Using Musical and Statistical Analysis of the Predominant Melody of the Voice to Create datasets from a Database of Popular Brazilian Hit Songs

André A. Bertoni and Rodrigo P. Lemos

Universidade Federal de Goiás, Brazil
engenheirobertoni@gmail.com - lemos@ufg.br

Abstract.

This work deals with the creation and optimization of a large set of features extracted from a database of 882 popular brazilian hit songs and non-hit songs, from 2014 to May 2019. From this database of songs, we created four datasets of musical features. The first comprises 3215 statistical features, while the second, third and fourth are completely new, as they were formed from the predominant melody of the Voice and previously there were no similar databases available for study. The second set of data represents the graph of the time-frequency spectrogram of the singer's voice during the first 90 seconds of each song. The third dataset results from a statistical analysis carried out on the predominant melody of the voice. The fourth is the most peculiar of all, as it results from the musical semantic analysis of the predominant melody of the voice, which allowed the construction of a table with the most frequent melodic sequences of each song. Our datasets use only Brazilian songs and focus their data on a limited and contemporary period. The idea behind these datasets is to encourage the study of Machine Learning techniques that require musical information. The extracted features can help develop new studies in Music and Computer Science in the future.

Categories and Subject Descriptors: H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Miscellaneous; H.3.3 [**Information Search and Retrieval**]: Retrieval models

Keywords: MUSIC INFORMATION RETRIEVAL, HIT SONGS, NON-HIT SONGS, BRAZILIAN MUSIC DATASETS, MUSICAL SEMANTIC INFORMATION DATASETS.

1. INTRODUCTION

First of all, it is important to mention that this article is an extension of the article published in DSW (Dataset Showcase Workshop - 2021), bringing several complementary information, as well as new databases.[1]. For several years now, we have seen a huge shift in the music entertainment industry. The fall of traditional formats, such as vinyl, CD and DVD, and the emergence of new formats, consumed exclusively by streaming, ended up changing forever the socioeconomic and cultural paradigms of how people listen to music. Digital platforms like Deezer[2], Spotify[3] and Itunes[4] now deliver thousands of songs to the palm of our hands through smartphones. To keep their subscriber base connected to their platforms for as long as possible, these companies have been learning that they must have a deep understanding of their products (songs) and also their users. It is from this idea that we begin to give more importance to studies on features extracted directly from songs. These techniques, commonly

[1]https://sol.sbc.org.br/index.php/dsw/article/view/17410
[2]Deezer - https://www.deezer.com/
[3]Spotify - https://www.spotify.com/
[4]Itunes - https://www.itunes.com/

known as MIR (Music Information Retrieval), aim to extract as much statistical information from music as possible through the use of specialized algorithms. Thus, for each song, MIR generates a set of musical information and stores it in a database that Streaming companies can later use to carry out statistical analyses in order to assess the performance of the songs and promote greater engagement among their users [Raieli 2013].

Below are some of the published works related to creating and manipulating data of this nature. Unfortunately, since the first known citation on the subject (2005), the databases used in each of the cited articles are not available for use by the scientific community. They only have their detailed constructions, but they are not available for public access. With the reading of these articles, it was possible to understand more deeply the real need to create and organize data such as those proposed in this work, because, unfortunately, there is no organization or availability of these databases for the scientific community.

Table I: Timeline published works on *Hit Song Science*

Font:By the Author, adapted from [Bertoni A. 2021]

2005	**Automatic Prediction of Hit Song** Dhanaraj - (ISMIR)[Dhanaraj and Logan 2005].
2006	**Predicting Success from music sales data** Chon - (1st ACM)[Chon et al. 2006].
2008	**Hit Song Science is not yet a science** Pachet - (ISMIR)[Pachet and Roy 2008].
2011	**"THE MILLION SONG DATASET"** Mahieux **"labROSA"**, Whitman **"The Echo Nest"** (ISMIR)[Bertin-Mahieux et al. 2011].
2011	**Hit Song Science Once Again a Science?** Yizhao Ni - (MML, Dez 17)[Ni et al. 2011].
2014	**Dance Hit Song Prediction** Herremans - September, 10 (Journal of New Music Research)[Herremans et al. 2014].
2014	**Hit Song Detection Using Lyric Features Alone** Singhi - 27-31 Out (ISMIR)[Singhi and Brown 2014].
2017	**Revisiting the Problem of Audio-Based Hit Song Prediction Using Convolutional Neural Networks** Yang - (IEEE-ICASSP)[Yang et al. 2017].
2018	**Musical Trends and Predictability of Success in Contemporary Songs In and Out of the Top Charts** Interiano - (Royal Society Open Science)[Interiano et al. 2018].

Some datasets have already been made available to the scientific community. Despite not having direct access to the aforementioned databases, the first known work on the subject is from 2005, developed by researchers Dhanaraj and Logan, who proposed the creation of the song bank to be used in a Hit Songs prediction project [Dhanaraj and Logan 2005]. It is also important to cite the work *The Million Song DataSet* (MSD) [Bertin-Mahieux et al. 2011]. The MSD was an article published by Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman and Paul Lamere in 2011 and can be considered a watershed for studies related to this subject. The *Million Song* dataset is a collection of freely available audio resources and metadata for one million tracks of contemporary popular music.

It was from this research that several other researches on *HSS* were developed. Another dataset is that of Olteanu, which provides information about a thousand songs, divided into ten musical genres, with approximately sixty audio features for each song [Olteanu 2020]. Another widely used dataset is that of Ay, which proposes the extraction of about twenty features for 175,000 songs from the years 1921 to 2020 that stood out in the best positions of the Billboard magazine [Ay 2018]. An important detail to note is that, as in the first dataset, these last two also have only international songs.

For this reason, the inexistence of datasets containing information about musical features of Brazilian songs, as well as the huge socio-cultural differences between the Brazilian market and other cultures, were decisive for the realization of this work.

Therefore, in this article we propose the creation of four datasets containing features extracted from a database of contemporary popular Brazilian songs. The datasets are temporally and culturally delimited as they comprise songs that were in vogue from January 2014 to May 2019. Half of the songs in the datasets were hit songs at that time and the others were non-hit songs.

The first proposed dataset brings together characteristics of both a qualitative and quantitative nature, with or without time dependence, for each of the selected songs. The second dataset stores the spectrograms of the "predominant melody" played by the singer's voice in those songs. For the production of the third dataset, we carried out a statistical analysis of the predominant melody and promoted the reduction of the dimensionality of the spectrograms by selecting their perceptually important points. To generate the fourth dataset, we used musical semantic analysis of the melodic ensemble associated with the singer's voice. To do this, we developed an algorithm that, using the support of known music theory, such as concepts about musical scales and harmonic field, managed to identify the most frequent melodic sequences in each analyzed song.

