FMA: A DATASET FOR MUSIC ANALYSIS

Michaël Defferrard[†] **Kirell Benzi**[†] **Pierre Vandergheynst**[†] **Xavier Bresson**[‡]
[†]LTS2, EPFL, Switzerland [‡]SCSE, NTU, Singapore Work conducted when XB was at EPFL.
{michael.defferrard,kirell.benzi,pierre.vandergheynst}@epfl.ch, xbresson@ntu.edu.sg

ABSTRACT

We introduce the Free Music Archive (FMA), an open and easily accessible dataset suitable for evaluating several tasks in MIR, a field concerned with browsing, searching, and organizing large music collections. The community's growing interest in feature and end-to-end learning is however restrained by the limited availability of large audio datasets. The FMA aims to overcome this hurdle by providing 917 GiB and 343 days of Creative Commons-licensed audio from 106,574 tracks from 16,341 artists and 14,854 albums, arranged in a hierarchical taxonomy of 161 genres. It provides full-length and high-quality audio, pre-computed features, together with track- and user-level metadata, tags, and free-form text such as biographies. We here describe the dataset and how it was created, propose a train/validation/test split and three subsets, discuss some suitable MIR tasks, and evaluate some baselines for genre recognition. Code, data, and usage examples are available at https://github.com/mdeff/fma.

1. INTRODUCTION

While the development of new mathematical models and algorithms to solve challenging real-world problems is obviously of first importance to any field of research, evaluation and comparison to the existing state-of-the-art is necessary for a technique to be widely adopted by research communities. Such tasks require open benchmark datasets to be reproducible. In computer vision, the community has developed established benchmark datasets such as MNIST [22], CIFAR [18], or ImageNet [4], which have proved essential to advance the field. The most celebrated example, the ILSVRC2012 challenge on an unprecedented ImageNet subset of 1.3M images [34], demonstrated the power of deep learning (DL), which won the competition with an 11% accuracy advantage over the second best [19], and enabled incredible achievements in both fields [21].

Unlike the wealth of available visual or textual content, the lack of a large, complete and easily available dataset for MIR has hindered research on data-heavy models such as DL. Table 1 lists the most common datasets used for content-based MIR. GTZAN [46], a collection of 1000 clips from 10 genres, was the first publicly available benchmark dataset for genre recognition (MGR). As a result, despites its flaws (mislabeling, repetitions, and distortions), it continues to be the most used dataset for MGR [43]. Moreover, it is small and misses metadata which e.g. prevents researchers to control for artists or album effects. Looking at Table 1, the well-known MagnaTagATune [20] and the Million Song Dataset (MSD) [3] as well as the newer AudioSet [10] and AcousticBrainz [32] appear as contenders for a large-scale reference dataset. MagnaTagATune, which was collected from the Magnatune label and tagged using the TagATune game, includes metadata, features and audio. The poor audio quality and limited number of songs does however limit its usage. MSD and AudioSet, while very large, force researchers to download audio clips from online services. AcousticBrainz's approach to the copyright issue is to ask the community to upload music descriptors of their tracks. Although it is the

dataset[1]	#clips	#artists	year	audio
RWC [12]	465	-	2001	yes
CAL500 [45]	500	500	2007	yes
Ballroom [13]	698	-	2004	yes
GTZAN [46]	1,000	$\sim$ 300	2002	yes
MusiClef [36]	1,355	218	2012	yes
Artist20 [7]	1,413	20	2007	yes
ISMIR2004	1,458	-	2004	yes
Homburg [15]	1,886	1,463	2005	yes
103-Artists [30]	2,445	103	2005	yes
Unique [41]	3,115	3,115	2010	yes
1517-Artists [40]	3,180	1,517	2008	yes
LMD [42]	3,227	-	2007	no
EBallroom [23]	4,180	-	2016	no[2]
USPOP [1]	8,752	400	2003	no
CAL10k [44]	10,271	4,597	2010	no
MagnaTagATune [20]	25,863[3]	230	2009	yes[4]
Codaich [28]	26,420	1,941	2006	no
FMA	**106,574**	**16,341**	**2017**	**yes**
OMRAS2 [24]	152,410	6,938	2009	no
MSD [3]	1,000,000	44,745	2011	no[2]
AudioSet [10]	2,084,320	-	2017	no[2]
AcousticBrainz [32]	2,524,739[5]	-	2017	no

