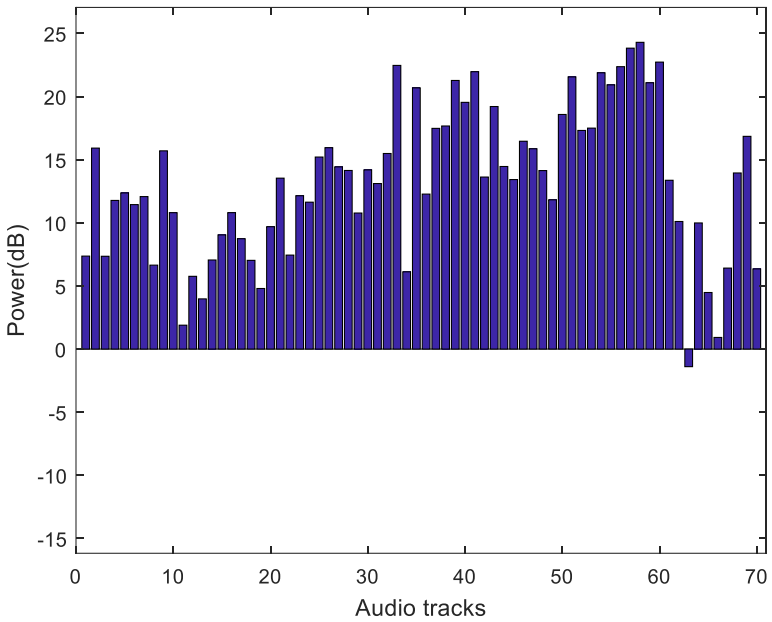

Figure 70: Power of the highest energy band of the 70 blues audio tracks

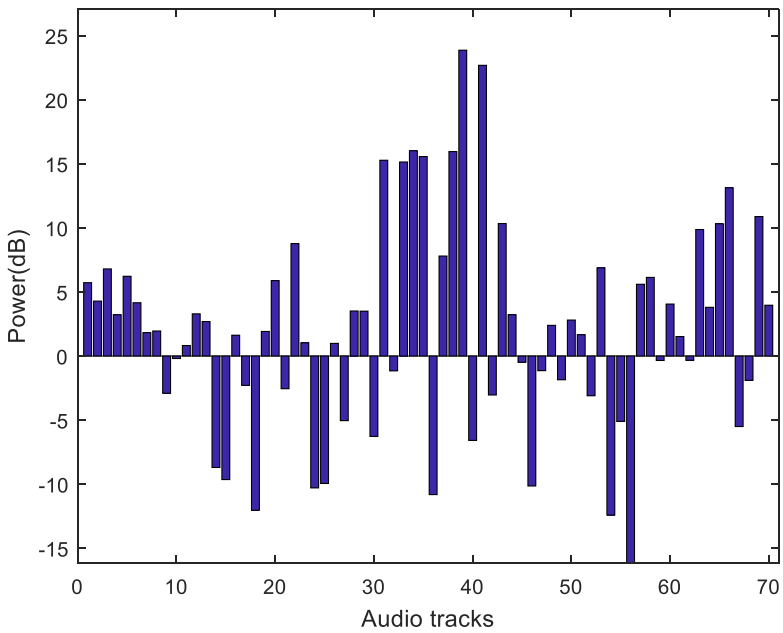

Figure 71: Power of the highest energy band of the 70 classical audio tracks

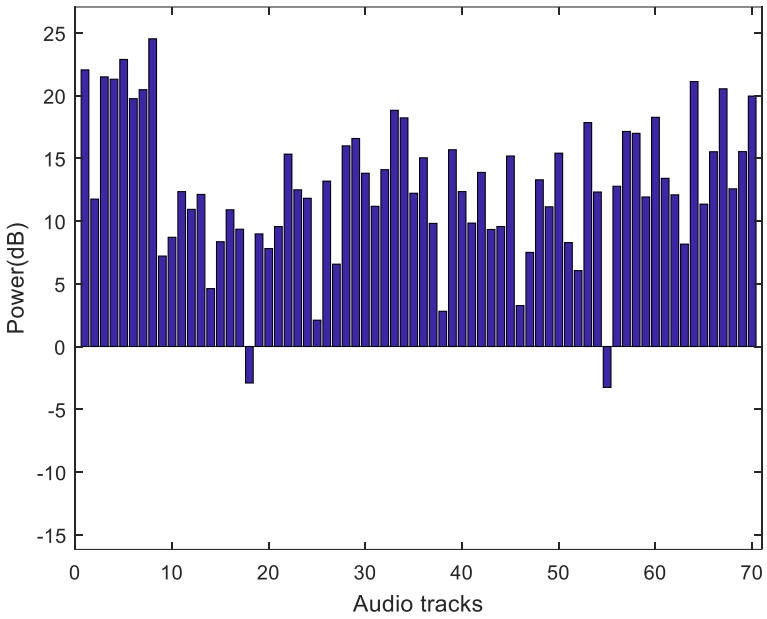

Figure 72: Power of the highest energy band of the 70 country audio tracks

