The Greek Audio Dataset

Dimos Makris, Katia L. Kermanidis, and Ioannis Karydis

Dept. of Informatics, Ionian University, Kerkyra 49100, Greece
{c12makr,kerman,karydis}@ionio.gr

Abstract. The Greek Audio Dataset (GAD), is a freely available collection of audio features and metadata for a thousand popular Greek tracks. In this work, the creation process of the dataset is described together with its contents. Following the methodology of existing datasets, the GAD dataset does not include the audio content of the respective data due to intellectual property rights but it includes MIR important features extracted directly from the content in addition to lyrics and manually annotated genre and mood for each audio track. Moreover, for each track a link to available audio content in YouTube is provided in order to support researchers that require the extraction of new feature-sets, not included in the GAD. The selection of the features extracted has been based on the Million Song Dataset in order to ensure that researchers do not require new programming interfaces in order to take advantage of the GAD.

Keywords: music, information retrieval, audio, Greek music, dataset

1 Introduction

Music Information Retrieval (MIR) research, in other words research on methods for Information Retrieval and Data Mining on musical data, has a number of requirements. Central to these requirements is the necessity to experiment with the methods on real musical data. This experimentation mainly achieves testing on the efficiency and effectiveness of the methods, as well as comparison of existing existing methods in order to show improvement.

In MIR research, the term musical data refers to sound recordings, sheet music as well as associated information to the musical content (i.e. metadata, social tags, etc). The process of accumulating a set of musical data (dataset) is very important for most scientific processed, including MIR. Datasets allow researchers to scientifically to compare and contrast their methods by testing these on commonly available collection of musical works.

Accordingly, MIR requires data for all kinds of music as results are not always intuitive due to the highly artistic nature of music. Although a number of widely used datasets do exist for the purposes of MIR research, most of these are collections of mainstream English language music. Local music has numerous differences from the these musical collections such us different instruments and rhythms. To the best of our knowledge, a standardized Greek audio dataset is not currently available.

1.1 Motivation and Contribution

To address the aforementioned requirements, we introduce the Greek Audio Dataset, a freely-available collection of Greek audio data for the purposes of MIR that, for each song it contains, offers:

- audio features for immediate use in MIR tasks,
- the lyrics of the song,
- manually annotated mood and genre labels,
- a link to YouTube for further feature extraction, on the basis of MIR research processes.

The rest of the paper is organised as follows: Section 2 presents the related work on MIR available datasets, while Section 3 discuses the dataset, its creation processes as well as a detailed analysis of its content. Next, Section 4 details future directions concerning the dataset that could ameliorate is usability and further support MIR research. Finally the paper is concluded in Section 5.

2 Related Research

Since the early years of the MIR research, there have been numerous efforts to build datasets facilitating the work of MIR researchers.

The dataset titled RWC [4] is a copyright-cleared music database that is available to researchers as a common foundation for research. RWC was one of the first large-scale music database containing six original collections on different genres.

CAL500 [19] is another widely used dataset, a corpus of 500 tracks of Western popular music each of which has been manually annotated by at least three human labelers with a total number of 1700 human-generated musical annotations.

Tzanetakis and Cook [20], introduced a dataset under the name GTZAN Genre Collection. The dataset consists of 1000 audio tracks, each 30 seconds long. It also contains 10 genres, each represented by 100 tracks and its size is approximately 1.2 Gb. The dataset was collected gradually with no titles from a variety of sources including personal CD collections or radio and microphone recordings, while it doesn't contain any music information or metadata.

USPOP2002 [1] was introduced in order to perform comparison between different acoustic-based similarity measurements. 400 artists were chosen for popularity and for representation, while overall, the dataset contains 706 albums and 8764 tracks.

The Swat10k [18] data set contains 10,870 songs that are weakly-labeled using a tag vocabulary of 475 acoustic tags and 153 genre tags. These tags have all been harvested from Pandora [13] result from song annotations performed by expert musicologists involved with the Music Genome Project.