[1] Names are clickable links to datasets' homepage.
[2] Audio not directly available, can be downloaded from ballroomdancers.com, 7digital.com, youtube.com.
[3] The 25,863 clips are cut from 5,405 songs.
[4] Low quality 16 kHz, 32 kbit/s, mono mp3.
[5] As of 2017-07-14, of which a subset has been linked to genre labels for the MediaEval 2017 genre task.

Table 1: Comparison between FMA and alternative datasets.

100% track_id	100% title	93% number
2% information	14% language_code	100% license
4% composer	1% publisher	1% lyricist
98% genres	98% genres_all	47% genre_top
100% duration	100% bit_rate	100% interest
100% #listens	2% #comments	61% #favorites
100% date_created	6% date_recorded	22% tags
100% album_id	100% title	
94% type	96% #tracks	
76% information	16% engineer	18% producer
97% #listens	12% #comments	38% #favorites
97% date_created	64% date_released	18% tags
100% artist_id	100% name	25% members
38% bio	5% associated_labels	
43% website	2% wikipedia_page	
	5% related_projects	
37% location	23% longitude	23% latitude
11% #comments	48% #favorites	10% tags[1]
99% date_created	8% active_year_begin	
	2% active_year_end	

[1] One of the tags is often the artist name. It has been subtracted.

Table 2: List of available per-track, per-album and per-artist metadata, i.e. the columns of `tracks.csv`. Percentages indicate coverage over all tracks, albums, and artists.

largest database to date, it will never distribute audio. On the other hand, the proposed dataset offers the following qualities, which in our view are essential for a reference benchmark.

Large scale. Large datasets are needed to avoid overtraining and to effectively learn models that incorporate the ambiguities and inconsistencies that one finds with musical categories. They are also more diverse and allows to average out annotation noise as well as characteristics who might be confounded with the ground truth and exploited by learning algorithms. While FMA features less clips than MSD or AudioSet, every other dataset with available quality audio are two orders of magnitude smaller (Table 1).

Permissive licensing. MIR research has historically suffered from the lack of publicly available benchmark datasets, which stem from the commercial interest in music by record labels, and therefore imposed rigid copyright. The FMA's solution is to aim for tracks which license permits redistribution. All data and code produced by ourselves are licensed under the CC BY 4.0 and MIT licenses.

Available audio. Table 1 shows that while the smaller datasets are usually distributed with audio, most of the larger do not. They either (i) only contain features derived from the audio, or (ii) provide links to download the audio from an online service. [1] The problem with (i) is that researchers are stuck with the chosen features and are prevented to leverage feature learning or end-to-end learning systems like DL. Moreover, we should be wary of proprietary features like those computed by commercial services such as Echonest. The problem with (ii) is that researchers have no control, i.e. we have no assurance that the files or services will not disappear or change without notice.

Quality audio. Distributed or downloadable audio are usually clips of 10 to 30 seconds and sometimes of low quality, e.g. 32 kbit/s for MagnaTagATune or an average of 104 kbit/s for MSD [37]. The problem with clips is that the beginning 30 seconds of tracks may yield different results than the middle or final 30 seconds, and that researchers may not have control over which part they get. In comparison, FMA comes with full-length and high-quality audio.

Metadata rich. The dataset comes with rich metadata, shown in Table 2. While not complete in any means, it compares favorably with the MSD which only provides artist-level metadata [3] or GTZAN which offers none.

Easily accessible. Working with the dataset only requires to download some `.zip` archives containing `.csv` metadata and `.mp3` audio. No need to crawl the web and circumvent rate limits or access restrictions. Besides, we provide some usage examples in the `usage.ipynb` Jupyter notebook to start using the data quickly.