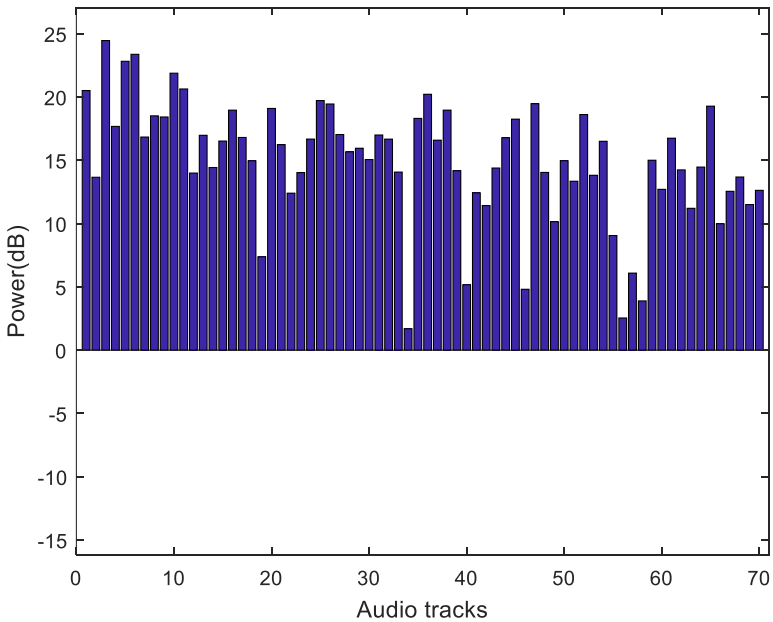

Figure 73: Power of the highest energy band of the 70 disco audio tracks

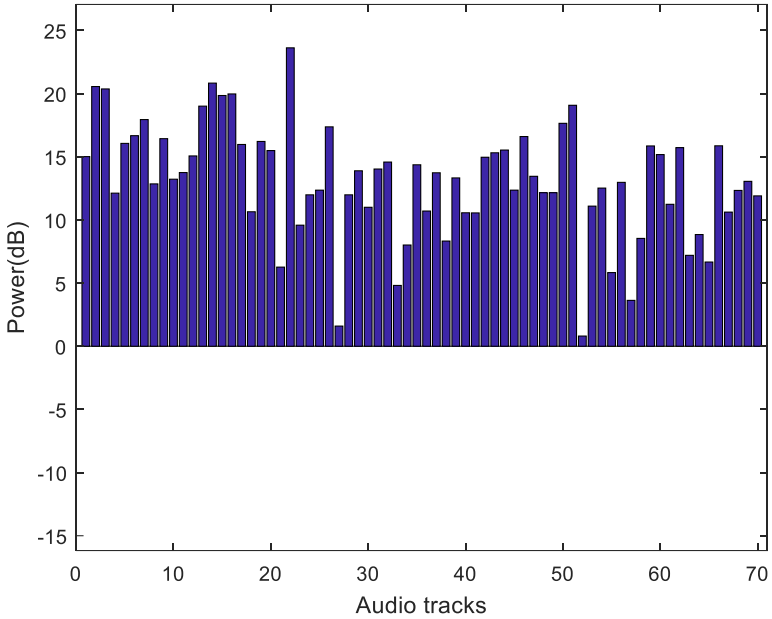

Figure 74: Power of the highest energy band of the 70 hip hop audio tracks

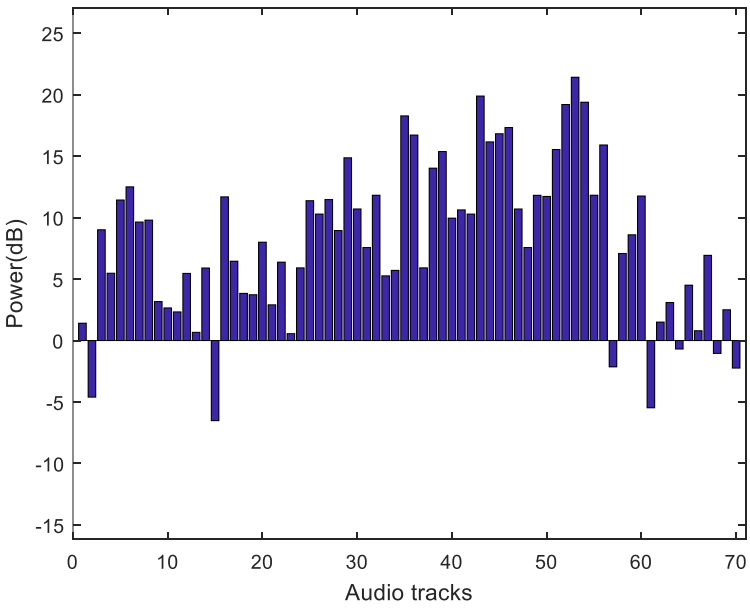

Figure 75: Power of the highest energy band of the 70 jazz audio tracks