The Magnatagatune dataset features human annotations collected by Edith Law's TagATune game [9] and corresponding sound clips encoded in 16 kHz/

32kbps /mono/mp3 format. It also contains a detailed analysis from The Echo Nest of the track's structure and musical content, including rhythm, pitch and timbre.

Schedl et al. [15], presented the MusiClef dataset, a multimodal data set of professionally annotated music. MusiClef contains editorial meta-data, audio features (generic and MIR oriented) are provided. Additionally MusiClef includes annotations such as collaboratively generated user tags, web pages about artists and albums, and the annotation labels provided by music experts.

Finally the Million Song Dataset (MSD) [2] stands out as the largest currently available for researchers with 1 million songs, 44,745 unique artists, 280 GB of data, 7,643 unique terms containing acoustic features such as pitch, timbre and loudness and much more.

Table 1 (adapted from [2]) collectively presents the aforementioned datasets by comparison of size and samples/audio availability.

dataset	# songs	includes samples/audio
CAL500	500	No
Greek Audio Dataset (GAD)	1000	No (YouTube links available)
GZTAN genre	1000	Yes
Magnatagatune	25863	Yes
Million Song Dataset (MSD)	1000000	No
MusiCLEF	1355	Yes
RWC	465	Yes
Swat10K	10870	No
USPOP	8752	No

Table 1. Comparison of existing datasets: size and samples/audio availability.

3 The Dataset

The musical tradition of Greece is diverse and celebrated through its history. Greek music can be separated into two major categories: Greek traditional music and Byzantine music, with more eastern sounds [7]. Greek traditional (folk) music or "*δημοτική μουσική*", as it is most commonly referred to, is a combination of songs, tempos and rhythms from a litany of Greek regions. These compositions, in their vast majority, have been created by unknown authors, are more than a century old and are the basis for the Modern Greek traditional music scene. GAD contains many Greek traditional music songs, starting from the 60's and 70's as well as some current variants that are very popular nowadays.

Greek music contains special characteristics that are not easily to be found on the existing datasets. The unique instruments (bouzouki, lyra, laouto, etc) and the unique rhythms reflect on the audio feature extraction while the complexity of the Greek language can be an area of study for linguistic experts on lyric feature extraction.

3.1 Creation process

The audio data collection covers the whole range of Greek music, from traditional to modern. We made every effort to make the song and genre selection balanced in GAD as much as possible. The selection of the music tracks was made from personal CD collections, while some of the songs' recordings were available in live performances. The audio feature extraction process was applied on CD quality wave files (44,1KHz, 16 bit) using jAudio [11] and AudioFeatureExtraction, a processing tool introduced by the MIR research team of Vienna University of Technology[21].

For each song, the GAD additionally includes its lyrics. The lyrics included in the dataset have been retrieved among various sources, such as the web site stixoi.info [6], and are used within for the purposes of academic and scientific research of MIR researchers. It should be noted that copyright of the lyrics remains with their respective copyright owners.

As far as genre classification is concerned, the following Greek musical culture oriented tags where used: Rembetiko, Laiko, Entexno, Modern Laiko, Rock, Hip Hop/R & B, Pop, Enallaktiko, that are described in the sequel (Section 3.2). Because of the different music styles that artists may adopt during their career, we performed listening tests to every song before choosing the appropriate genre tag. For mood annotation, the Thayer model [17] has been adopted. To record the mood categories, we invited 5 annotators to listen and read the lyrics for one couple and refrain for each song. Then we computed a standard F-measure as a measure of the inter-annotator agreement [3] and we found a value of 0.8 approximately. For clusters of mood with smaller F-measure, a discussion between the annotators was taken part in order to reduce controversy.

The dataset additionally includes a YouTube [22] link for each song that allows for access to the audio content. In order to identify the best match of the YouTube available alternative songs, for every song we attempted to identify the YouTube URL using the following criteria: number of views, number of responses, best audio quality and as close as possible performance as in the wave library where the features were extracted from.

