4MuLA: A Multitask, Multimodal, and Multilingual Dataset of Music Lyrics and Audio Features

Angelo Cesar Mendes da Silva
Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo
São Carlos, SP, Brazil
angelo.mendes@usp.br

Diego Furtado Silva
Departamento de Computação - Universidade Federal de São Carlos
São Carlos, SP, Brazil
diegofs@ufscar.br

Ricardo Marcondes Marcacini
Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo
São Carlos, SP, Brazil
ricardo.marcacini@usp.br

ABSTRACT

We present a new benchmark dataset of songs with structured information to be applied in various machine learning tasks. The data comes from a platform focused on lyrics information, but contain several other annotations provided by their users. Our dataset, called 4MuLA (Multitask, Multimodal, and Multilingual Music Lyrics and Audio features dataset), includes features extracted from 96,458 songs distributed by 15,310 artists in 76 genres. In particular, our dataset contains latin music genres that are often overlooked in other benchmark datasets. For each track, we make available various acoustic features, extracted tags, and lyrics in English, Portuguese, or Spanish. With these features, researchers can use our dataset for, at least, lyrics-, audio- or multimodal-based genre classification, music and artist similarity, and popularity regression. Moreover, we can perform cross- or multilingual text analysis on lyrics, such as discourse analysis or measuring the differences between emotion transmitted by audio and lyrics.

CCS CONCEPTS

• **Information systems → Multimedia databases**.

KEYWORDS

music dataset, multimodal musical dataset, latin musical dataset

ACM Reference Format:
Angelo Cesar Mendes da Silva, Diego Furtado Silva, and Ricardo Marcondes Marcacini. 2020. 4MuLA: A Multitask, Multimodal, and Multilingual Dataset of Music Lyrics and Audio Features. In *Brazilian Symposium on Multimedia and the Web (WebMedia '20), November 30-December 4, 2020, São Luís, Brazil.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3428658.3431089

1 INTRODUCTION

The lack of public datasets is a relevant issue for the Music Information Retrieval (MIR) research since its first endeavors. Copyrights, large volumes of data, and the specific needs of different tasks are clear reasons for this inconvenience. For instance, datasets for tagging and classification may publish the original audio if they use Creative Commons-licensed music, such as FMA [7] and Jamendo [3]. On the other hand, datasets for cover song identification, such as DA-TACOS [25], do not have Creative Commons data as an option since music protected by copyrights are the most commonly covered.

For this reason, datasets for varied tasks were created. In the last edition of the International Society for Information Retrieval Conference, for instance, more than ten novel datasets for specific tasks were presented [2, 4, 6, 8, 11–13, 15, 18, 21, 23, 25]. However, it leads to a considerable heterogeneity of quality, organization, and type of data. These characteristics increase the difficulty in curate data, reproducing results, and other research and practical problems.

In this paper, we presented a dataset to be used for different tasks and applied in varied scenarios. The data comes from a platform focused on lyrics information but contains several other annotations provided by their users. Our dataset comprises lyrics from different languages and audio features of a thousand songs labeled with one or more genres. The main features of our dataset, which gives it the name 4MuLA – a Multitask, Multimodal, and Multilingual Music Lyrics and Audio features dataset – are:

- Multitask: Our dataset has annotations of the genre (multi-label and ordered by relevance), artist, the similarity between songs and artists, popularity index. These labels can be used as the target for many different MIR tasks.
- Multimodal: We make available the lyrics of each song and various features extracted from the audio. So, researchers or practitioners may investigate how these modalities complement each other.
- Multilingual: The lyrics of the songs in our dataset are from three different languages: English, Spanish, and Portuguese. This way, it is possible to use multilingual or cross-lingual techniques, as well as subsets containing one single language.

With these features, researchers can use our dataset for, at least, lyrics-, audio- or multimodal-based genre classification, music and artist similarity, and popularity regression. Also, one may be interested in studying the differences between regional genres, given we provide features for genres like hip-hop (Northern American), bossa nova (Brazilian), and salsa (Central/Latin American). Moreover, we can perform cross- or multilingual text analysis on lyrics, such as discourse analysis or measuring the differences between emotion transmitted by audio and lyrics.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WebMedia '20, November 30-December 4, 2020, São Luís, Brazil
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8196-3/20/11…$15.00
https://doi.org/10.1145/3428658.3431089

2 RELATED WORK

The number of commercial music releases protects by copyrights is much higher than the number of available music on open repositories or under Creative Commons license. To build learning models with functional data generalization capacity, we need to process the audio signal for the songs, lyrics, and other metadata to extract relevant features representing the music's characteristics, but this is a difficult task due to the limitations of access to music.

