The Greek Music Dataset

Dimos Makris
Department of Informatics,
Ionian University
Corfu, 49100
Greece
c12makr@ionio.gr

Ioannis Karydis
Department of Informatics,
Ionian University
Corfu, 49100
Greece
karydis@ionio.gr

Spyros Sioutas
Department of Informatics,
Ionian University
Corfu, 49100
Greece
sioutas@ionio.gr

ABSTRACT

Music Information Research (MIR) requires musical data in order to test methods and to compare results. Greek music presents a number of unique characteristics that make its musical pieces distinct from popular tracks existing in currently available datasets, leading thus to the MIR requirement of Greek datasets. This work presents the Greek Music Dataset (GMD), a collection of musical information pertaining to Greek musical pieces. GMD is a significant extension of the Greek Audio Dataset by addition of symbolic information, both features and raw MIDI files, inclusion of multi-label manual genre categorisation of the content as well as by extension of the included tracks and balancing of the content in terms of genre. GMD includes information for 1400 Greek tracks, while for each track, the dataset includes pre-computed audio, lyrics & symbolic features for immediate use in MIR tasks, manually annotated labels pertaining to mood & genre styles of music, generic objective metadata, a manually selected MIDI file (available for 500 of the tracks) and a manually selected link to a performance / audio content in YouTube for further research.

CCS Concepts

•**Information systems** → **Test collections;** *Question answering; Clustering and classification;* •**Computing methodologies** → Feature selection;

Keywords

information retrieval, data mining, audio, symbolic, genre, mood, Greek music, dataset

1. INTRODUCTION

Music Information Research (MIR), in other words research on methods for Information Retrieval and Data Mining on musical data, has a number of requirements. The requirement to experiment with the methods on real musical data is central. In MIR, the term musical data refers to sound recordings, sheet music, lyrics as well as associated information to the musical content (i.e. metadata, social tags, etc). Although a number of widely used datasets do exist for the purposes of music information research, most of these are collections of mainstream English language music. In addition, most existing datasets concentrate on providing feature sets extracted from audio data while available lyric and symbolic datasets are scarce.

The musical tradition of Greece is diverse and celebrated through its history. The Greek traditional music (folk) contains musical, rhythmic, structural and literary characteristics which are unknown on the rest of the musical world [13]. Genres like “ρεμπέτικο”, “λαϊκό” and “έντεχνο”, which form the backbone of Greek music, are unique and there are currently very little contextual information for that kind of content in the MIR field. Our attempt to make Greek music available to MIR researchers does not start from scratch, as it is a continuation and extension of the Greek Audio Dataset [18], a freely available collection of audio features and metadata for a thousand popular Greek tracks.

1.1 Motivation and Contribution

To address the aforementioned requirements, we introduce the Greek Music Dataset (GMD), a freely available collection of features and metadata for 1400 popular Greek tracks, following the original Greek Audio Dataset as an extension. For each song GMD contains:

- pre-computed audio, lyrics & symbolic features for immediate use in MIR tasks,
- manually annotated labels pertaining to mood & genre styles of music,
- metadata,
- a manually selected MIDI file for the track (currently available for 500 of the tracks),
- a manually selected link to a performance / audio content in YouTube is provided for further research, following the methodology of existing datasets, not to include content from the respective audio data due to intellectual property rights.

The rest of the paper is organized as follows: Section 2 presents the related work on MIR available datasets, while Section 3 details the dataset, its creation processes as well as an analysis of Greek Tradition music. Next, Section 4 presents a detailed analysis of its content with all the available features sets and metadata. Finally the paper is concluded in Section 5.

16th EANN workshops, September 25 - 28, 2015, Rhodes Island, Greece
© 2015 ACM. ISBN 978-1-4503-3580-5/15/09…$15.00
DOI: http://dx.doi.org/10.1145/2797143.2797175

2. RELATED RESEARCH

Since the early years of the MIR, there have been numerous efforts to build datasets facilitating the work of music information researchers. Due to the inherent subjectivity of music perception, despite the largely diverse information of importance to MIR process, there are no generally accepted standards for tags describing genre and mood categories. Moreover, the lack of widely available manually annotated musical scores complicates the construction of datasets containing the composer's high level features, which can be derived exlusively from symbolic data.