The GAD is available as a download from the webpage of the Informatics in Humanistic and Social Sciences Lab of the Ionian University[1].

3.2 The content

The GAD contains audio and lyrics features and metadata for selected Greek Songs:

- 1000 songs
- 1000 txt files with the lyrics
- 1000 best YouTube links
- 277 unique artists

[1] http://di.ionio.gr/hilab/gad

– 8 Genres. We give the necessary explanations to distinguish the unique Greek genres.
  - Ρεμπέτικο: 65 tracks. In urban centers with strong Greek presence, among other genres, it appeared a kind of folk called "Rembetiko". Major schools of Rembetiko were: Smyrnaiiki and the Piraeus school classic Rembetiko [12]. "Zeibekiko", "karsilamas" and "hasapiko" are a few characteristic rhythms.
  - *Λαικό*: 186 tracks. As an evolution of Rembetiko are considered Greek Folk songs of decades of 1950-1960, which continues to evolve and sound today [10]. The transition to the so-called "Laiko" music is evident in the imposition of European instrument tuning, rhythms and now that the author can write songs with "harmony".
  - Εντεχνο: 195 tracks. A complex musical work of art that combines Modern Greek music with poetry [16]. It differs from the Laiko mainly in verse, but also in music (instrumentation, style).
  - Μοντέρνο *Λαικό*: 175 tracks. The "Modern Laiko" is considered as the current evolution of popular music. Several pop elements, electronic sounds can be identified and is now the most common music heard in Greek live stages. The style and the rhythms have not changed much and the theme of the songs is adapted to the current daily problems.
  - Rock: 195 tracks. It also includes 80s Pop-Rock.
  - Hip Hop/R & B: 60 tracks.
  - Pop: 63 tracks. It also includes Dance-Club music style and older Greek disco hits.
  - Εναλλακτικό: 60 tracks. Although with the term "Enallaktiko" it is usually considered as "Alternative Rock" [14], we include tracks fusing Greek modern kinds of music. Pop Rock and Entexno elements can be found in this class.
– 16 Mood taxonomies: 2 dimensions, valence and arousal, which divide a 2-dimensional emotive plane into 4 parts by having positive/high and negative/low values respectively. Arousal and valence are linked to energy and tension, respectively. Arousal values correspond to moods such as "angry" and "exciting" to "tired" and "serene" from 1-4. Valence values correspond to moods such as "sad" and "upset" to "happy" and "content" from A-D. Figure 1 shows the distribution of tracks by mood and we observe that the tracks with positive emotions (valence C, D), have variations in arousal and either are very calm (arousal 1) or intense stress (arousal 4). We also include all the lyrics for further feature extraction. The accumulated lyrics contain:
  - 32024 lines
  - 143003 words
  - 1397244 characters

The data is available in two formats, HDF5 and CSV.

HDF5 is efficient for handling the heterogeneous types of information such as audio features in variable array lengths, names as strings, and easy for adding

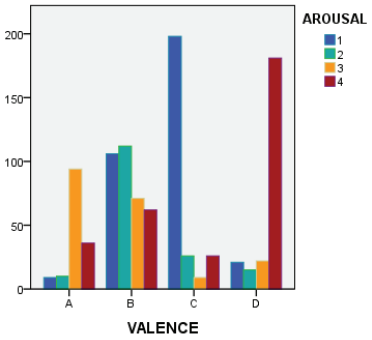

Fig. 1. Valence and Arousal on the GAD.

new types of features. Every song is described by a single file the contents of which are as shown in Figure 2. Each file has 2 groups: the folder "Audio Features" containing the three most important types of features, such as Timbral, Rhythm and Pitch [20] and the "Metadata" containing all the other information that is related to the song (i.e. title, artist, mood, genre and YouTube link).

CSV is compatible for processing with Weka [5], RapidMiner [8] and other similar data mining platforms. The GAD provides the commonly used, on the discipline of MIR, audio feature sets in separate CSV files. Weka, besides the well-known options for preprocessing and classification, additionally offers attribute visualization. Figure 3 shows one such mapping between two timbral characteristics with sorting class the genre and more specifically, Rock and Entexno.