We can easily observe these limitations in traditional datasets [1, 9, 10, 24] that have audio files, but do not contain a large amount of music, concentrate a large part of them in a small standard set of genres, or do not have a multimodal representation. Besides, these datasets are usually limited to Western music.

In general, most public music datasets neglect Latin music. In the absence of datasets that contain Latin songs with information and quantity that allows a consistent musical representation, some authors build specific datasets to use in their studies. Usually, these efforts are limited to one task. Besides, they are specific for Latin music, i.e., do not include music from other countries in the same dataset. The BR_Mus [22] is an exception to the latter limitation. The dataset contains songs from twelve genres, including but not limited to Brazilian genres.However, the only available metadata is the genre information, and the dataset only comprises 1,008 recordings.

The Latin Music Database (LMD) proposed by Silla et al. [14] contains 3,160 pieces of music distributed across ten genres empirically selected by the authors and labeled by experts. The available labels represent a small amount of the existing genres.

In another effort to include Brazilian music in open datasets, the SAMBASET [15] comprises 493 Samba Enredo songs. However, the dataset is limited to that genre, and the metadata focuses on tempo identification and beat tracking.

With only Brazilian genres, The Brazilian Music Dataset (BMD) [16] has 210 songs grouped in 7 genres, some not existing in LMD. Also, with songs chosen empirically and not available, the set was developed to compare a model's performance in solving a classification problem using GTZAN [24] as a dataset with "global" genres and BMD as regional. Disregarding the question of GTZAN being a dataset with the global representation of genres, the amount of songs and genres existing in the BMD limits the scope of regional representation of the dataset.

Using the Brazilian portal Vagalume (c.f. Section 3), Lima et al.[5] propose a dataset with lyrics information labeled by artists and genres. The authors provide information on 138,368 Brazilian songs distributed in 14 genres.

Based on these limitations, we propose a new dataset using commercial music represented by acoustic features, lyrics, metadata, and categorical labels. Our dataset is composed of western music, including American and Latin music, with great emphasis on Brazilian artists and genres, commonly neglected in digital music research due to the absence of these genres in traditional datasets.

3 THE DATASET

The main source of data for 4MuLA is the Brazilian music portal Vagalume. This website, created in 2002, provides thousands of lyrics, links to YouTube videos, genre information, the similarity between songs and artists, among other data. The portal's users are responsible for creating and modifying this information, which implies a data bias that reflects the users' behavior and demands. Besides, the company has a public API[1] to grant access to all the data.

We make available our dataset in a public repository[2], with instructions for use, logs of changes, download links, metadata, and code for reproducing the acquisition process.

3.1 Data Acquisition

The data acquisition to build the 4MuLA begins with searching for all Vagalume artists and information about music. The Vagalume does not provide the audio of the songs. However, there is a YouTube video of the song on the lyrics page where we can access the audio. We make available some acoustic features, identifiers, and names of the artists and songs, lyrics, genres, related artists, categorical audio tags, and the artist's position in the Vagalume ranking.

The artists are grouped by genres or by the initial of their name. We chose to search using the initials to search for all artists, as we observed that the same artist might have an affinity with multiple genres. This search strategy using initials eliminated artists' redundancies and resulted in 177,685 artists.

For the initial version of our dataset, we chose to work with only songs by artists with some position in the ranking and have lyrics in Brazilian Portuguese, Portuguese, English, and Spanish. This selection criterion eliminated 117,247 artists who are not ranked and other 44,797 artists who do not have songs or their songs have different languages, not defined by us. The remaining 15,310 artists have a total of 96,458 songs that make up the dataset. The audio used in our database comes from a YouTube video the users associate with each song.

3.2 Features Extraction

Copyrights protect most of the songs in 4MuLA, so we can not provide the audio of each track in the data set. However, we acknowledge that different MIR tasks or algorithms may consider various features. Therefore, we provide a set of distinct features extracted from each recording. Specifically, we offer features that reveal different characteristics of the songs.