Mainly due to the abovementioned reasons, as well as the popularity of the content-based similarity MIR process, there has been a tendency to create datasets mostly focused on audio collections. The difficulties to collect ground truth data for lyrics and symbolic representations is yet another reason preventing researchers to construct such datasets in structured form. Subsequently, many datasets avoid containing unprocessed music data while focusing on metadata and information that has been extracted directly from the performance audio and are thus immediately ready for use by researchers in the MIR field.

One of the very popular datasets is under the name "GTZAN Genre Collection" as introduced by Tzanetakis and Cook [28]. The dataset consists of 1000 audio tracks, each 30 seconds long. It contains 3 audio genre hierarchies, and its size is approximately 1.2 Gb. The authors also introduced one of the first speech music datasets for the purposes of music/speech discrimination. The dataset consists of 120 tracks, each 30 seconds long.

RWC [8] dataset is a copyright-cleared music database that is available to researchers as a common foundation for research. It provides audio samples together with extensive metadata and is well suited for the evaluation of many kinds of audio processing tasks. RWC was one of the first large-scale music database containing four original collections on different genres. It is one of the few datasets that is based on the originally-recorded music compact discs, standard MIDI files, and text files of lyrics. Unfortunately the size of the dataset is currently considered small and it does not contain any further metadata.

The Swat10k [26] 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 [1] result from song annotations performed by expert musicologists involved with the Music Genome Project.

USPOP2002 [3] 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 dataset does'nt include any music data due to intellectual property rights, but only some audio MFCC features. Hu et al. [11] collected social tags of single adjective words from last.fm for this particular dataset, and manually selected 19 mood related terms which then reduced to three latent mood categories using multidimensional scaling.

CAL500 [27] 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. For each song the dataset provides various features that have been extracted from the audio.

The Magnatagatune dataset features human annotations collected by Law's TagATune game [16]. The deployed version of TagATune is an instantiation of the input-agreement mechanism. The dataset also contains a detailed analysis from "The Echo Nest" of each track's structure and musical content, including rhythm, pitch and timbre.

Schedl et al. [24], presented the MusiClef dataset, a multimodal data set of professionally annotated music. MusiClef contains editorial meta-data, while audio features (generic and MIR oriented) are also provided. Additionally, MusiClef includes annotations such as collaboratively generated user tags, web pages about artists and albums and annotation labels provided by music experts.

Finally the Million Song Dataset (MSD) [4], a freely-available collection of audio features and metadata for a million contemporary popular music tracks stands out as the largest currently available such dataset for researchers. It contains 44,745 unique artists, 280 GB of data, 7,643 unique terms containing acoustic features and much more.

Table 1 (adapted from [4]) collectively presents the aforementioned datasets by comparison of size, lyrics data, symbolic data, metadata and samples/audio availability.

3. THE DATASET

The construction of the first Greek Music Dataset did not start from scratch. An early version of GMD exists when drafted and used as the first attempt for a collection of audio features and metadata (single label tags concerning music genre and mood) for a thousand popular Greek tracks [18]. Unfortunately no other information or metadata where available in that dataset. In cotrast, GMD contains many Greek traditional music songs, starting from the 1940's and 50's as well as some current variants that are very popular nowadays.

3.1 Greek Music and its importance

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 containing more eastern sounds [22]. Greek traditional (folk) music or "*δημοτική μουσική*" in Greek, as it is most commonly referred to, is a combination of songs, tempos and rhythms from a litany of Greek regions. The traditional folk music is a creation of closed rural communities and reflects the experiences, ideas and feelings of the local people. Commonly it is characterized by the area in which the composers originate ("*νησιώτκα*", "*ηπειρώτικα*", and "*ποντιακά*") and often are related to the music of neighbouring communities. Except for the regions, traditional songs and tunes can be classified based on their content and the circumstances under which played (carnival or marriage celebrations, emigration, etc.) or musical, rhythmic, structural and literary characteristics ("*στροφικά*", "*επτασύλλαβα*", "*καρσιλαμάδες*").