This work describes the creation procedures for these four datasets and is organized as follows: Section 2 describes the construction of the song database; Section 3 deals with the formation of the first data set, with 3215 statistical characteristics; Section 4 deals with extracting the predominant melody from the voice and how it was used to build the second dataset; Section 5 explains the fundamentals of extracting statistical features from the spectrograms of the predominant melody of the voice; Section 6 details the construction of the fourth dataset from melodic semantic information extracted from the predominant melody of the voice. Finally, Section 7 brings the conclusions of this work.

2. CREATION OF THE SONG DATABASE

The first stage consisted of surveying the set of songs that were in the best positions within a ranking system - and that could be classified as a Hit Song - within the period of analysis. The main problem in this stage was choosing the best way to measure the performance of the songs, such as: a) the number of views through social media such as Youtube, Instagram or Facebook; b) if the songs were trending topics [5]; c) the collection and distribution of copyrights of the songs by the Brazilian Central Collection and Distribution Office (ECAD)[6]; d) the number of streams verified on major digital music platforms: Spotify, Deezer, Itunes; e) the number of times the songs were played on Brazilian radio stations within the period of analysis.

In this work, we chose to use the parameter Number of Executions on Brazilian radios, obtained from Connectmix[7] [ConnectMIX 2019], as this parameter is the most common way to check the

[5] ***Trending Topic*** (TT) is a trending topic. This means that a large number of tweets with a hashtag or word(s) related to this topic have been spread by a large number of people over a period of time. When this happens, the subject enters a Twitter ranking of most popular subjects and becomes a Twitter Trending Topic;

[6] **The Central Collection and Distribution Office (ECAD)** is a Brazilian private office responsible for collecting and distributing the copyright of music to its authors, with its headquarters located in Rio de Janeiro.

[7] **Connectmix** All over the world there are companies specialized in this task in monitoring the number of radio plays.

Table II: Ranking 100+ ConnectMix - Year 2014

Font:By the Author, adapted from [Bertoni A. 2021]

year	ranking	artist	song title	style	times played
2014	1º	Marcos e Belutti	Domingo de Manhã	Sert.	384067
2014	2º	Zezé Di Camargo e Luciano	Flores Em Vida	Sert.	273933
2014	3º	Cristiano Araújo	Cê Que Sabe	Sert.	246369
2014	4º	Eduardo Costa	Os 10 M. Do Amor	Sert.	245669
2014	5º	Jorge e Mateus	Calma	Sert.	239337
⋮	⋮	⋮	⋮	⋮	⋮
2014	100º	Fred e Gustavo	Tó Sou Seu	Sert.	59060

performance of new songs by artists and companies linked to the artistic segment, providing greater reliability in data acquisition. Using this performance criterion, we organized the data in our main database as shown in Table II.

Another relevant information is that the methodology used to build our song database is similar to that used by the **Billboard Magazine**[8] [Billboard 2019], which is based mainly on the analysis of how long the songs remain in the top positions of the ranking of each week. As ConnectMix's databases did not deliver information with so much detail, it was decided to use the percentage factor of radio plays as a comparison parameter for each song, in each year of observation.

For the study, an observation period that started in January 2014 until May 2019 was considered, totaling approximately five years. For each year observed, the 100 most played songs on radio stations in Brazil were catalogued. The percentage factor chosen as a comparison index meant that the songs from the last year (2019) were not penalized as much compared to the others years in which the observation had already completed 12 months. This did not harm the data analysis as the main focus was not the classification by number of executions, but the percentage of executions year by year. This, at first, apparently would cause distortions, as not always the most played songs throughout the period (from 2014 to 2019) would emerge in the best positions in the general ranking. This was because the comparison was made year by year, comparing all songs played in the same period. As an example, we can mention the number one song in the ranking of our work, (Domingo de Manhã - Marcos e Belutti), which had only 384,067 plays in 2014 - against 1,191,735 plays in the second place (Apelido Carinhoso - Gusttavo Lima) in the year 2018. This apparent discrepancy occurred because, in the year of analysis of the song that ranked first, the overall number of plays among all songs in that year was much lower than in subsequent years. We could also say that the number of times the songs played on the radio in 2014 were much more diluted by a greater number of artists, while in 2018 the number of plays were more concentrated in a smaller proportion of artists. Another reason for adopting this methodology is the fact that the songs cataloged in the period could be observed for all the years following their release. In other words, a song that was ranked better in the first year of observation could have its performance verified in the following years; even songs that didn't do so well in the year of release, because they weren't released in such a favorable time frame, could rank better in subsequent years.

For the construction of the non-hit song database, in order to generate a greater balance within the song database, we started from the premise that such songs should, in principle, belong to the same set of verified artists in the first class, but not listed in any position in the general ranking of the

In Brazil, this service is also provided by the company Connectmix, which offers a self-titled real-time monitoring tool for auditing and managing Broadcasting for radio and television stations.

[8]Belonging to the American group Prometheus Global Media, Billboard is one of the oldest magazines in the world, founded in 1894 and specializing in information about the music industry. Its best-known ranking, the "Hot 100", shows the top 100 best-selling singles played on radio and is often used in the United States as the primary way of measuring the popularity of artists as well as a song.

most played songs, according to ConnectMIX's own system. As a second criterion, considering the balance between the experiment databases, the same number of songs was taken in the two databases (hit songs and non-hit songs) for each artist, that is, if an artist appeared 5 times (with five songs) within the hit songs database, then, there should be, necessarily, 5 songs by the same artist within the non-hit songs database. Another important piece of information is that all the non-hit songs were also selected within the same evaluation period as the non-hit songs, that is, from 2014 to 2019. The explanation for this attitude was that the intention was to delimit the observation of data in such a way that it takes place temporally within a specific period - and for a group of artists who were more in vogue during the period of observation, generating a solid balance in the dataset.

The choice was, to a certain extent, random. That's because there are no parameters of comparison or ranking that classify the worst songs. If they are classified as heavily played songs, they are hit-songs, if they are not played or not played much, they are non-hit songs. In this way, a ranking was set up with the 600 best positioned songs, considering an observation interval from January 2014 to May 2019, with the 100 most played songs of each year.

With the 600 most played songs on the radio in this period, it was necessary to eliminate some inconsistencies in the database: a) songs that were repeated in more than a year of analysis; b) different versions of the same songs (live or studio); c) songs recorded by more than one artist in the period; d) international songs (which were not exclusively in Portuguese), as they are not part of the scope of this study. Thus, the number of 882 songs was reached, 441 being labeled hit songs and 441 as non-hit songs.

The complete list of all songs used in this work, as well as some additional statistical information can also be consulted in the repository `https://github.com/tocaestudio/JIDM_2021`.

2.1 Adopting a standard duration for each song

In order to reduce the computational effort of extracting the features, it is convenient to reduce the observation time interval of each song, because, as the melody, harmony and lyrics of the songs are repeated more than once throughout the musical records, we could take repeated snippets of analysis. Furthermore, as these records originally have different sizes, it is necessary to standardize the length of the observation time interval so that the same amounts of features extracted from each song are produced. To define the duration of this time interval, a brief statistical study was carried out on the approximate time of the musical structures of the selected songs.