We extracted and made available 128-bins melspectrogram, 20 MFCCs, 12-bin Constant-Q chromagram, tonnetz, and tempogram. For all cases, we use the Python library Librosa [17], keeping the default parameter values. Besides, we extracted the 3 top tags using the neural network for music audio tagging musicnn [20]. For this purpose, we use 30 seconds of the recording, skipping its initial 15 seconds (as suggested in [19]). In addition to not disrespecting copyrights, we reduce the space needed to store the data and avoid the overhead for calculating the features to be used.

Still, recognizing that other researchers may need different features extracted from the audio, we can extract additional features on demand. In this case, we commit to making the new features available to increase our dataset, so other researchers can also take advantage of this new data.

[1] http://api.vagalume.com.br/docs/
[2] https://github.com/4mulaDataset/4mula

Finally, we emphasize that both the already available features and those extracted on demand can not rely on an exact inverse transform. In other words, we are committed to not providing features that can undergo the audio reconstruction.

In addition to the acoustic information, we have descriptive information that can also represent the songs. We make its identifier, title, language, lyrics, and a list of related songs available for each song. Regarding the artists, we also have their identifier, name, primary and related genres organized by similarity to the artist's style, related artists, and their position in a popularity ranking.

3.3 Possible Applications

Our dataset allows researching on varied tasks from this data and metadata, which give 4MuLA the multitask character. For instance, the genre information will enable us to study novel models to recognize music genres automatically. Other labels, such as the artist performing the song, can also be used to target classification tasks.

We can also develop regression and clustering techniques using 4MuLA. For example, the artist's ranking metadata can represent the target attribute for the regression task. Grouping music by content or metadata and analyzing the factors that led to such a partition can help gain insights into the relationship between music and artists.

Finally, the similarity between artists and songs is another metadata that serves as a basis for some tasks, such as creating machine learning models to estimate the similarity between these entities. Researchers interested in visualization or complex network analysis can also use the similarities included in 4MuLA.

We recall that every task associated with the 4MuLA, including those presented in this section, can potentially be performed using the lyrics, the audio features, or taking advantage of its multimodality.

4 EXPLORATORY DATA ANALYSIS

Each song in 4MuLA has up to six genres ordered by their relevance, assigned according to the artist's genres. For the sake of simplicity, we refer to the first of these labels as the main genre. To explore 4MuLA data, we start by analyzing the data based on the genres linked to each recording. In total, we have 76 different genres that label 15,310 artists and 96,458 songs. Figure 1 and Figure 2 show the number of artists and songs, respectively, in each genre, sorted by the number of occurrences.

In both cases, it is possible to observe the significant dominance of some genders. Notably, the charts' left part is dominated by Gospel/Religious, Sertanejo (a popular genre in the Brazilian countryside), Rock, Pop, and, curiously, Unknown genre. We can also notice that the number of unlabeled artists is very prominent in the first position (Figure 1). However, the number of Gospel/religious songs is much higher than any other genre (Figure 2). We can see that our dataset comprises several artists with unknown genres but with relatively few songs each. On the other hand, the gospel artists in our dataset have a large average number of songs.

Figure 3 reports the ranking score, regarding the relevance of the songs by each genre.

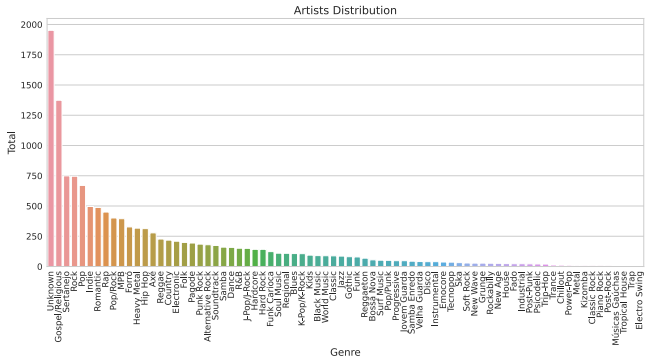


Figure 1: Number of artists in each genre.

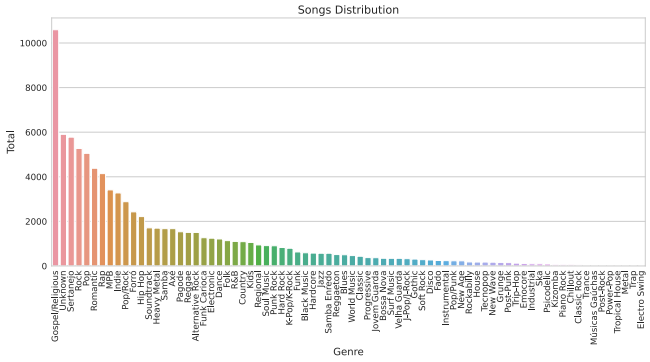


Figure 2: Number of songs in each genre.