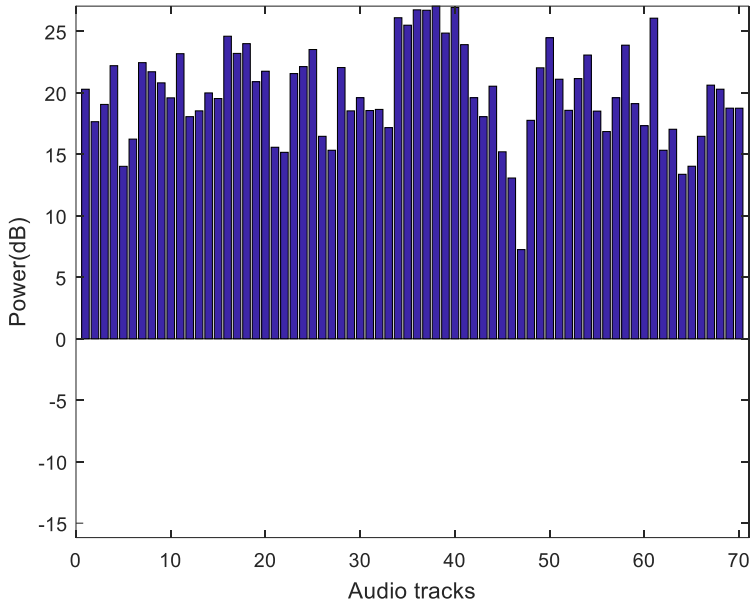

Figure 76: Power of the highest energy band of the 70 metal audio tracks

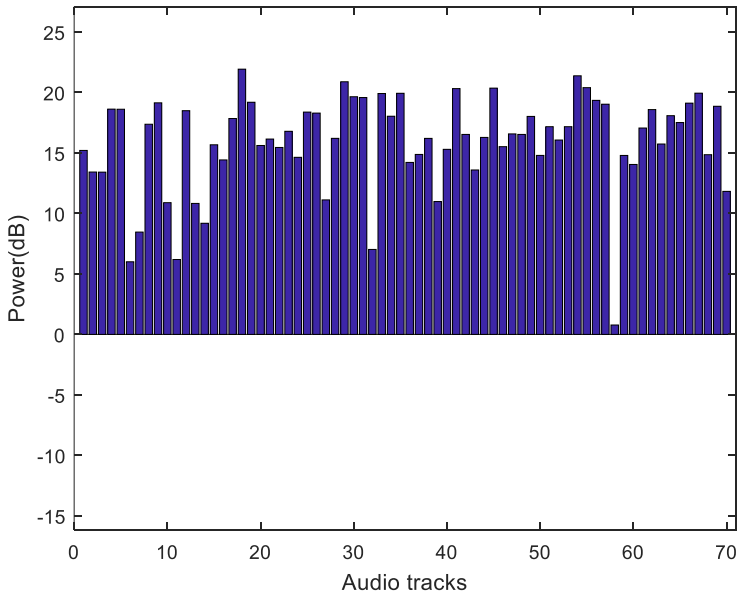

Figure 77: Power of the highest energy band of the 70 pop audio tracks

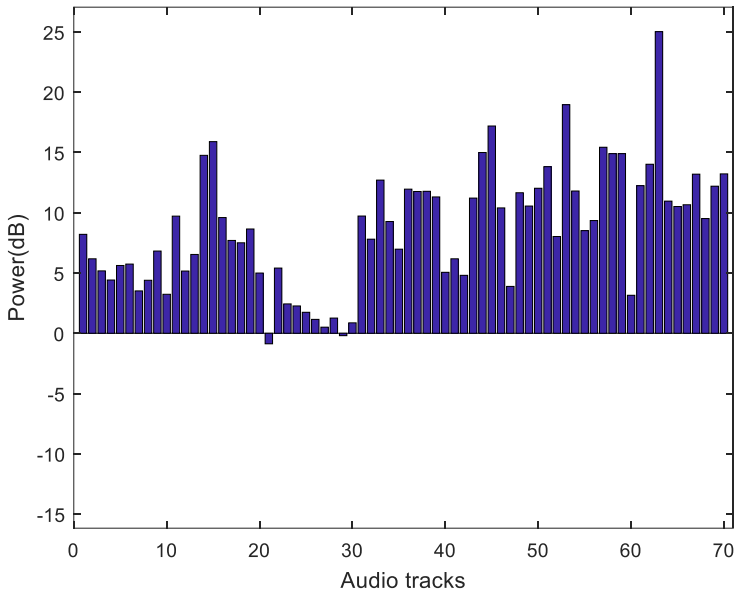

Figure 78: Power of the highest energy band of the 70 reggae audio tracks