The main acoustic features are Timbral, Rhythm and Pitch. Most of the features were extracted using jAudio which also calculates derived features. These features are created by applying metafeatures to primary features.

- Timbral Texture Features: features used to differentiate mixture of sounds based on their instrumental compositions when the melody and the pitch components are similar. (251 features)
  - FFT: Spectral Centroid, Flux, Rolloff, Low Energy, Zero Crossings, Compactness, Spectral Variability, Root Mean Square. 16 features
  - MFCC: 26 features
  - Spectrum: Power & Magnitum Spectrum. 199 features
  - MoM: Method Of Moments. 10 features
- Rhythm Features: Rhythmic content features are used to characterize the regularity of the rhythm, the beat, the tempo, and the time signature. These features are calculated by extracting periodic changes from the beat histogram. (63 features)
  - Beat and Freq: Strongest beat, sum, strength + Strongest FREQ, Zero spectral, FFT + Peak Finder. 13 features

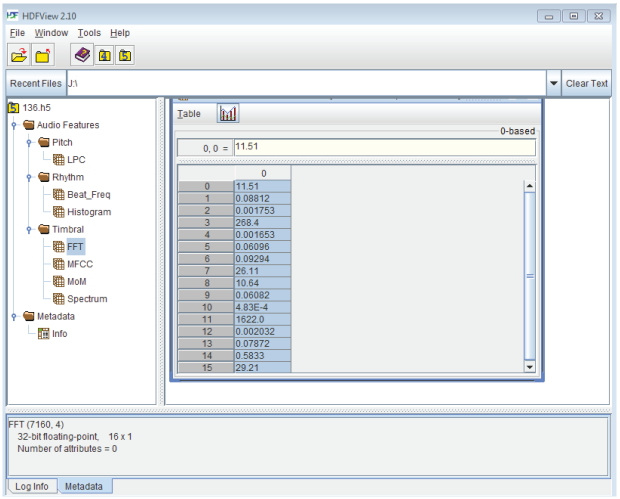

Fig. 2. View of a single track using HDFView.

- Beat Histogram. 60 features

– Pitch Content Features: The pitch content features describe the distribution of pitches. (18 features)
  - LPC: Linear Predictive Coding for voice recognition.

4 Future Direction of the Dataset

The GAD is not without issues that can be ameliorated in future versions. One of these issues pertains to the balance of the distributions of genre types in the data. In addition, classification of Greek music into genres can be confusing when use of only one tag is assumed. An extension of the dataset by addition of more data as well as a multi-label version of the dataset will eliminate those problems and it is already under development.

Some of the main future actions that would greatly enhance the GAD, to name a few, are:

– the inclusion of user generated tags (from tagging games or web-services),
– the collection of labels for mood and genre based on more users,
– the expansion of the number of songs (i.e. include latest top-chart songs),
– the refinement of genres by adding more detailed labels with descriptions,
– the balancing of moods and/or genres,
– the inclusion of scores for each song,
– the development of programming language wrappers.

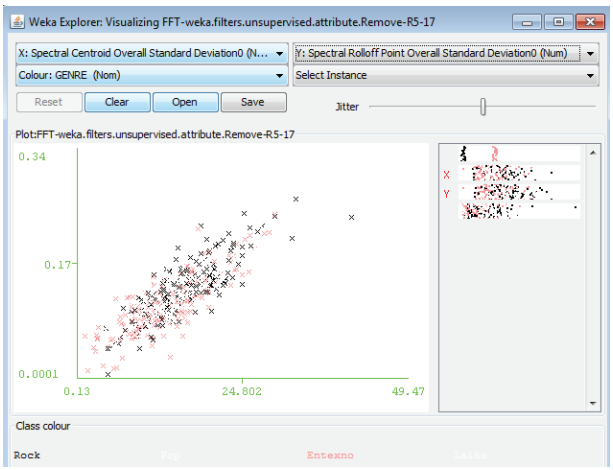

Fig. 3. Visualizing Centroid and Rolloff for genres Rock and Entexno.