Starting from the premise that songs performed on radios have a purely commercial purpose, these songs almost always end up sharing a very similar musical structure (metric), being composed, basically, almost always by the same structures.: Introduction, Verse A/Pre-Chorus, Chorus, Solo, Verse B/Pre-Chorus, Final Chorus e Final Solo, as illustrated in Figure 1.

The analysis revealed that, on average, the first 90 seconds of each song cover: Introduction, Verse A/Pre-Chorus and Chorus, which bring together the main characteristics of interest of each musical record, since after Chorus, songs usually repeat the previous passages until the end of the song. With that, it is concluded that the songs can only be considered technically original until the end of the first Chorus of the Song. From this part onwards they are just repetitions of the first part, until the song comes to an end. Table III represents the statistical calculation of the ten songs analyzed in Figure 1.

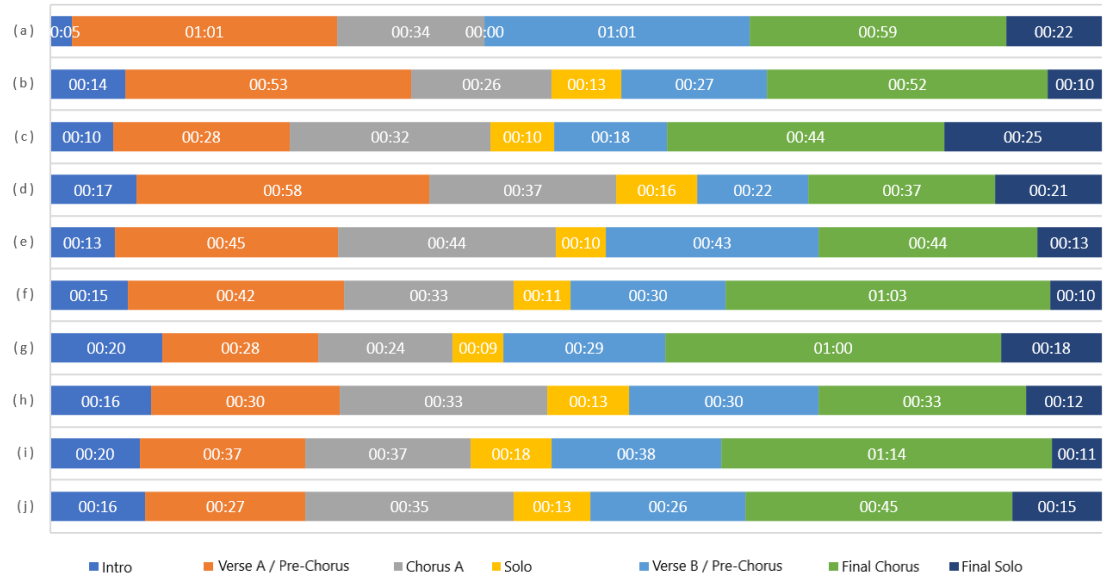


Fig. 1: Arrangement and duration of the structural elements of the hit songs (a, b, c, d, e) and non-hit songs (f, g, h, i, j)
Source: By the Author, adapted from [Bertoni A. 2021]

Table III: Minimum, Maximum and Average - Musical Structure

Font:By the Author, adapted from [Bertoni A. 2021]

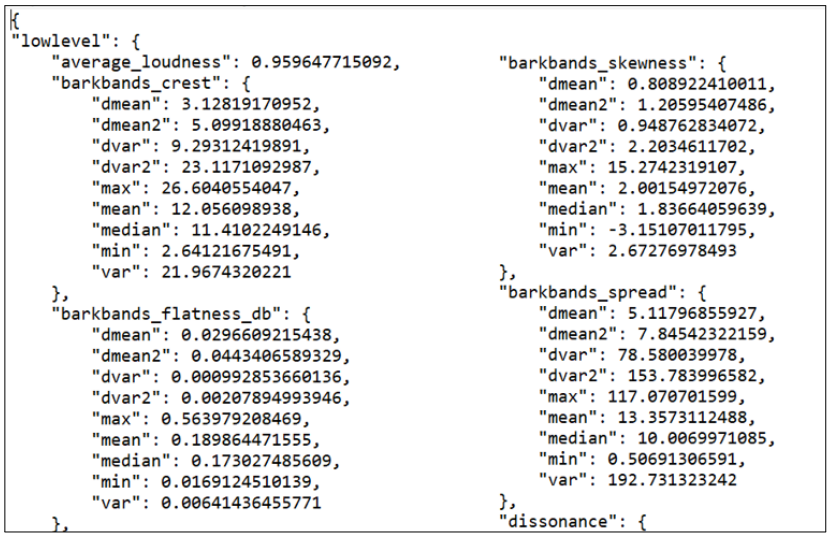
lenght	Intro	Verse A	Chorus	Solo	Verse B	End Chorus	End Solo
Min	00:05	00:27	00:24	00:00	00:18	00:33	00:10
Max	00:20	01:01	00:44	00:18	01:01	01:14	00:25
Average	00:15	00:41	00:34	00:11	00:32	00:51	00:16

Average Duration of Songs until the end of the 1st. Chorus	**01:29**

3. THE FIRST BRAZILIAN SONG FEATURES DATASET

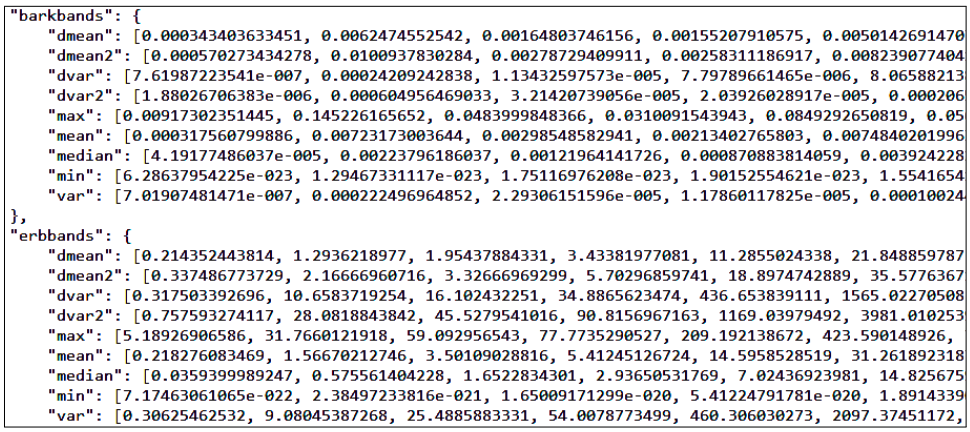
For the extraction of features from the songs, the Streaming Extractor Music application was used, which makes up the Essentia package [Bogdanov et al. 2013]. Essentia is an open source C++ library with Python and JavaScript bindings for audio-based musical information analysis and retrieval. The codes were executed on the Linux Operating System (V. 18.04.4 LTS), which is the environment, according to the developers, where there is greater compatibility between the python libraries and the proposed codes.