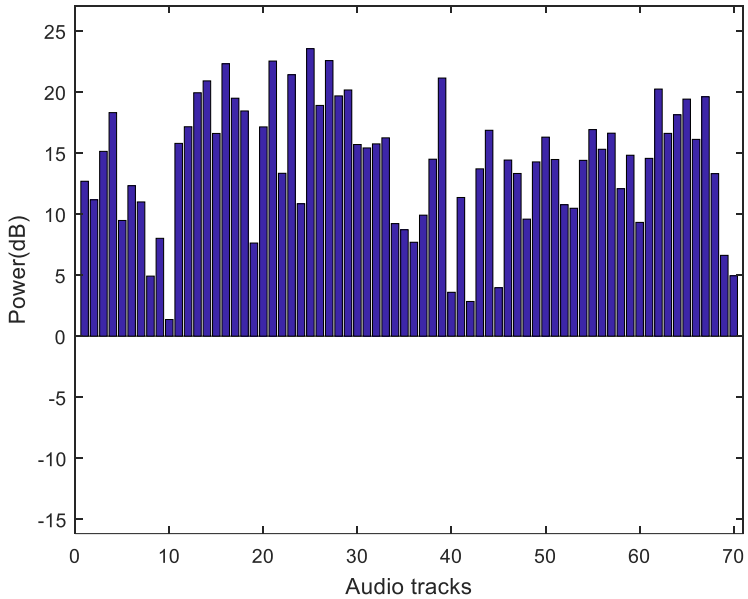

Figure 79: Power of the highest energy band of the 70 rock audio tracks

D.5 Standard deviation of the bark scale bands across the 70 audio tracks

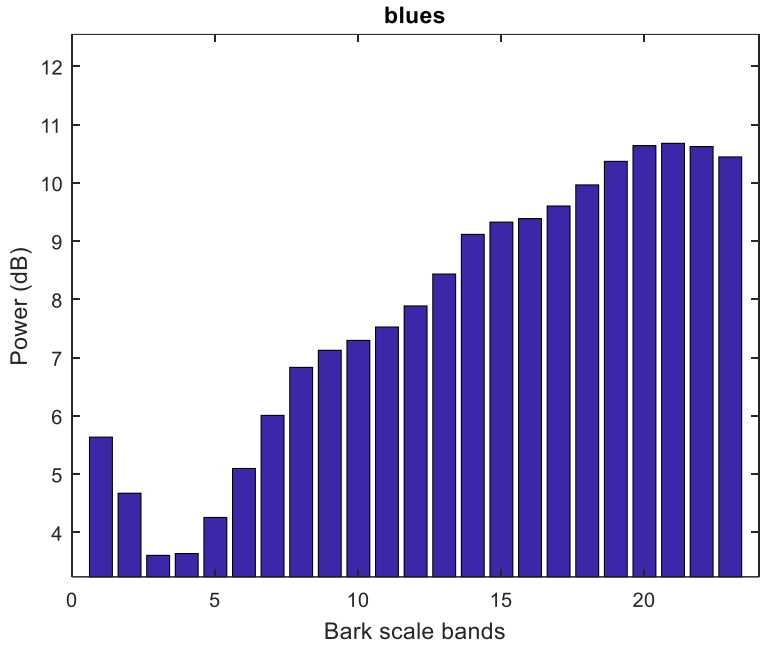

Figure 80: Average for blues audio tracks of the standard deviation of the bark bands power