Greek music contains special characteristics that are not easily to be found on the existing datasets. Unusual chord progressions, pentatonic structures, unique traditional instruments (bouzouki, lyra, laouto, etc) and the unique rhythms (9/8 time signatures) reflect significantly on MIR processes and especially on feature extraction. The co-existence of audio and MIDI files as well as high level symbolic features will hopefully provide for musicologists seeking for available

[1] http://www.pandora.com/

dataset	# songs	audio & features	lyrics & features	symbolic & features	metadata
CAL500	500	No/Yes	No/Yes	No/No	No
Greek Audio Dataset (GAD)	1000	No (see notes)/Yes	Yes/No	No/No	Yes
Greek Music Dataset (GMD)	1400	No (see notes)/Yes	Yes/Yes	Yes/Yes	Yes
GZTAN genre	1000	Yes/No	No/No	No/No	No
Magnatagatune	25863	Yes/Yes	No/No	No/No	Yes
Million Song Dataset (MSD)	1000000	No/Yes	No/No	No/No	Yes
MusiCLEF	1355	Yes/Yes	No/No	No/No	Yes
RWC	465	Yes/No	Yes/No	Yes/No	Yes
Swat10K	10870	No/Yes	No/No	No/No	Yes
USPOP	8752	No/Yes	No/No	No/No	Yes

Notes. GAD & GMD although not providing audio content, do include YouTube links.

Table 1: Comparison of existing datasets: size, lyrics data, symbolic data, metadata and samples/audio availability.

and annotated research data. Finally the complexity of the Greek language and the subjects of the songs can be an area of study for linguistic experts on lyrics feature extraction.

3.2 Gathering the Content

The audio data collection covers a broad range of Greek music, from traditional to modern. Compared with the first version of the dataset we removed 100 songs and added 500 new songs, in addition to the corresponding MIDI files. The changes in tracks were made in order to make the song and genre selection balanced in GMD as much as possible. The new songs that were added, mainly belong to the traditional Greek music genres (“ρεμπέτικο”, “λαϊκό” and “έντεχνο”).

The dataset includes a manually selected for its quality YouTube[2] 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.

The MIDI files were collected from various sources in addition to the free Greek Midi database[3]. These were preprocessed and checked manually for the music & performance’s precise correspondence to the audio content selected from YouTube. However in some cases, different instrument tunings or tempos were detected in comparison to actual audio performance. Thus, the preprocessing stage was included in order to compensate for MIR process pertaining to MIDI alignment [21].

Finally for each song, the GMD additionally includes its lyrics. The lyrics included in the dataset have been retrieved among various sources, such as the web site stixoi.info[4], and are used within for the purposes of academic and scientific research of MIR researchers. It should be noted that copyright of the symbolic data as well as lyrics remain with their respective copyright owners.

3.3 Genre and Mood tags

The genre and mood annotation was a tedious and tricky effort as it was manually performed. The early version of GMD includes single tag mood and genre annotation. For this current version we include single tag mood annotation and multi label annotation for genre music style.

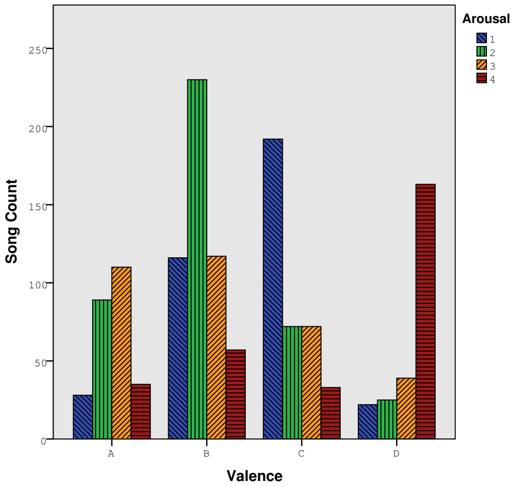

Figure 1: Valence and Arousal distribution on GMD.

3.3.1 Mood Annotation

The procedure is the same for the new songs as it was done in previous works [18]. A group of annotators were invited and were annotating the same song the same time. They were listening and reading the lyrics for each song. For mood modeling and mapping, the model of Thayer [25] was adopted. In this model there exist 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. Annotations are deemed reliable if annotators agree sufficiently for relevant purposes, that is they consistently make the same decisions. The reliability of an annotation was calculated as a measure of the inter-annotator agreement [5]. Figure 1 shows the distribution of tracks by mood and it may be observed that tracks with positive emotions (valence C, D), have variations in arousal and either are very calm (arousal 1) or intense stress (arousal 4).

3.3.2 Genre Annotation

[2] http://www.youtube.com/

[3] http://www.greekmidi.com/

[4] http://www.stixoi.info/

Imitating the procedure of single label mood annotation we invited three students of the Music Department of the Ionian University and two sound producers to annotate. The annotators had to select for every song one or more of the 8 available tags of GMD. The genre tags are the same on the early version of the dataset and in the sequel the necessary explanations to distinguish the unique Greek genres are provided.