Besides the Essentia package, codes were developed to automate the feature extraction process. As extraction applications generate as output a single **json** file containing all the features, it was necessary to write code that could load, extract and save the files in an automated way. Figure 2 shows a part of the structure of the file **json**, showing time-invariant characteristics, that is, they are calculated based on the entire length of the audio file. The extractors also offer time-varying characteristics, calculated using temporal windowing, which are shown in Figure 3.



```
{
"lowlevel": {
    "average_loudness": 0.959647715092,
    "barkbands_crest": {
        "dmean": 3.12819170952,
        "dmean2": 5.09918880463,
        "dvar": 9.29312419891,
        "dvar2": 23.1171092987,
        "max": 26.6040554047,
        "mean": 12.056098938,
        "median": 11.4102249146,
        "min": 2.64121675491,
        "var": 21.9674320221
    },
    "barkbands_flatness_db": {
        "dmean": 0.0296609215438,
        "dmean2": 0.0443406589329,
        "dvar": 0.000992853660136,
        "dvar2": 0.00207894993946,
        "max": 0.563979208469,
        "mean": 0.189864471555,
        "median": 0.173027485609,
        "min": 0.0169124510139,
        "var": 0.00641436455771
    },
    "barkbands_skewness": {
        "dmean": 0.808922410011,
        "dmean2": 1.20595407486,
        "dvar": 0.948762834072,
        "dvar2": 2.2034611702,
        "max": 15.2742319107,
        "mean": 2.00154972076,
        "median": 1.83664059639,
        "min": -3.15107011795,
        "var": 2.67276978493
    },
    "barkbands_spread": {
        "dmean": 5.11796855927,
        "dmean2": 7.84542322159,
        "dvar": 78.580039978,
        "dvar2": 153.783996582,
        "max": 117.070701599,
        "mean": 13.3573112488,
        "median": 10.0069971085,
        "min": 0.50691306591,
        "var": 192.731323242
    },
    "dissonance": {
```


Fig. 2: Example of an excerpt from the json file generated by the Essentia Extraction tool, in which some features and the respective values of their statistical descriptors appear.
Source: By the Author, adapted from [Bertoni A. 2021]



```
"barkbands": {
    "dmean": [0.000343403633451, 0.0062474552542, 0.00164803746156, 0.00155207910575, 0.0050142691470
    "dmean2": [0.000570273434278, 0.0100937830284, 0.00278729409911, 0.00258311186917, 0.008239077404
    "dvar": [7.61987223541e-007, 0.00024209242838, 1.13432597573e-005, 7.79789661465e-006, 8.06588213
    "dvar2": [1.88026706383e-006, 0.000604956469033, 3.21420739056e-005, 2.03926028917e-005, 0.000206
    "max": [0.00917302351445, 0.145226165652, 0.0483999848366, 0.0310091543943, 0.0849292650819, 0.05
    "mean": [0.000317560799886, 0.00723173003644, 0.00298548582941, 0.00213402765803, 0.0074840201996
    "median": [4.19177486037e-005, 0.00223796186037, 0.00121964141726, 0.000870883814059, 0.003924228
    "min": [6.28637954225e-023, 1.29467331117e-023, 1.75116976208e-023, 1.90152554621e-023, 1.5541654
    "var": [7.01907481471e-007, 0.000222496964852, 2.29306151596e-005, 1.17860117825e-005, 0.00010024
},
"erbbands": {
    "dmean": [0.214352443814, 1.2936218977, 1.95437884331, 3.43381977081, 11.2855024338, 21.848859787
    "dmean2": [0.337486773729, 2.16666960716, 3.32666969299, 5.70296859741, 18.8974742889, 35.5776367
    "dvar": [0.317503392696, 10.6583719254, 16.102432251, 34.8865623474, 436.653839111, 1565.02270508
    "dvar2": [0.757593274117, 28.0818843842, 45.5279541016, 90.8156967163, 1169.03979492, 3981.010253
    "max": [5.18926906586, 31.7660121918, 59.092956543, 77.7735290527, 209.192138672, 423.590148926,
    "mean": [0.218276083469, 1.56670212746, 3.50109028816, 5.41245126724, 14.5958528519, 31.261892318
    "median": [0.0359399989247, 0.575561404228, 1.6522834301, 2.93650531769, 7.02436923981, 14.825675
    "min": [7.17463061065e-022, 2.38497233816e-021, 1.65009171299e-020, 5.41224791781e-020, 1.8914339
    "var": [0.30625462532, 9.08045387268, 25.4885883331, 54.0078773499, 460.306030273, 2097.37451172,
```


Fig. 3: Example of excerpt generated by Essentia's feature extractor tool, which shows some features and the respective values of their statistical descriptors extracted by Temporal Windowing
Source: By the Author, adapted from [Bertoni A. 2021]

Finally, the code extracts features represented by categorical variables, as shown in Figure 4, into the Harmonic Field of the song ***(chords_key and chords_scale)***, which are described by values of type *string*. As these variables do not assume defined numerical values, the technique of transforming into binary variables was used, that is, for each category a new predictor is created that can assume binary values (0 or 1). This type of treatment is quite common in Data Science, and is known as Dummy Variables.

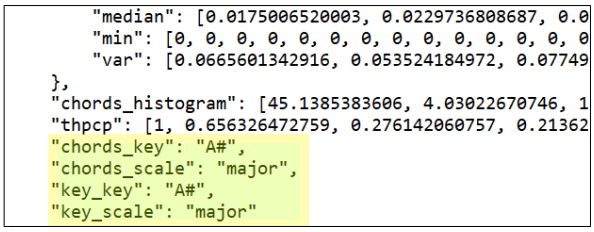


Fig. 4: Example of an excerpt from the json file in which the categorical variables and their estimated values were highlighted in yellow, extracted using the Essentia package tool
Source: By the Author, adapted from [Bertoni A. 2021]

3.1 ***Parsing***

This process consists of extracting and reorganizing the data contained in the **json** files, converting them into a table, for which the CSV format was used. The *Parsing* technique is very important in this process, because only from this organization will it be possible to properly manipulate the data obtained, enabling the treatment of possible inconsistencies, which, in Data Science, is very common and almost always necessary .

3.2 Treatment using the Pandas library

Only after performing Parsing was it possible to use it as a fully functional database within Pandas. At first, a visual search was performed, trying to find the most common inconsistencies. They are: unwanted features, missing data (NaN), null, divergent, duplicated, outliers and, finally, categorical variables. After all these treatments, it was finally possible to format the forecast data matrix and the data vector that represents the class. The final Matrix of data was as follows: 882 Lines (One line for each Song); 3215 Columns (Predictors - "Characteristics"); 1 Column (Classifier - hit song "0" / non-hit Song "1").