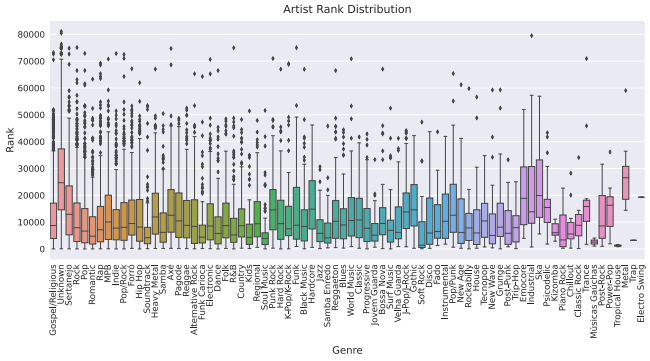


Figure 3: The artist rank distribution in each genre.

The presentation order of genres in Figure 3 follows the same sequence as shown in Figure 2. However, it is not possible to see a clear pattern between consecutive boxplots. This observation shows that the number of songs is not directly related to the relevance of each genre. For instance, the soft rock appears in the 47th position but has scores concentrated in lower values than religious music, in the first position.

Finally, we recall that we can represent our data by graphs where the vertices represent the artists (or the songs), and an edge is inserted between node A and B if B appears in the list of the most

related node to A or vice versa. Thereby, we can study its local characteristics, that is, how it relates to its neighbors. Table 1 shows information from the top ten artists with the highest degree in the artist graph, which average degree is 12.8618. The relative degree is the proportion of the degree in the whole graph.

Table 1: Summary of the nodes representing the ten artists with higher absolute degree.

Artist	absolute degree	relative degree	betweenness	closeness
Adele	1440	0.094	0.123	0.361
Luan Santana	989	0.065	0.050	0.339
Jorge e Mateus	820	0.054	0.034	0.333
Paula Fernandes	720	0.047	0.034	0.329
Glee	719	0.047	0.029	0.330
Katy Perry	676	0.044	0.021	0.329
Bruno Mars	592	0.039	0.022	0.322
Rihanna	567	0.037	0.017	0.323
Lady Gaga	537	0.035	0.017	0.324
Gusttavo Lima	525	0.034	0.015	0.317

This type of analysis may lead to the observation of some interesting factors. For example, recall Figures 1 and 2. In Sertanejo, which appears with a large number of artists and songs, it is common for a few artists to gain much fame and others not to be very successful, which can be seen by the wide distribution of their ranks in Figure 3. Consequently, these less successful artists are commonly compared with more prominent artists, causing the famous ones to have a higher degree in the similarity graph. This observation explains because Table 1 presents several artists of this genre: Luan Santana, Jorge e Mateus, Paula Fernandes, and Gusttavo Lima.

5 FINAL REMARKS

As an alternative to music data limitations for research, this work presents a new dataset with musical information in different modalities with implications in various tasks and applications. We also described and explored the organization of data, revealing some of its possibilities.

As future work, we intend to establish strong baselines for the tasks we can perform using 4MuLA. For this, we will apply well-known algorithms for each task and make available the results in 4MuLA's repository and scientific publications. Besides, we will maintain the evolution of 4MuLA. Every new version of the dataset will create a separate subset to avoid overwriting stable versions and, consequently, maintain the reproducibility property.

REFERENCES

[1] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*. ISMIR, Miami, USA, 591–596.

[2] Dmitry Bogdanov, Alastair Porter, Hendrik Schreiber, Julián Urbano, and Sergio Oramas. 2019. The AcousticBrainz genre dataset: Multi-source, multi-level, multi-label, and large-scale. In *International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, 360–367.

[3] Dmitry Bogdanov, Minz Sanghee Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The MTG-Jamendo dataset for automatic music tagging.

[4] Keunwoo Choi and Kyunghyun Cho. 2019. Deep unsupervised drum transcription. In *International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, 183–191.

[5] Raul de Araújo Lima, Rômulo César Costa de Sousa, Simone Diniz Junqueira Barbosa, and Hélio Cortês Vieira Lopes. 2020. Brazilian Lyrics-Based Music Genre Classification Using a BLSTM Network. arXiv:2003.05377 [cs.CL]

[6] Reinier de Valk, Ryaan Ahmed, and Tim Crawford. 2019. JOSQUINTAB: a dataset for content-based computational analysis of music in lute tablature. In *International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, 431–438.