- “ρεμπέτικο”: In urban centers with strong Greek presence, among other genres, this genre appeared as kind of folk. Major schools of “ρεμπέτικο” were: Smyrnaiiki and the Piraeus school classic Rembetiko [22].
- “λαϊκό”: As an evolution of “ρεμπέτικο” are considered Greek Folk songs of decades of 1950-1960, which continue to evolve and are listened to even today. The transition to the so-called “λαϊκό” music is evident in the imposition of European instrument tuning, rhythms and now that the author can write songs with *harmony*.
- “έντεχνο”: A complex musical work of art that combines modern Greek music with poetry [14]. It differs from the “λαϊκό” mainly in verse, but also in music (instrumentation, style).
- “Modern λαϊκό”: During 1980’s the influence of western music is adopted from artists of “λαϊκό music. It is considered as the current evolution of Greek popular music. The style and the rhythms have not changed much and the theme of the songs is adapted to the current daily problems.
- “Rock”: Rock and Roll conquered the world in the 1950’s and 60’s, entering Greece in the mid-60’s. At the time, a tendency was observed to combine rock music with traditional folk (“ρεμπέτικο”) or “έντεχνο” trying to put lyric poetry through the lyrics.
- “Pop”: Pop is a more generative descriptor characterizing milder forms of foreign music, or corresponding Greek creations on the same standards of rhythm and melody. It should not be confused with “λαϊκό” since the differences in orchestrations are evident.
- “Hip-Hop / R&B”: Hip hop music is a stylized rhythmic music that commonly accompanies rapping, rhythmic samples and rhyming speech that has many exponents in the Greek music scene.
- “εναλλακτικό”: Although it is common for the term to be considered as “Alternative Rock”, in Greek terms it can include tracks fusing Greek modern kinds of music. These tracks usually present something new in the rhythm, melody even in the lyrics which makes them stand apart from the other genres.

The GMD totals 2421 annotations for 1400 songs. After gathering the annotations, it was necessary to use a specific methodology on how to utilize the results. There was no restriction about the number of genre tags assigned in each song following the methodology of [15]. Table 2 shows the number of annotations for each genre. To summarize, GMD includes:

- 40 different genre combinations
- 2421 annotations
- 521 single label annotations from the 8 genre classes
- 748 double label annotations from 17 different combinations
- 119 triple label annotations from 15 different combinations
- 12 quad label annotations from 8 different combinations

Genre	Annotations
“λαϊκό”	580
“ρεμπέτικο”	195
“έντεχνο”	423
“Modern λαϊκό”	448
“Rock”	263
“Pop”	188
“εναλλακτικό”	265
Hip-Hop / R&B	59

Table 2: Number of annotations for each Genre class.

3.4 Format of the Dataset

The data is available in two formats, HDF5 and CSV. Following the methodology of existing datasets, the proposed 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 MIDI files with manually annotated genre and mood for each track.

The HDF5 format is efficient for handling the heterogeneous types of information such as features in variable array lengths, names as strings, and easy for adding new types of data. The structure of the HDF5 format provided in GMD has been based on the Million Song Dataset in order to ensure that researchers will find a familiar interface on GMD. Every song is described by a single file the contents of which are divided in separate categories to the corresponding audio, lyric, symbolic features and metadata.

The provision of CSV format alternative was to ensure compatibility with a wide range of processing software such as Weka [10] and other similar data mining platforms. The GMD provides the commonly used, on the discipline of MIR, feature sets in separate CSV files [17]. The dataset can be found, as a download from the webpage of the Informatics in Humanistic and Social Sciences Lab of the Ionian University [5].

4. FEATURES & METADATA

In addition to the audio, lyrics and symbolic feature sets, described in detail in the sequel, the GMD additionally includes for 621 of its tracks their equivalent Last.fm id aiming to facilitate information collection using the Last.fm’s. This id can be used directly in API calls to Last.fm’s public API methods in order to retrieve more information. The collection of the ids was made by manual processing in order to compensate for misspelled artists’ name and tracks’ titles as well as for multiple versions of tracks available at Last.fm.