4. DATASET WITH THE PREDOMINANT MELODY OF THE SINGING VOICE

According to Jason Blume, a renowned international composer, melody is the main key to the success of any song [Blume 2019]. In order to be able to evaluate this statement, we propose a dataset containing the predominant melodic lines of the singing voice for each of the 882 songs analyzed in our database.

4.1 Basic Music Theory

This section will explain the entire set of Musical rules that will be necessary for the creation and development of the algorithms for extracting and treating the Semantic Melodic features database, which will be explained in detail later.

4.1.1 *Tempered Musical Scale.* The Tempered Scale represents a scale with twelve semitones equally distributed across the octave. In this scale, the interval between C and C♯,[9] is the same between C♯ and D. In addition, the so-called **enharmonic** notes (with intervals different from the semitone), start to have the same frequency, which does not occur in the Pythagorean scales[10], fair and medium tone [Rossing et al. 2014].

[9]the musical symbol ♯ (sharp) is used to represent a half-step increase from the original tone, while the symbol ♭ is used to lower half a step from the original key.
[10]Its construction is based on the superposition of fifths (ratio of 3/2) and their inversions, the fourths (ratio of 4/3)

If i is the interval between each tempered scale semitone, a fifth interval (7 semitones) is $i^{\,7}$, a fourth interval (5 semitones) is $i^{\,5}$, an interval second (2 semitones) is $i^{\,2}$, and so on. The octave interval (12 semitones), given by $i^{\,12}$, has the ratio of 2/1, so:

$$i^{12} = \frac{2}{1} \rightarrow i = 2^{\frac{1}{12}} = \mathbf{1.05946} \tag{1}$$

This is the interval value of a tempered semitone. Similarly, any other tempered scale interval can be calculated using the expression $i_n = 2^{\frac{n}{12}}$, where n is the number of semitones contained in the interval. For example, to calculate the frequency of an E fifth above (7 semitones) of an A at 440 Hz, we have:

$$F_i = f_o \cdot 2^{\frac{n}{12}} = 440 \cdot 2^{\frac{7}{12}} = 440 \cdot 1.498 = 659,25Hz \tag{2}$$

The tempered scale will be used in this work within the range from $C2 = 130,812$ Hz to $B5 = 1975,533$ Hz as described in Table IV. The main reason for choosing this range was primarily to look for a number of complete octaves, but that were also approximately within the range of fundamental frequencies emitted by the human voice for popular Brazilian songs. Furthermore, after a brief statistical study of the frequencies obtained in the main dataset, it was found that the frequencies chosen ($C2$ to $B5$), are really suitable for the creation of the bandpass filter to which it will later be submitted, without significant loss of information. The reason for constructing the bandpass filter is to avoid *outliers*.

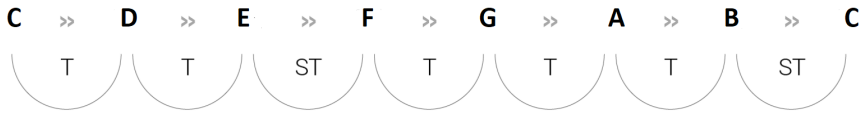
Table IV: Octaves Used (Notes and Frequencies)

Font:By the Author, adapted from [Bertoni A. 2021]

musical note	**Frequency (Hz)**	**musical note**	**Frequency (Hz)**
C2	130.812775	C3	261.625519
C♯2	138.591324	C♯3	277.182648
D2	146.832367	D3	293.664734
D♯2	155.563492	D♯3	311.126984
E2	164.813782	E3	329.627533
F2	174.614105	F3	349.228241
F♯2	184.997208	F♯3	369.994385
G2	195.997711	G3	391.995392
G♯2	207.652344	G♯3	415.304688
A2	220	A3	440
A♯2	233.081848	A♯3	466.163788
B2	246.941635	B3	493.883301
C4	523.251099	C5	1046.502075
C♯4	554.365234	C♯5	1108.730591
D4	587.329529	D5	1174.659058
D♯4	622.253906	D♯5	1244.507935
E4	659.255127	E5	1318.510254
F4	698.456482	F5	1396.912964
F♯4	739.988831	F♯5	1479.977539
G4	783.990845	G5	1567.981812
G♯4	830.609375	G♯5	1661.21875
A4	880	A5	1760
A♯4	932.327576	A♯5	1864.654785
B4	987.766602	B5	1975.533325

4.1.2 *Major and Minor Scale.* s

All major scales are formed from the idea of the C (C Major) scale, which is taken from the following idea:



Font: By the Author, adapted from [Bertoni A. 2021]

Fig. 5: C Major Scale - “base formula”

From the formation structure of the C Major scale, any other scales are formed. Just understand that there is a logical sequence of tonal distances between each note on the scale. For example: between the first and second notes of the C major scale, there is a full tone (T). Between D and E, there is also an integer tone (T). Between E and F, there is only one semitone (ST).
The idea is then repeated until the entire octave cycle from **C to C** is completed. It can then be said that the formula for building the C major scale is:

$$T - T - ST - T - T - T - ST \tag{3}$$

In fact, this formula is used to compose all other Major scales. Therefore, if from Note D, the formulation described above in 3 is applied, it can then be concluded that the scale of D Major is given by:

Font: By the Author, adapted from [Bertoni A. 2021]

Fig. 6: D Major Scale

It can now be seen that the notes F and C will be Sustained (half a step above the natural), because when applying the formula 3, it will be necessary to increase a tone between E and F, resulting in F♯ and no longer in F, as in the C natural scale, because in this scale the natural difference from E to F would only be half a step.

The same reasoning is repeated for all the other Major Scales, always respecting the formula 3.

The Natural Minor scale is a scale that can be represented by the following tonal distribution:

Font: By the Author, adapted from [Bertoni A. 2021]

Fig. 7: A Minor Scale

Therefore, the formula that represents the Natural Minor Scales are:

$$T - ST - T - T - ST - T - T \tag{4}$$

and can be used in the construction of any other Natural Minor scales using the same reasoning.

As melody is characterized by the variation of musical notes (that is, frequencies) over time, spectrogram throughout the 90 seconds of the song allow the visualization and time-frequency analysis of each song's melody, as shown in an excerpt in 8 (B), where the waveform in (A) and the extracted melody in (C) are also exemplified. Except for a few purely instrumental parts, notably in the intro arrangement, the predominant melody of the songs is mainly performed by the singing voice.