Future proof and reproducible. All files and archives are checksummed and hosted in a long-term digital archive. Doing so alleviates the risks of songs to become unavailable. Moreover, we share all the code used to (i) collect the data, (ii) analyze it, (iii) generate the subsets and splits, (iv) compute the features and (v) test the baselines. The developed code can serve as a starting point for researchers to compute their own features or evaluate their methods. Finally, anybody can recreate or extend the collection, thanks to public songs and APIs.

Note that an alternative to open benchmarking is the approach taken by the MIREX evaluation challenges: the evaluation (by the organizers) of submitted algorithms on private datasets [6]. This practice however incurs an approximately linear cost with the number of submissions, which put the long-term sustainability of MIREX at risk [26]. By releasing this open dataset, we realize part of the vision of McFee *et al*. in "a distributed, community-centric paradigm for system evaluation, built upon the principles of openness, transparency, and reproducibility".

2. DATASET

2.1 The Free Music Archive

The dataset, both the audio and metadata, is a dump of the Free Music Archive, a free and open library directed by WFMU, the longest-running freeform radio station in the United States. Inspired by Creative Commons and the open-source software movement, the FMA provides a platform for curators, artists, and listeners to harness the potential of music sharing. The website provides a large catalog of artists and tracks, hand-picked by established audio curators. Each track is legally free to download as artists decided to release their works under permissive licenses. While there exists other sources of CC-licensed music, notably Jamendo, FMA is unique as it combines user-generated content with the curatorial role that WFMU and others have always played. [2]

2.2 Creation

As of April 1st 2017, when the dataset was gathered, the online archive largest track id was 155,320, of which

[1] Going to the source distributor is a way to adhere with copyright.

[2] Interview with Jason Sigal of the Free Music Archive, Rhizome.

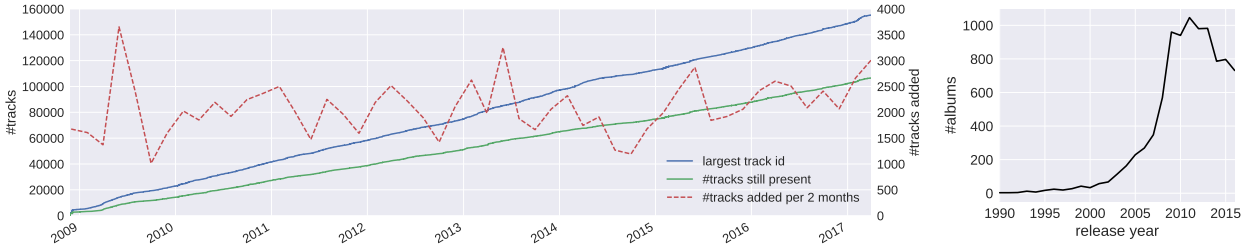

Figure 1: (left) Growth of the archive, created in 11/2008. (right) Number of albums released per year (min 1902, max 2017).

track						album			artist	
track_id	title	genres_all	genre_top	dur.	listens	title	listens	tags	name	location
150073	Welcome to Asia	[2, 79]	International	81	683	Reprise	4091	[world music, dubtronica, fusion]	DubRaJah	Russia
140943	Sleepless Nights	[322, 5]	Classical	246	1777	Creative Commons Vol. 7	28900	[classical, alternate, soundtrack, piano, ...	Dexter Britain	United Kingdom
64604	i dont want to die alone	[32, 38, 456]	Experimental	138	830	Summer Gut String	7408	[improvised, minimalist, noise, ...	Buildings and Mountains	Oneonta, NY
23500	A Life In A Day	[236, 286, 15]	Electronic	264	1149	A Life in a Day	6691	[idm, candlestick, romanian, candle, ...	Candlestickmaker	Romania
131150	Yeti-Bo-Betty	[25, 12, 85]	Rock	124	183	No Life After Crypts	3594	[richmond, fredericksburg, trash rock, ...	The Crypts!	Fredericksburg

Table 3: Some rows and columns of the metadata table, stored in `tracks.csv`.