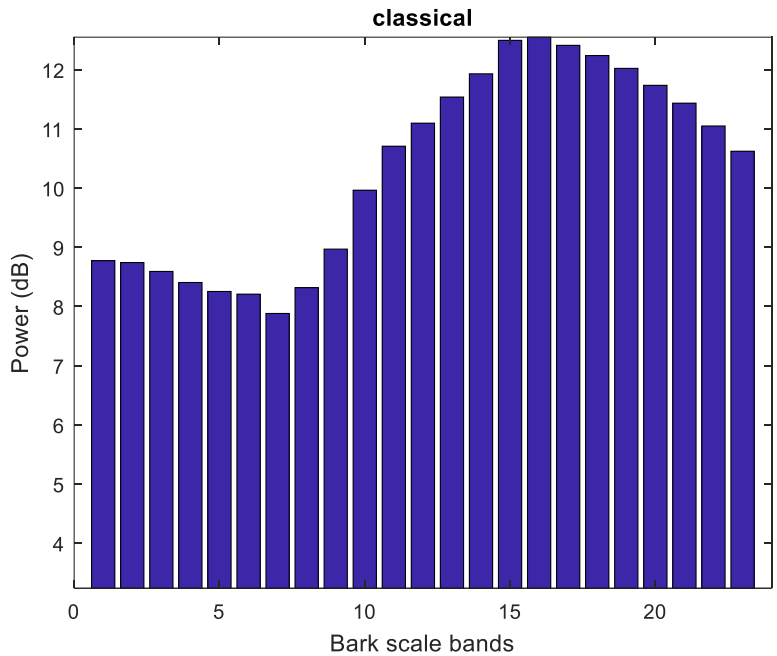

Figure 81: Average for classical audio tracks of the standard deviation of the bark bands power

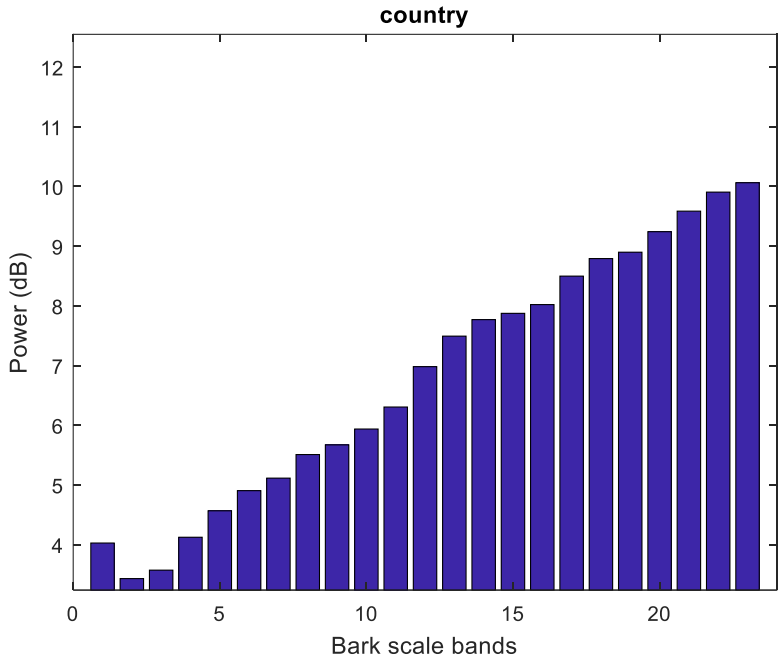

Figure 82: Average for country audio tracks of the standard deviation of the bark bands power

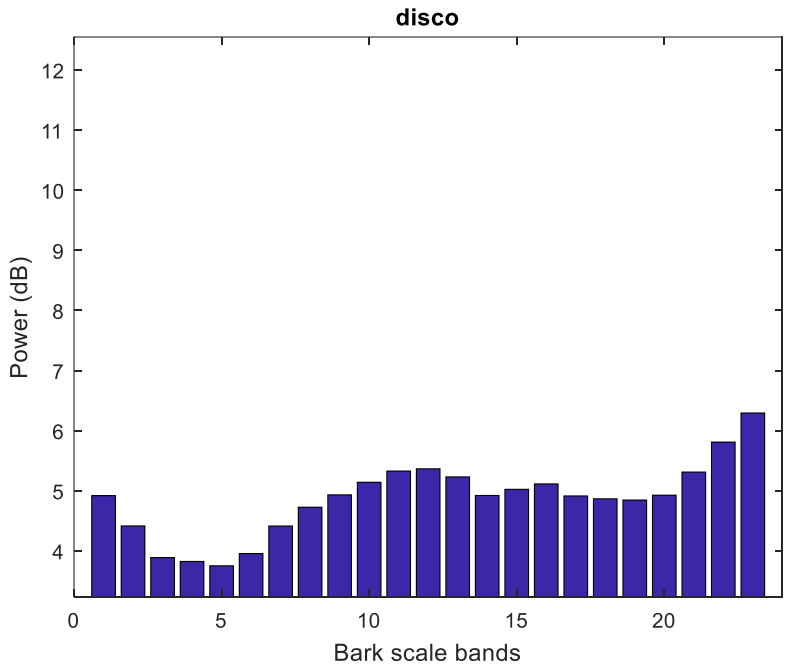

Figure 83: Average for disco audio tracks of the standard deviation of the bark bands power