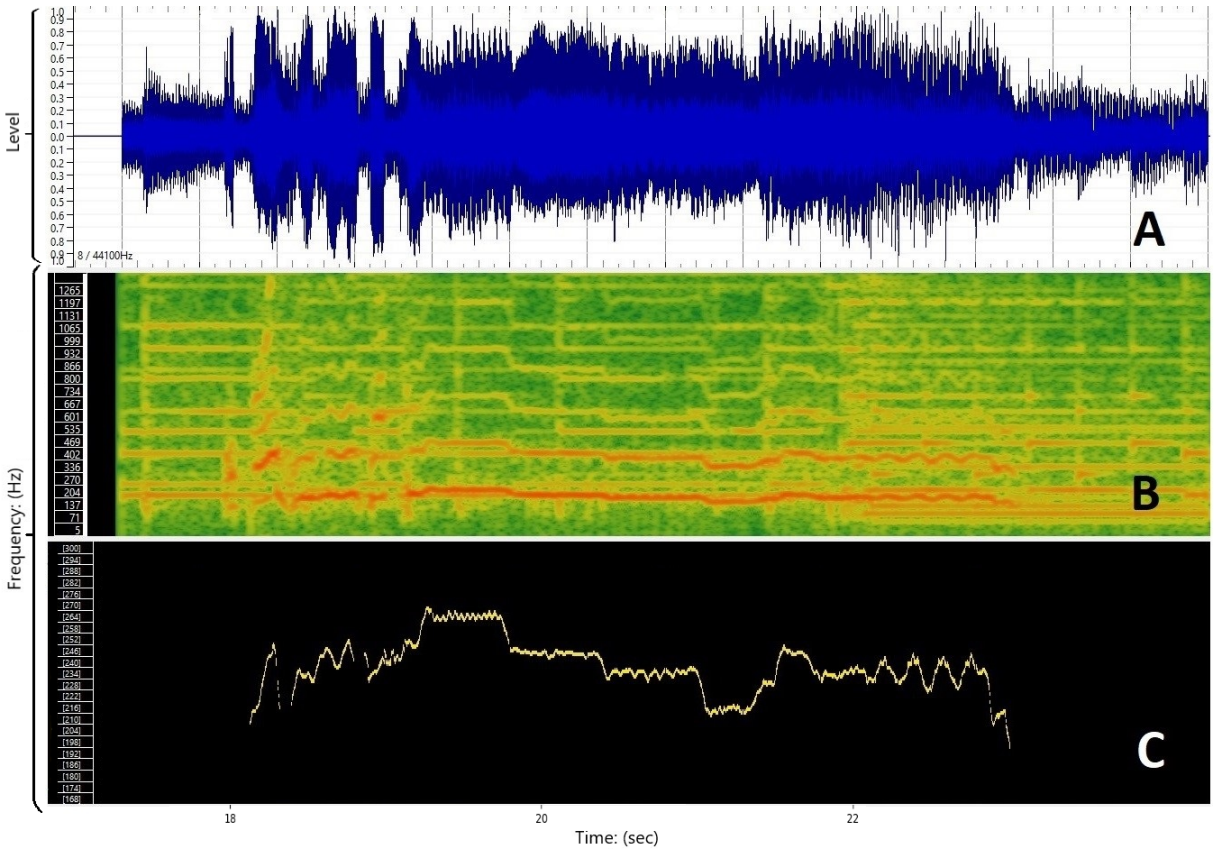


Fig. 8: Sample excerpt containing the waveform (A), the spectrogram (B) and the melody of the main voice (C).
Source: By the Author, adapted from [Bertoni A. 2021]

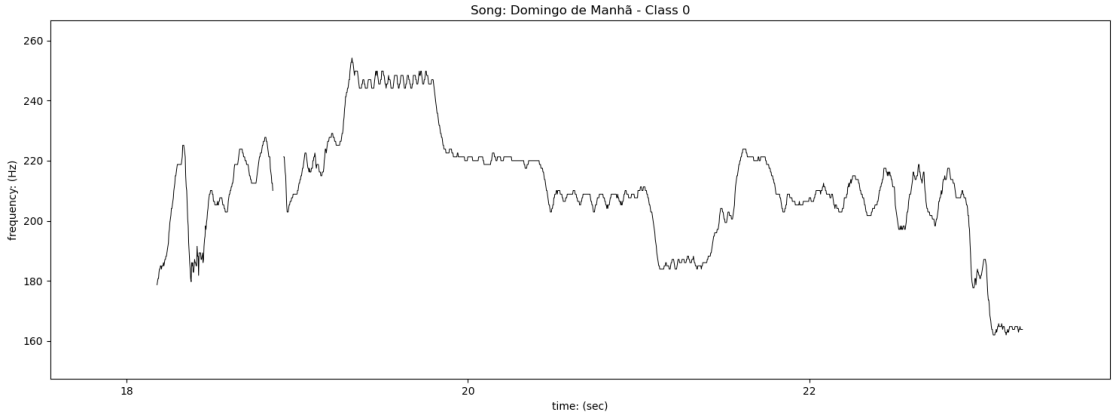


Fig. 9: Sample excerpt containing the melody frequency variation over time
Source: By the Author, adapted from [Bertoni A. 2021]

So, to extract the predominant melody of the singing voice from the spectrograms, we propose to apply the method introduced by Justin Salomon [Salamon and Gómez 2012]. As this method is already implemented and available in the Essentia package, we used this tool to perform the proposed task. Figure 9 illustrates one of the time-frequency traces of the predominant melody given by Essentia and stored in CSV format [Salamon 2013].

5. DATASET WITH STATISTICAL FEATURES OF THE PREDOMINANT MELODY

Since the predominant melody traces are actually time series of frequencies, we propose to extract statistical information using the tool tsfresh [Christ et al. 2018]. However, due to the high memory cost of processing time-series data, the 31728 points of each of the 882 melody traces ended up harming the processing and overflowing the IDE's memory (Spyder). This made the analysis and extraction of new features unfeasible.

On the other hand, visual examination of the predominant melody traces reveals that only some of the data points actually contribute to the overall shape of the time series, while most others can even be discarded. In order to reduce the time-series dimensionality and thus the memory cost, we propose to use the concept of Perceptually Important Points (PIP) [Fu et al. 2017]. In this way, we developed a PIP code to detect and maintain only the zigzag corner points of the melody traces as shown in 10. This procedure allowed us to reduce the amount of predominant melody points from 31728 to just 2,000 per song.