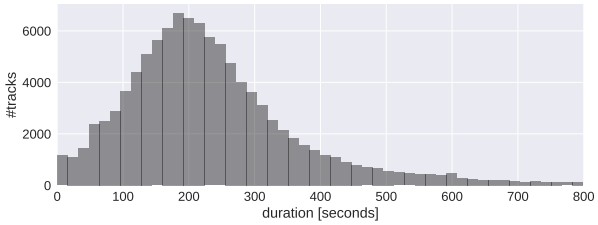

Figure 2: Track duration (min 0, max 3 hours).

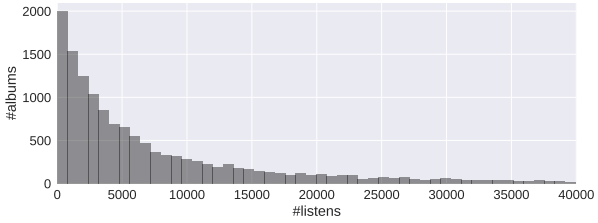

Figure 3: Album listens (min 0, max 3.6 millions).

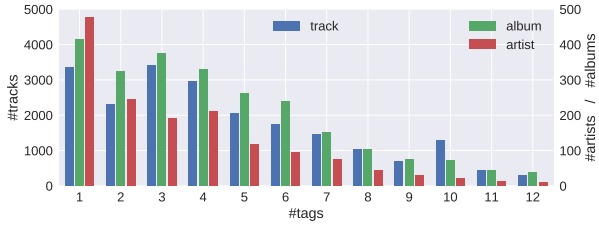

Figure 4: Per-track, album and artist tags (min 0, max 150).

109,727 were valid. The missing 45,594 ids probably correspond to deleted tracks. Figure 1 illustrates the growth of the dataset. In addition to per-track metadata, the used hierarchy of 161 genres and extended per-album (480 not found) and per-artist (250 not found) metadata were collected via the available API.[3] Finally, mp3 audio was downloaded over HTTPS. Out of all collected track ids, 180 mp3s could not be downloaded, 286 could not be trimmed by ffmpeg, and features could not be extracted from 71. Finally, the license of 2,616 tracks prohibited their redistribution, leaving us with 106,574 tracks.

While it may be argued that the dataset should be cleaned, we wanted it to resemble real world data. As such, we did not remove tracks which have too many genres, are too long, belong to rare genres, etc. Moreover, it is hard to set a threshold, algorithms shall handle outliers, and the small number of outliers will not impact performance much anyway. Researchers are obviously free to discard any track for training.

[3] See `webapi.ipynb` to query the API with our helpers.

2.3 Content

The collected metadata[4] was cleaned, uniformly formatted, merged and stored in `tracks.csv`[5] which Table 3 shows an excerpt. That file is a relational table where each row represents a track and columns are listed in Table 2. For ease of use, we kept all the metadata in a single table despite the redundancy incurred by the fact that all tracks from a given artist share all artist related columns. The problem is mitigated in practice by compression for storage and by *categorical variables* for memory usage.

All the metadata available through the API has been archived. It includes song title, album, artist, and per-track genres; user data such as per-track/album/artist favorites, play counts, and comments; free-form text such as per-track/album/artist tags, album description and artist biography. Coverage varies across fields and is reported in Table 2. Note that all that metadata has been produced by the artists when uploading their music and that while the content is curated, the curators focus on the musical content not the metadata. Figures 1, 2 and 3 show the distribution of albums per year, track durations, and play counts per album. See the `analysis.ipynb` notebook for a much more detailed analysis of the content.

The audio for each track is stored in a file which name is the track id. All tracks are mp3-encoded, most of them with sampling rate of 44,100 Hz, bit rate 320 kbit/s (263 kbit/s on average), and in stereo.

[4] `raw_tracks.csv`, `raw_albums.csv`, `raw_artists.csv`

[5] See `creation.ipynb` for the code which created the dataset.

id	parent	top_level	title	#tracks
38	None	38	Experimental	38,154
15	None	15	Electronic	34,413
12	None	12	Rock	32,923
1235	None	1235	Instrumental	14,938
25	12	12	Punk	9,261
89	25	12	Post-Punk	1,858
1	38	38	Avant-Garde	8,693

Table 4: An excerpt of the genre hierarchy, stored in `genres.csv`. Some of the 16 top-level genres appear in the top part, while some second- and third-level genres appear in the bottom part.