[7] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2017. FMA: A Dataset For Music Analysis. In *International Society for Music Information Retrieval Conference*. ISMIR, Suzhou, China, 316–323.

[8] Felipe Falcao, Baris Bozkurt, Xavier Serra, Nazareno Andrade, and Ozan Baysal. 2019. A dataset of rhythmic pattern reproductions and baseline automatic assessment system. In *International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, 439–445.

[9] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. 2002. RWC Music Database: Popular, Classical, and Jazz Music Databases. In *International Society for Music Information Retrieval Conference*. ISMIR, Paris, France, 287–288.

[10] Helge Homburg, Ingo Mierswa, Bulent Moller, Katharina Morik, and Michael Wurst. 2005. A Benchmark Dataset for Audio Classification and Clustering. In *International Society for Music Information Retrieval Conference*. ISMIR, UK, 528–531.

[11] Cheng-Zhi Anna Huang, Curtis Hawthorne, Adam Roberts, Monica Dinculescu, James Wexler, Leon Hong, and Jacob Howcroft. 2019. The Bach Doodle: Approachable music composition with machine learning at scale. In *International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, 793–800.

[12] Berit Janssen, Tom Collins, and Iris Yuping Ren. 2019. Algorithmic ability to predict the musical future: datasets and evaluation. In *International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, 208–215.

[13] Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Kyogu Lee, and Juhan Nam. 2019. VirtuosoNet: a hierarchical RNN-based system for modeling expressive piano performance. In *International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, 908–915.

[14] Carlos N. Silla Jr, Alessandro L. Koerich, and Celso A.A. Kaestner. 2008. The Latin Music Database. In *Proceedings of the 9th International Conference on Music Information Retrieval*. Lulu.com, Philadelphia (USA), 451–456. https://kar.kent.ac.uk/24000/

[15] Lucas S Maia, Magdalena Fuentes, Luiz WP Biscainho, Martín Rocamora, and Slim Essid. 2019. SAMBASET: a dataset of historical samba de enredo recording for computational music analysis. In *International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, 628–635.

[16] J. Martins de Sousa, E. Torres Pereira, and L. Ribeiro Veloso. 2016. A robust music genre classification approach for global and regional music datasets evaluation. In *2016 IEEE International Conference on Digital Signal Processing (DSP)*, Vol. 21. IEEE, Beijing (China), 109–113.

[17] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, Vol. 8. Scipy, Austin, Texas, 18–24. https://librosa.github.io/librosa/

[18] Oriol Nieto, Matthew McCallum, Matthew EP Davies, Andrew Robertson, Adam Stark, and Eran Egozy. 2019. The harmonix set: beats, downbeats, and functional segment annotations of western popular music. In *International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, 565–572.

[19] Elias Pampalket. 2006. *Computational models of music similarity and their application in music information retrieval*. Ph.D. Dissertation. Vienna University of Technology, Vienna, Austria. http://www.ofai.at/{~}elias.pampalk/publications/pampalk06thesis.pdf

[20] Jordi Pons and Xavier Serra. 2019. musicnn: Pre-trained convolutional neural networks for music audio tagging. arXiv:1909.06654 [cs.SD]

[21] Zhengshan Shi, Craig Stuart Sapp, Kumaran Arul, Jerry McBride, and Julius O Smith III. 2019. SUPRA: digitizing the Stanford University piano roll archive. In *International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, 517–523.

[22] Eduardo Simas Filho, Elmo Borges Jr., and Antonio Fernandes Jr. 2018. Genre Classification for Brazilian Music using Independent and Discriminant Features. *Journal of Communication and Information Systems* 33, 1 (May 2018), 104–112. https://doi.org/10.14209/jcis.2018.11

[23] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. 2019. AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing. In *International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, 501–510.

[24] George Tzanetakis and Perry Cook. 2002. Musical Genre Classification of Audio Signals. *IEEE Transactions on Speech and Audio Processing* 10 (08 2002), 293 – 302. https://doi.org/10.1109/TSA.2002.800560

[25] Furkan Yesiler, Chris Tralie, Albin Andrew Correya, Diego F Silva, Philip Tovstogan, Emilia Gómez Gutiérrez, and Xavier Serra. 2019. Da-TACOS: A dataset for cover song identification and understanding. In *International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, 327–334.