[5] http://di.ionio.gr/hilab/gmd

4.1 Audio Features

Feature extraction is the process of computing a compact numerical representation that can be used to characterize a segment of audio. The commonly used acoustic features can be divided into three categories: timbral, rhythm and pitch. The GMD includes several individual feature sets proposed in the MIR and audio analysis literature as well as some common combinations of them. For every feature set, means and variances are calculated too. A total of 454 features are available in the dataset. The following content based audio feature sets have been obtained using the Marsyas software [28].

Timbral Texture Features features used to differentiate mixture of sounds based on their instrumental compositions when the melody and the pitch components are similar. Timbral features are generally used for music-speech discrimination and speech recognition.

- Standard Timbral Set (68 features): One of the most high specificity and commonly used feature set provided by Marsyas. It includes Time ZeroCrossings, Spectral Centroid, Flux and Rolloff, and Mel-Frequency Cepstral Coefficients (MFCC).
- Other Timbral Features (264 features): A combination of other timbral feature sets which Marsyas tool offers and focus in magnitude spectrum. It contains Spectral Flatness Measure, Spectral Crest Factor and Line Spectral Pair.

Rhythm Features Rhythmic content features are used to characterize the regularity of the rhythm, the beat, the tempo, and the time signature [1]. The feature set for representing rhythm structure is based on detecting the most salient periodicities of the signal and is calculated by extracting periodic changes from the beat histogram.

- Beat Histogram (18 features): A vector containing the most commonly rhythmic features (detecting and measuring peaks, bpm etc.)

Pitch (Chroma) Content Features The pitch content features describe the distribution of pitches and show a high degree of invariance to timbral fluctuations.

- Chroma Set (104 features): A combination of Chroma feature set (representing the power spectrum of every note) and Linear Prediction Cepstral Coefficients set (which is widely used for voice recognition).

4.2 Lyrics Features

The complexity of the Greek language has long been an area of study for linguistic experts researching lyrics feature extraction methods. Lyrics feature sets provide a valuable source of information for classification experiments and thus, a number of studies on music mood classification, solely based on lyrics, have spawned in recent years [9, 12]. Stimulated by initial attempts for lyrics feature extraction in Greek music for the purposes of music information research [6], this work describes the selection of 5 feature sets based on the bag-of-words (BOW) model from Greek song lyrics.

Bag-of-words (BOW) are collections of unordered word terms. Each term is assigned a value that can represent, among others, the frequency of the word, its TF-IDF weight, its normalized frequency or a Boolean value indicating presence or absence. Among these variations the GMD includes TF-IDF term weighting, which is the most widely used element in text analysis and MIR with high performance [29] as well as the simple term frequency which indicates the number of occurrences of each term in each song's lyrics. The most popular BOW features are various unigram, bigram, and trigram representations. Unigram terms indicate that metrics for one word have been computed while bigram and trigram terms are for two- and three-word sequences respectively. In addition, to preserve the information of lyrics for customized use, stemming operations are not applied. All in all, the GMD dataset includes the following feature sets:

1. A unigram set of the top 250 words with the most occurrences. 17771 unique words were counted to the total number of 160634 word tokens. This feature set contains also "Function words" words (also called "stop words"), i.e. words that carry no or very little meaning (e.g. articles, pronouns, prepositions). The remaining words are referred to as "Content words".
2. A unigram set of the top 60 words with the most occurrences without counting the Content Words. A special dictionary with Greek "stop words" was adopted from Saroukos' work [23]. This time 17435 unique words were counted to the total number of 82635 word tokens.
3. A bigram set of the top 100 bigram words with the most occurrences. For bigrams and trigrams, function words were not eliminated, as content words are usually connected via function words e.g. in "$\sigma\epsilon$ $\alpha\gamma\alpha\pi\acute{\omega}$" (I love you) where "$\sigma\epsilon$" (you) is a function word and "$\alpha\gamma\alpha\pi\acute{\omega}$" (I love) a content word. 80550 unique bigrams were counted to the total number of 159239 bigram tokens.
4. A trigram set of the top 60 trigram words with the most occurrences. The usefulness of having bigram and trigram tokens is indicated by higher-order BOW features capturing more useful semantics as far as mood classification is concerned than individual n-grams [12]. 121185 unique trigrams were counted to the total number of 157844 trigram tokens.
5. A unigram set of the top 60 function words with the most occurrences. As previously mentioned although function words carry little meaning, these have been shown to be effective in text style analysis [2] and for this reason are included in this separate feature set. 462 unique stop words were counted to the total number of 85344 stop word tokens.