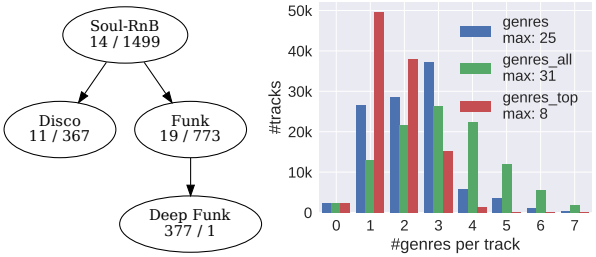

Figure 5: (left) Example of genre hierarchy for the top-level Soul-RnB genre. Left number is the `genre_id`, right is the number of tracks per genre. (right) Number of genres per track. A 3 genres limit has been introduced early on by the administrators.

2.4 Genres

The FMA is especially suited for MGR as it features fine genre information, i.e. multiple (sub-)genres associated to individual tracks, has a built-in genre hierarchy (Table 4), and is annotated by the artists themselves. While the artists are the best placed to judge the positioning of their creations, they might be inconsistent and motivated by factors not necessarily objective, such as achieving a higher play count. As labeling noise is unavoidable, those labels should ideally be one of many ground truths, to be complemented by crowd-sourcing and experts (from different music metadata websites).

While there is no agreement on a taxonomy of genres [35], we followed the hierarchy used by the archive, which is the one the authors had in mind when annotating their tracks. That hierarchy is composed of 161 genres of which 16 are roots, the others being sub-genres. Table 4 shows an excerpt of that information along with the number of tracks per genre and the associated top-level genre, that is the root of the genre tree. Figure 5 shows an excerpt of the tree.

In the per-track table, the `genres` column contain the genre ids indicated by the artist. Then, given such hierarchical information, we constructed a `genres_all` column which contains all the genres encountered when traversing the tree from the indicated genres to the roots. The root genres are stored in the `genres_top` column. Figure 5 and 6 shows the number of genres per track and tracks per genre.

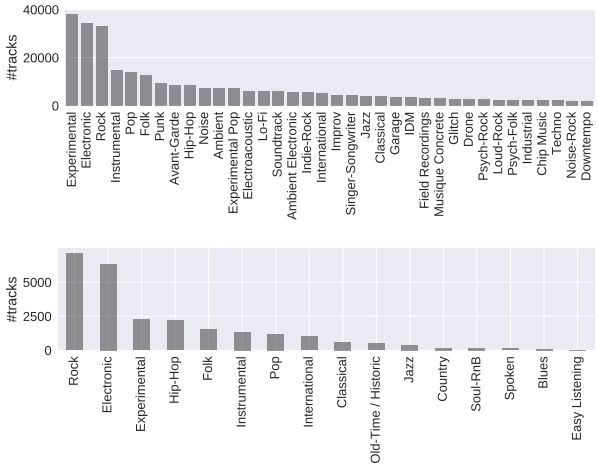

Figure 6: (top) Tracks per (sub-)genre on the full set (min 1, max 38,154). (bottom) Tracks per all 16 root genres on the medium subset (min 21, max 7,103). Note how experimental music is much less represented in the curated medium subset.

2.5 Features

To allow researchers to experiment without dealing with feature extraction, we pre-computed the features listed in Table 6. These are all the features the librosa Python library, version 0.5.0 [25], was able to extract. Each feature set (except zero-crossing rate) is computed on windows of 2048 samples spaced by hops of 512 samples. Seven statistics were then computed over all windows: the mean, standard deviation, skew, kurtosis, median, minimum and maximum. Those 518 pre-computed features are distributed in `features.csv` for all tracks. [6]

2.6 Subsets

For the dataset to be useful as a development set or for people with lower computational resources, we propose the following sets, each of which is a subset of the larger set:

1. **Full**: the complete dataset, described above. All 161 genres, unbalanced with 1 to 38,154 tracks per genre (Figure 6) and up to 31 genres per track (Figure 5).
2. **Large**: the full dataset with audio limited to 30 seconds clips extracted from the middle of the tracks (or entire track if shorter than 30 seconds). That trimming reduces the size of the data by a factor 10.
3. **Medium**: while root genre recognition should be treated as a multi-label problem in general, we constructed this subset for the simpler problem of single-label prediction. It makes sense as half the tracks have a single root genre (Figure 5). As such, we selected those tracks with only one top genre and sampled the clips according to the completeness of their metadata and their popularity, hoping to select tracks of higher quality. That selection left us with 25,000 30s clips, genre unbalanced with 21 to 7,103 clips per top genre (Figure 6), but only one of the 16 top genres per clip.

[6] See `features.py` for the code which computed the features.

4. **Small**: to construct a balanced subset, we selected with the same process the top 1,000 clips from the 8 most popular genres of the medium set. The subset is thus composed of 8,000 30s clips from 8 top genres, balanced with 1,000 clips per genre, 1 root genre per clip. This subset is similar to the very popular GTZAN [46] with the added benefits of the FMA, that is metadata, pre-computed features, and copyright-free audio.

Table 5 highlights the main differentiating factors between the proposed subsets.

2.7 Splits

We propose an 80/10/10% split into training, validation and test sets to make research on the FMA reproducible. Training and validation shall be merged if cross-validation is used instead. Below are the followed constraints:

1. Stratified sampling to preserve the percentage of tracks per genre (important for minority genres). Each root genre is guaranteed to be represented in all splits, but the ratio is only exact for the small subset (800/100/100). The seven smallest sub-genres (less than 20 tracks in total) are however not guaranteed to appear in all splits of the full and large sets.
2. An artist filter for artists to be part of one set only, thus avoiding any artist and album effect. It has been shown that the use of songs from the same artist in both training and test sets leads to over-optimistic accuracy and may favor some approaches [8, 29].

The above constraints are satisfied for all subsets, and a track is assigned to the same split across all of them.[5] The 2,231 tracks without genre label are assigned to the training set (full and large sets) as they might be useful as additional training samples for semi-supervised algorithms.

3. USAGE

With its rich set of metadata, user data, audio and features, the FMA is amenable to many tasks in MIR. We share below some possible uses which serve to illustrate the breadth of data available in the dataset.

3.1 Music Classification and Annotation

Music classification is a key problem in MIR with many potential applications. For one, a classification system enables end users to search for the types of music they are interested in. On the other hand, different music types are managed more effectively and efficiently once they are categorized into different groups [9]. The classification tasks which can readily be evaluated on FMA include genre recognition, artist identification, year prediction, and automatic tagging. Automatic tagging [2] is a classification problem which covers different semantic categories, where tags are labels which can be any musical term that describes the genre, mood, instrumentation, and style of the song. It helps to convert the music retrieval problem to text retrieval by substituting songs with tags. In addition to supervised methods which classify music given an arbitrary taxonomy, another approach is to cluster data in an unsupervised way so that a categorization will emerge from the data itself based on objective similarity measures. Then, does genre or another taxonomy naturally come up?

dataset	clips	genres	length	size	
			[s]	[GiB]	#days
small	8,000	8	30	7.4	2.8
medium	25,000	16	30	23	8.7
large	106,574	161	30	98	37
full	106,574	161	278	917	343

Table 5: Proposed subsets of the FMA.

feature set	dim.	LR	kNN	SVM	MLP
1 Chroma [11]	84	44	44	48	49
2 Tonnetz [14]	42	40	37	42	41
3 MFCC [33]	140	58	55	61	53
4 Spec. centroid	7	42	45	46	48
5 Spec. bandwidth	7	41	45	44	45
6 Spec. contrast [17]	49	51	50	54	53
7 Spec. rolloff	7	42	46	48	48
8 RMS energy	7	37	39	39	39
9 Zero-crossing rate	7	42	45	45	46
3 + 6	189	60	55	63	54
3 + 6 + 4	273	60	55	63	53
1 to 9	518	61	52	63	58

Table 6: Test set accuracies of various features and classifiers for top genre recognition on the medium subset.