5 Conclusion

The Greek Audio Dataset is, to the best of the authors' knowledge, the first complete attempt to create an annotated Greek audio dataset. Although its size is relatively small compared to latest datasets, it constitutes a good start for researchers who want to study how the Greek music reflects in the context of MIR. Following the methodology of the, ubiquitous in MIR research, Million Song Dataset, the GAD does not include the audio content and it uses the HDF5 format to store all the information. This makes it easy for any future enrichment with more features and musical information as well as for the utilisation of existing methodologies in order to access the data. In order to enhance further the dataset, it also contains lyrics information so as to provide for linguistic methods on MIR. In addition, the dataset is also available in CSV format that makes it very easy to explore it in data mining experiments using widely available software as WEKA and RapidMiner.

References

1. Berenzweig, A., Logan, B., Ellis, D.P.W., Whitman, B.P.W.: A large-scale evaluation of acoustic and subjective music-similarity measures. Comput. Music J. 28(2), 63–76 (2004)
2. Bertin-Mahieux, T., Ellis, D.P., Whitman, B., Lamere, P.: The million song dataset. In: Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011) (2011)
3. Boisen, S., Crystal, M., Schwartz, R.M., Stone, R., Weischedel, R.M.: Annotating resources for information extraction. In: LREC (2000)

4. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: Rwc music database: Popular, classical, and jazz music databases. In: Proceedings of 3rd International Conference on Music Information Retrieval. pp. 287–288 (2002)
5. Holmes, G., Donkin, A., Witten, I.H.: Weka: a machine learning workbench. pp. 357–361 (1994)
6. stixoi info: Greek lyrics for songs and poetry, http://www.stixoi.info/
7. Institute for research on music and acoustics: Greek traditional music, http://www.musicportal.gr/greek_traditional_music/
8. Jungermann, F.: Information extraction with rapidminer. In: Proceedings of the GSCL Symposium 'Sprachtechnologie und eHumanities'. pp. 50–61 (2009)
9. Law, E., von Ahn, L.: Input-agreement: A new mechanism for collecting data using human computation games. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1197–1206 (2009)
10. Liavas, L.: The greek song: from 1821 to the 1950s. Emporiki Bank Of Greece (2009)
11. Mcennis, D., Mckay, C., Fujinaga, I., Depalle, P.: jaudio: An feature extraction library. In: Proceedings of the 6th International Conference on Music Information Retrieval (2005)
12. Ordoulidis, N.: The greek laiko (popular) rhythms: Some problematic issues. In: Proceedings 2nd Annual International Conference on Visual and Performing Arts (2011)
13. Pandora: A free personalized internet radio, http://www.pandora.com/
14. di Perna, A.: Brave noise-the history of alternative rock guitar. Guitar World (1995)
15. Schedl, M., Liem, C.C., Peeters, G., Orio, N.: A Professionally Annotated and Enriched Multimodal Data Set on Popular Music. In: Proceedings of the 4th ACM Multimedia Systems Conference (2013)
16. Sideras, A.: The sung poetry. Musicology 3, 89–106 (1985)
17. Thayer, R.: The biopsychology of mood and arousal. Oxford University Press, USA (1989)
18. Tingle, D., Kim, Y.E., Turnbull, D.: Exploring automatic music annotation with "acoustically-objective" tags. In: Proceedings of the International Conference on Multimedia Information Retrieval. pp. 55–62 (2010)
19. Turnbull, D., Barrington, L., Torres, D., Lanckriet, G.: Towards musical query-by-semantic-description using the cal500 data set. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 439–446 (2007)
20. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. Speech and Audio Processing, IEEE Transactions on 10(5), 293–302 (2002)
21. Vienna University of Technology: Audio feature extraction web service, http://www.ifs.tuwien.ac.at/mir/webservice/
22. YouTube: Share your videos with friends, family and the world, http://www.youtube.com/