Accordingly, the Greek Music Dataset offers 5 linguistic feature sets with a total of 530 features. The dataset also includes all the lyrics accompanied with MATLAB functions for further feature extraction.

4.3 Symbolic Features

The majority of to date research projects on music classification and similarity analysis have focused on extracting features from audio content. These are known as low-level features. Most features of this type do not provide information that seems intuitively musical but have dominated most aspects of music information related research. Examples include MFCCs, spectral flux, zero-crossing rate and RMS.

On the other hand, high-level features emphasize on the musical characteristics and contain information that consists of musical abstractions meaningful to musically trained individuals. Examples include instruments present, melodic contour, chord frequencies and rhythmic density. Many high-level features cannot currently be reliably extracted from audio recordings and have important musicological and music theoretical value that audio features cannot provide. McKay and Fujinaga [19] showed that including high level features relating to instrumentation are of particular importance when distinguishing between genres.

Although high-level features can be relatively easily extracted from music recorded in symbolic formats (e.g., MIDI), MIR researchers tend to concentrate more on audio formats as these are more easily accessible as well as due to the lack of existing symbolic datasets. The Greek Music Dataset provides the opportunity of high-level feature extraction for a significant subset of the dataset. More specifically GMD offers 2 feature sets for 500 tracks, including mainly traditional Greek folk music genres (“*ρεμπέτικο*”, “*λαϊκό*” and “*έντεχνο*”), accompanied by the corresponding MIDI files for further feature extraction.

The feature extraction was done by Music21 [7], an object-oriented toolkit for analyzing, searching, and transforming music in symbolic forms. Music21 offers two feature sets for high-level features, the native Music21 feature set and the implemented jSymbolic feature set.

- jSymbolic Set (78 features): One of the most useful aspects of the Features module of music21 library, is the integration of 57 features of the 111 implemented in McKay’s jSymbolic toolkit [20]. jSymbolic was the first tool for extracting high-level musical features from symbolic music representations, specifically MIDI files, as well as for iteratively developing and sharing new features. It includes features regarding the instrumentation, voice texture (polyphonic, monophonic), rhythm (time intervals, note durations, meter and tempo), dynamics (loudness), chords (chord progression, harmonic movements) and detecting melody variations or patterns.

- Native Music21 Set (17 features): In addition to the feature set of jSymbolic toolkit, music21 offers a small but specialized and very high-level feature set by taking the advantage of the built-in analytical capabilities and its ability to work with notational aspects. Although it requires a high level of musical harmony knowledge, detecting and calculating percentages of diminished seventh chords and diminished triads could be useful for some music researchers.

5. CONCLUSION

This work presents the Greek Music Dataset (GMD), an extension of the Greek Audio Dataset (GAD), compiled for the purposes of Music Information Research. GMD constitutes a significant extension of GAD as it addresses numerous of the latter’s issues by balancing the content’s categorisation and by incorporating raw and processed symbolic information as well as manual multi-label genre annotation. The dataset contains information for 1400 Greek tracks. For each track, the dataset includes pre-computed audio, lyrics & symbolic features for immediate use in MIR tasks, manually annotated labels pertaining to mood & genre styles of music, generic objective metadata, a manually selected MIDI file (available for 500 of the tracks) and a manually selected link to a performance / audio content in YouTube for further research. Actual audio content is not provided, following the methodology of existing datasets, due to intellectual property rights.

Future directions include the addition of the remaining tracks’ symbolic information, the incorporation of contextual information for each track from social networks as well as experimentation on data mining tasks using the dataset.

6. ACKNOWLEDGMENTS

This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thales. Investing in knowledge society through the European Social Fund.

European Union
European Social Fund

OPERATIONAL PROGRAMME
EDUCATION AND LIFELONG LEARNING
investing in knowledge society
MINISTRY OF EDUCATION & RELIGIOUS AFFAIRS
MANAGING AUTHORITY

NSRF
2007-2013
programme for development
EUROPEAN SOCIAL FUND

Co-financed by Greece and the European Union

7. REFERENCES

[1] *Factors in automatic musical genre classification of audio signals*, 2003.

[2] S. Argamon, M. Šarić, and S. S. Stein. Style mining of electronic messages for multiple authorship discrimination: First results. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 475–480, 2003.

[3] A. Berenzweig, B. Logan, D. P. W. Ellis, and B. P. W. Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Comput. Music J.*, 28(2):63–76, 2004.