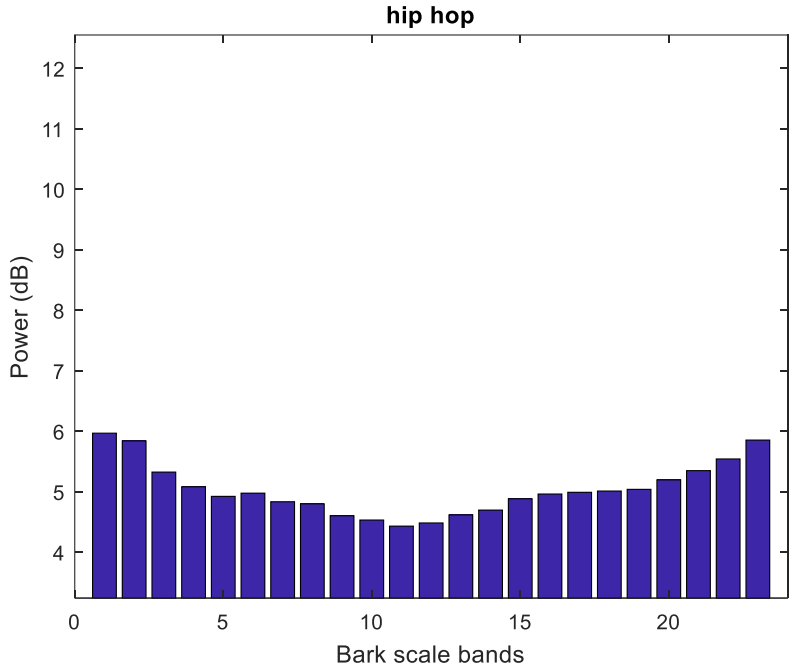

Figure 84: Average for hip hop audio tracks of the standard deviation of the bark bands power

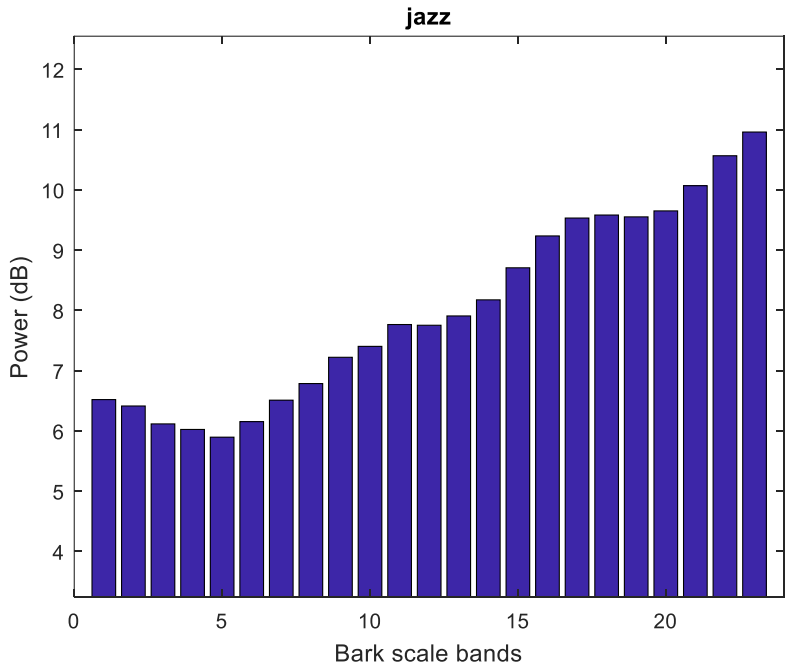

Figure 85: Average for jazz audio tracks of the standard deviation of the bark bands power