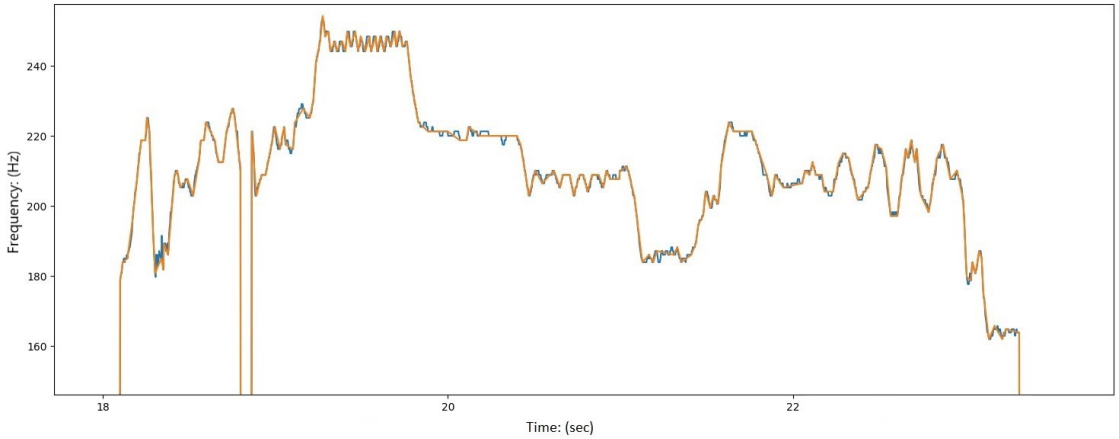


Fig. 10: Sample audio clip containing the original waveform (in blue) and the waveform after downsizing to two thousand points (in orange).
Source: By the Author, adapted from [Bertoni A. 2021]

After performing the dimensionality reduction, we were able to use tsfresh. Tsfresh automatically extracts hundreds of features from Time Series. Those features describe basic characteristics of the time series such as the number of peaks, the average or maximal value or more complex features such as the time reversal symmetry statistic. These features can be used as predictors in Artificial Neural Networks, together with the previously proposed databases. The reason for choosing this feature extractor was mainly linked to the fact that the intention was to make datasets more robust, offering more possibilities for combinations and analysis by Artificial Neural Networks.

6. DATASET WITH SEMANTIC FEATURES OF THE PREDOMINANT VOICE MELODY

Another way of extracting useful information from melody is through Musical Semantic Analysis. In this sense, we then propose an algorithm to find similarity melodic patterns along the predominant melody, in such a way that it would be possible to create a new dataset with semantic information of each song.

However, Figure 9 allows us to observe that, in addition to the song's melodic line, the traces of the predominant melody of the singing voice present frequency oscillations associated with the vibrato effect. Singers use vibrato as a vocal expression and support to keep the pitch throughout the interpretation. Legato vocal articulations can also be identified, associated with links between notes, both rapidly moving up and down a scale. These effects hinder Musical Semantic Analysis due to the large amount of notes outside the main melody found in short time intervals.

To smooth the outline of the predominant melody, we propose using Musical Theory to mitigate these unwanted artifacts. First, we limited frequencies to the 125 to 1500 Hz range, which concentrates the highest power spectral density and best represents the voices in the song databases. As the predominant melody trace amplitudes were already in Hz, no filtering was necessary. All we needed to do was exclude trace amplitudes that were outside the specified frequency range [Christiano and Fitzgerald 2003].

Next, as simple moving average filters allow waveform smoothing and noise removal [Lima 2015], we computed the moving average of predominant melody frequencies over 65-point windows time-shifted by one point. This procedure allowed us to minimize vibrato oscillations, as shown in 11.

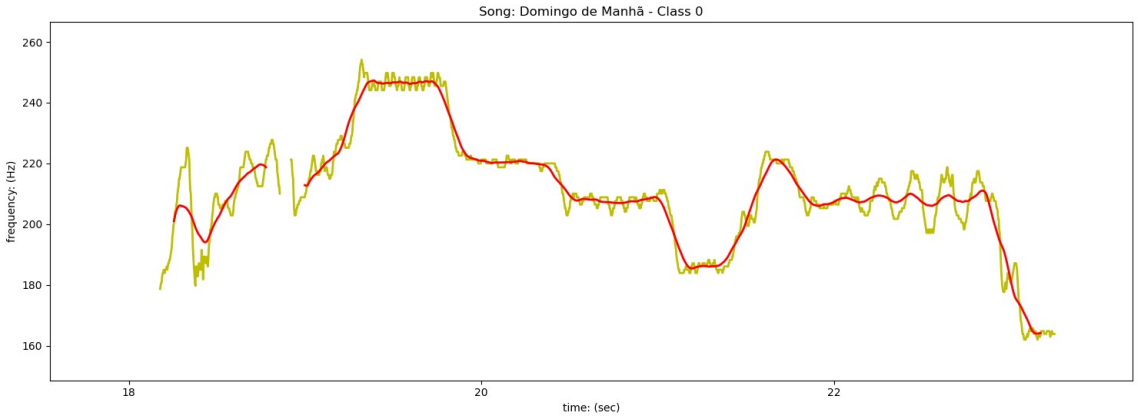


Fig. 11: Audio segment containing the result of the first filtering with a range from 125 Hz to 1500 Hz (in yellow); and the result of applying the moving average filter (in red).
Source: By the Author, adapted from [Bertoni A. 2021]

To remove outliers from the melodic line of each song, the concept of tempered musical scale was used to preserve only the frequency values covered by the full octaves of C2 (C II), 130.812 Hz, up to B5 (Si V), 1975,533 Hz. Table IV shows the complete range used in the analysis.

Then, the values of frequencies of the melody line were quantized by the frequencies of the notes of the harmonic field of each song. In this way, the melodic line assumed the shape shown in Figure 12.

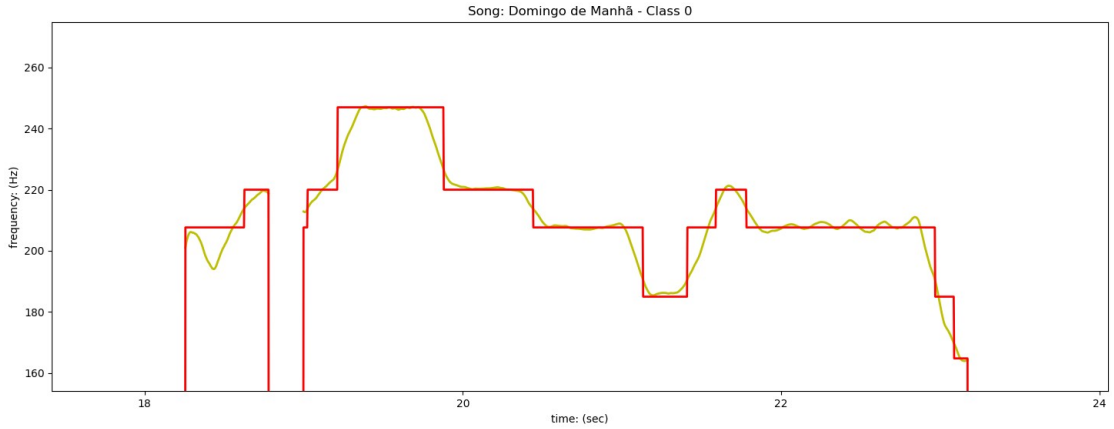


Fig. 12: Melody frequency variation over time after oscillation smoothing (vibratos e legatos)
Source: By the Author, adapted from [Bertoni A. 2021]