[4] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proceedings of International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[5] S. Boisen, M. Crystal, R. M. Schwartz, R. Stone, and R. M. Weischedel. Annotating resources for information extraction. In *LREC*, 2000.

[6] S. Brilis, E. Gkatzou, A. Koursoumis, K. Talvis, K. L. Kermanidis, and I. Karydis. Mood classification using lyrics and audio: A case-study in greek music. In *Proceedings of Advances in Computation and Intelligence*, pages 421–430, 2012.

[7] M. S. Cuthbert and C. Ariza. Music21: A toolkit for computer-aided musicology and symbolic music data. In *Proceedings of International Society for Music Information Retrieval Conference*, pages 637–642, 2010.

[8] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. Rwc music database: Popular, classical, and jazz music databases. In *Proceedings of International Conference on Music Information Retrieval*, pages 287–288, 2002.

[9] H. He, J. Jin, Y. Xiong, B. Chen, W. Sun, and L. Zhao. Language feature mining for music emotion classification via supervised learning from lyrics. In *Proceedings of Advances in Computation and Intelligence*, pages 426–435, 2008.

[10] G. Holmes, A. Donkin, and I. H. Witten. Weka: a machine learning workbench. pages 357–361, 1994.

[11] X. Hu, M. Bay, and J. S. Downie. Creating a simplified music mood classification ground-truth set. In *Proceedings of International Conference on Music Information Retrieval*, pages 309–310, 2007.

[12] Y. Hu, X. Chen, and D. Yang. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *Proceedings of International Society for Music Information Retrieval Conference*, pages 123–128, 2009.

[13] E. Kapsomenos. *Greek Folk music - A different approach*. Patakis, 1999.

[14] P. Konstantinidis. When progress fails, try greekness: From manolis kalomiris to manos hadjidakis and mikis theodorakis. In *Proceedings of International Musicological Conference*, pages 314–320, 2013.

[15] P. Lamere. Social tagging and music information retrieval. *Journal of New Music Research*, 37(2):101–114, 2008.

[16] E. Law and L. von Ahn. Input-agreement: A new mechanism for collecting data using human computation games. In *Proceedings of Conference on Human Factors in Computing Systems*, pages 1197–1206, 2009.

[17] T. Li and M. Ogihara. Towards intelligent music information retrieval. *IEEE Transactions on Multimedia*, 8(3):564–574, 2006.

[18] D. Makris, K. Kermanidis, and I. Karydis. The greek audio dataset. In *Artificial Intelligence Applications and Innovations*, volume 437 of *IFIP Advances in Information and Communication Technology*, pages 165–173. Springer Berlin Heidelberg, 2014.

[19] C. McKay and I. Fujinaga. Automatic music classification and the importance of instrument identification. In *Proceedings of Conference on Interdisciplinary Musicology*, 2005.

[20] C. McKay and I. Fujinaga. jsymbolic: A feature extractor for midi files. In *Proceedings of International Computer Music Conference*, 2006.

[21] Y. Meron and K. Hirose. Automatic alignment of a musical score to performed music. *Acoustical science and technology*, 22(3):189–198, 2001.

[22] N. Ordoulidis. Athens institute for education and research 2nd annual international conference on visual and performing arts, 2011.

[23] S. Saroukos. Enhancing a greek language stemmer - efficiency and accuracy improvements. Master's thesis, Dept. of Computer Sciences, University of Tampere, Finland, 2008.

[24] M. Schedl, C. C. Liem, G. Peeters, and N. Orio. A Professionally Annotated and Enriched Multimodal Data Set on Popular Music. In *Proceedings of Multimedia Systems Conference*, 2013.

[25] R. Thayer. *The biopsychology of mood and arousal*. Oxford University Press, USA, 1989.

[26] D. Tingle, Y. E. Kim, and D. Turnbull. Exploring automatic music annotation with "acoustically-objective" tags. In *Proceedings of International Conference on Multimedia Information Retrieval*, pages 55–62, 2010.

[27] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Towards musical query-by-semantic-description using the cal500 data set. In *Proceedings of International Conference on Research and Development in Information Retrieval*, pages 439–446, 2007.

[28] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE Transactions on*, 10(5):293–302, 2002.

[29] M. v. Zaanen and P. Kanters. Automatic mood classification using tf*idf based on lyrics. In *Proceedings of International Society for Music Information Retrieval Conference*, pages 75–80, 2010.