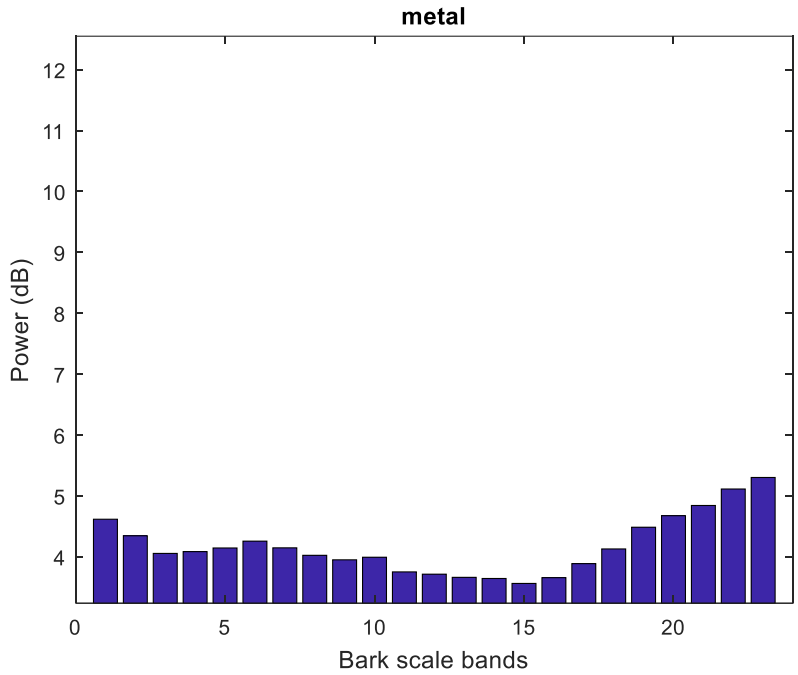

Figure 86: Average for metal audio tracks of the standard deviation of the bark bands power

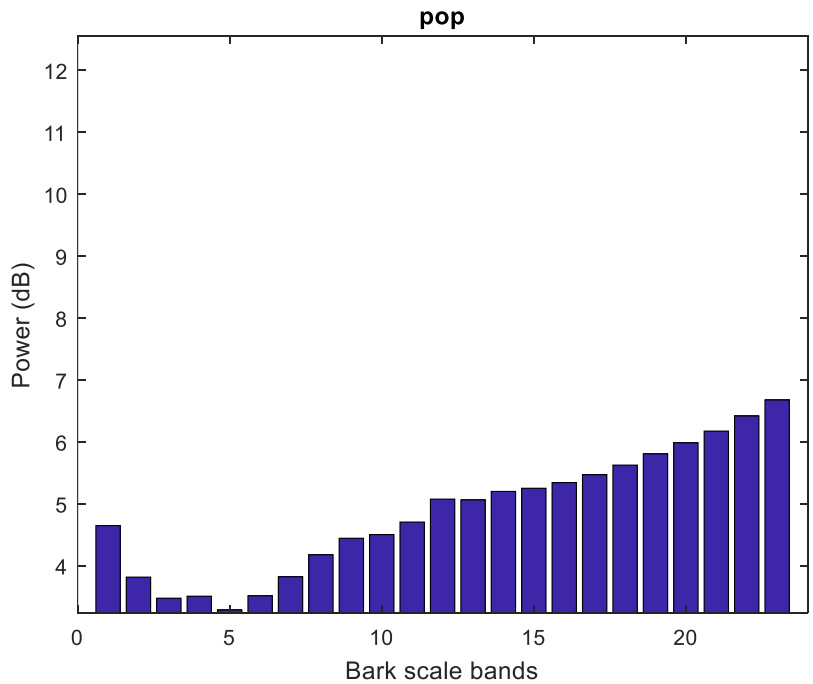

Figure 87: Average for pop audio tracks of the standard deviation of the bark bands power

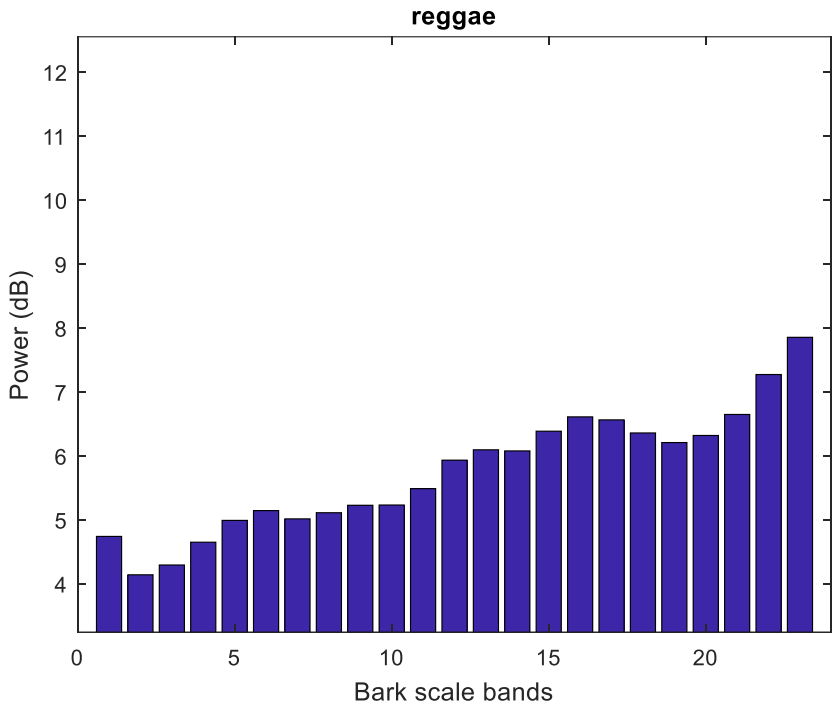

Figure 88: Average for reggae audio tracks of the standard deviation of the bark bands power