After smoothing the entire Predominant Melody traces, it was possible to propose an algorithm to extract semantic information from the predominant melody of the singing voice. To accomplish this task, it performed a point-to-point scan over the first 90 seconds of each song. After sequentially checking each frequency in the melody line, the algorithm creates a string for each sequence of musical notes found in a certain interval, ignoring repetitions. The stopping point, which delimits the end of each melodic sequence, is established when frequencies equal to zero are found, which naturally indicate the singer's breathing pause. When a new sequence is found, a new column with the name of the melodic sequence is created in the semantic information database, assigning the value 1 to this column, as it is the first sequence found with that note pattern. When an already known sequence is

found, the value present in the column corresponding to that sequence is incremented. The dataset has 882 rows and 9694 columns. Each row represents a respective song, while the 9694 columns represent all melodic sequences found. The first 441 lines refer to hit songs, while the remaining 441 refer to non-hit songs. The figure 13 represents a small part of the dataset of melodic sequences.

Index	A2_A_Sust2	G2	C2_A_Sust2_C2_C3	A2_A_Sust2_A2	C3	C3_C2_A_Sust2_A2_A_Sust2_C2_C3	C3_A_Sust3	D4_D_Sust4_D4_C4_D4
0	0	0	0	0	0	0	0	0
1	2	2	1	1	7	1	1	2
2	0	4	0	0	2	0	0	0
3	3	3	0	0	6	0	0	0
4	1	1	0	0	1	0	0	0
5	0	8	0	0	0	0	0	0
6	0	0	0	0	7	0	0	0
7	0	0	0	0	0	0	0	0
8	0	1	0	0	3	0	0	0
9	1	2	0	0	0	0	0	0

Fig. 13: Excerpt extracted from the dataset of Melodic Sequences.
Source: By the Author, adapted from [Bertoni A. 2021]

7. CONCLUSIONS

Data Science is a segment of Computing that has been growing a lot in recent years. The lack of new databases has always been a challenge, especially for studies involving Music. The objective of this work is to contribute with the emerging research in Data Science related to music. We believe that the offer of new datasets would be of great value.

Most of datasets available in the Internet focus mainly on the North American market, comprising hundreds of thousands of songs that only cover the English language, that is, little information (features) of a very large number of songs, limited to practically a single language and cultural context. Unlike the others, the four datasets developed here not only focus on a temporally reduced period, but also reflect the musical preferences of the largest country in Latin America, whose culture has impacted the world music scene for decades and currently constitutes a market of 220 millions of people[IBGE 2021].

Our datasets add more than 3,200 new different features to those currently available. The first encompasses all the possible features offered by the Essentia package. The second offers new features, which are the predominant melodies of the voice - something never proposed and freely available on the internet for studies. The third dataset is also innovative as it extracts statistical features from a new feature, never before proposed. Finally, the fourth dataset is the most innovative, as it uses the melodic sequence of the voice to extract new characteristics from a musical point of view using music theory.

The authors understand that there are numerous points to improve. The extraction of semantic information proves to be a very broad field in this type of work, as it is, in a way, information of an interpretive nature, that is, by changing the way of interpreting the analyzed musical arguments, it is possible to create a new dataset for analysis. And that is certainly a point to explore. In future work, we intend to obtain new information from the dataset of predominant melodies of the voice by relating harmony and melody.

All datasets described here can be found at `https://github.com/tocaestudio/JIDM_2021`.

REFERENCES

Ay, Y. E. Spotify dataset 1921-2020, 160k+ tracks, 2018.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. The million song dataset, 2011.

Bertoni A., L. R. P. Três datasets criados a partir de um banco de canções populares brasileiras de sucesso e não-sucesso de 2014 a 2019, 2021.

Billboard. Billboard magazine., 2019.

Blume, J. What makes a song a hit?, 2019.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. R., and Serra, X. Essentia: an audio analysis library for music information retrieval. In *International Society for Music Information Retrieval Conference (ISMIR'13)*. ESSENTIA - UPF - Universitat Pompeu Fabra, Curitiba, Brazil, pp. 493–498, 2013.

Chon, S. H., Slaney, M., and Berger, J. Predicting success from music sales data: a statistical and adaptive approach. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. ACM, pp. 83–88, 2006.

Christ, M., Braun, N., Neuffer, J., and Kempa-Liehr, A. W. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing* vol. 307, pp. 72–77, 2018.

Christiano, L. J. and Fitzgerald, T. J. The band pass filter. *international economic review* 44 (2): 435–465, 2003.

ConnectMIX. Connectmix, monitoramento, auditoria e gestão de áudio em tempo real em rádios e tvs., 2019.

Dhanaraj, R. and Logan, B. Automatic prediction of hit songs. pp. 488–491, 2005.

Fu, T.-c., Hung, Y.-k., and Chung, F.-l. Improvement algorithms of perceptually important point identification for time series data mining. In *2017 IEEE 4th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE, IEEE, https://ieeexplore.ieee.org/document/8279589, pp. 11–15, 2017.

Herremans, D., Martens, D., and Sörensen, K. Dance hit song prediction. *Journal of New Music Research* 43 (3): 291–302, 2014.

IBGE. População do brasil, 2021.

Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., and Komarova, N. L. Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society open science* 5 (5): 171274, 2018.

Lima, J. N. A utilização de filtros digitais em séries temporais gnss. In *Comunicação apresentada em VIII Conferência Nacional de Cartografia e Geodesia: VIII CNCG*. VIII CNCG, *https* : *//viiicncg.ordemengenheiros.pt/fotos/editor*2/$VIIICNCG/cncg2015_comunicao_80.pdf$, 2015.

Ni, Y., Santos-Rodriguez, R., Mcvicar, M., and De Bie, T. Hit song science once again a science. In *4th International Workshop on Machine Learning and Music*. Citeseer, 2011.

Olteanu, A. Gtzan dataset - music genre classification, 2020.

Pachet, F. and Roy, P. Hit song science is not yet a science. pp. 355–360, 2008.

Raieli, R. *Multimedia Information Retrieval: theory and techniques*. Philadelphia, PA : Chandos Pub., Oxford, UK, 2013.

Rossing, T. D., Moore, F. R., and Wheeler, P. A. *The science of sound*. Pearson, 2014.

Salamon, J. *Melody Extraction from Polyphonic Music Signals*. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2013.

Salamon, J. and Gómez, E. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing* 20 (6): 1759–1770, 2012.

Singhi, A. and Brown, D. G. Hit song detection using lyric features alone. *Proceedings of International Society for Music Information Retrieval*, 2014.

Yang, L.-C., Chou, S.-Y., Liu, J.-Y., Yang, Y.-H., and Chen, Y.-A. Revisiting the problem of audio-based hit song prediction using convolutional neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 621–625, 2017.