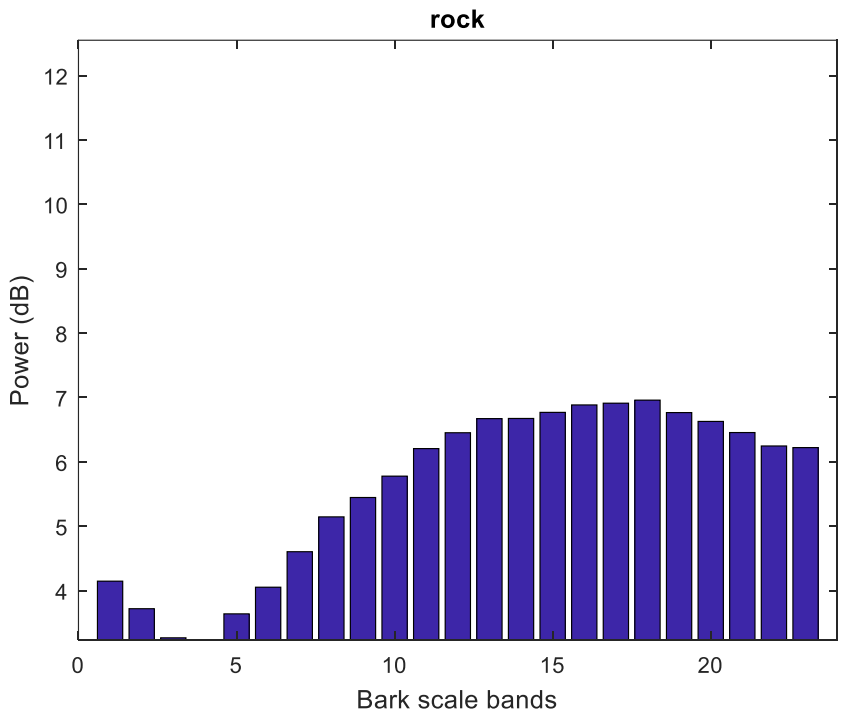

Figure 89: Average for rock audio tracks of the standard deviation of the bark bands power

Appendix E: Confusion matrix and evaluation mesures used

E.1 Confusion matrix

	Blues	*Classical*	*Country*	*Disco*	*Hip hop*	*Jazz*	*Metal*	*Pop*	*Reggae*	*Rock*
Blues	13	0	6	1	1	3	1	0	1	4
Classical	1	28	0	0	0	0	0	0	0	1
Country	1	0	24	0	0	1	0	1	0	3
Disco	0	0	3	17	1	1	1	1	2	4
Hip hop	2	0	0	1	19	1	1	4	1	1
Jazz	1	2	2	0	0	25	0	0	0	0
Metal	0	0	1	0	0	0	29	0	0	0
Pop	2	0	2	1	1	0	0	18	2	4
Reggae	2	2	3	2	1	0	0	2	17	1
Rock	1	3	5	1	0	2	5	4	1	8

Table 6: Confusion matrix of configuration 1

	Blues	*Classical*	*Country*	*Disco*	*Hip hop*	*Jazz*	*Metal*	*Pop*	*Reggae*	*Rock*
Blues	14	0	5	1	1	3	1	0	1	4
Classical	2	27	0	0	0	0	0	0	0	1
Country	1	0	23	0	0	1	0	1	0	4
Disco	0	0	3	17	1	1	1	0	2	5
Hip hop	3	1	0	1	18	1	1	3	1	1
Jazz	1	2	1	0	0	26	0	0	0	0
Metal	0	0	1	0	0	0	29	0	0	0
Pop	2	0	2	1	1	0	0	18	2	4
Reggae	2	1	3	2	1	0	0	2	18	1
Rock	1	2	6	1	0	2	5	3	1	9

Table 7: Confusion matrix of configuration 2

E.2 Evaluation measures used

The measures used to evaluate the classifier are:

- Accuracy: description of systematic errors

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

- Precision: fraction of relevant instances among the retrieved instances

$$Precision = \frac{tp}{tp + fp}$$

- Recall: fraction of relevant instances that have been retrieved over total relevant instances

$$Recall = \frac{tp}{tp + fn}$$

- F-score: measure of a test's accuracy

$$F - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Where tp (true positive), tn (true negative), fp (false positive) and fn (false negative). This values are computed comparing the original results with the classified labels. A perfect classification is achieved when Accuracy=Precision=Recall=F-score=1.