3.2 Genre Recognition

Music genres are categories that have arisen through a complex interplay of cultures, artists, and market forces to characterize similarities between compositions and organize music collections. Yet, the boundaries between genres still remain fuzzy, making the problem of music genre recognition (MGR) a nontrivial task [35]. While its utility has been debated, mostly because of its ambiguity and cultural definition, it is widely used and understood by end-users who find it useful to discuss musical categories [27]. As such, it is one of the most researched areas of MIR. We propose the following prediction problems of increasing difficulty:

1. Single top genre on the balanced small subset.
2. Single top genre on the unbalanced medium subset.
3. Multiple top genres on the large / full set.
4. Multiple (sub-)genres on the large / full set.

Table 6 reports accuracies for problem 2 with nine mainstream feature sets and some combinations as well as four standard classifiers using scikit-learn, version 0.18.1 [31]. Specifically, we employed linear regression (LR) with an L^2 penalty, k-nearest neighbors (kNN) with $k = 200$, support vector machines (SVM) with a radial basis function (RBF) kernel and a multilayer perceptron (MLP) with 100 hidden neurons. All classifiers were tested with otherwise default settings.[7] Reported performance should not be taken as the state-of-the-art but rather as

[7] See `baselines.ipynb` for all details.

a lower-bound and an indication of the task's difficulty. Moreover, the developed code can serve as a reference and is easily modified to accommodate other features and classifiers.

A major motivation to construct this dataset was to enable the use of the powerful DL set of techniques to music analysis, an hypothesized cause of stagnation on MIREX tasks [38]. With availability of audio, DL architectures such as convolutional neural networks and recurrent neural networks can be applied to the waveform to avoid any feature engineering. While those approaches have fallen behind learning from higher-level representations such as spectrograms [5], a greater exploration of the design space will hopefully provide alternatives to solving MIR challenges [16].

3.3 Data Analysis

While our intention was to release a large volume of audio for machine learning algorithms, analyzing audio is certainly of interest to musicologists and researchers who want to study relations with higher-level representations. Moreover, the availability of complete tracks allows proper study of music properties, for example music structure analysis. Finally, the metadata is surely a valuable addition to existing datasets (e.g. MusicBrainz, AllMusic, Discogs, Last.fm) for metadata analysis.

4. DISCUSSION

While the FMA can be used to evaluate many tasks, metadata is missing for e.g. mood classification or instrument recognition. However, a more thorough investigation of the available tags may reveal their feasibilities. Similarly, cover song detection may be doable if multiple versions of many songs are featured. While the present dump only captures listening and downloading counts in aggregates, [8] the lists of which songs, albums and artists a user marked as favorites or commented are public, as well as *user mixes*. While not public, listening and downloading activities are logged and might be shared after anonymization. [9] Moreover, users form a public social network via *friend requests*. Collecting this information would open the possibility of a large-scale evaluation of content-based recommender systems. Cover images for tracks, albums, and artists are another public asset which may be of interest. Finally, we can expect the dataset to be cross-referenced with other resources to unlock additional tasks, as has happened for example with the MSD and AllMusic, last.fm and beaTunes for genre recognition [37, 39], musixmatch for lyrics, SecondHandSongs for cover songs, or This Is My Jam for user play counts.

Diversity is another issue. As suggested by Figure 6, this collection is biased toward experimental, electronic, and rock music. Moreover, it does not contain mainstream music and few commercially successful artists. A common criticism of basing research on CC-licensed music is that the music is of substantially lower "quality". Moreover, it is unknown whether datasets made up of mainstream or non-mainstream music have similar properties and if algorithms tailored on one perform similarly on the other. While those points are valid for high-level tasks such as recommendation (which depend on a variety of factors beyond the acoustic content), this is a much more tenuous case for the majority of tasks, in particular perceptual tasks. Nevertheless, algorithms should ideally be evaluated on multiple datasets, which will help answer such questions.

[8] That information can be useful to e.g. analyze and predict hits.

[9] Private discussion with the website administrators.

5. CONCLUSION AND PERSPECTIVES

Benchmarking is an important aspect in experimental sciences — results reported by individual research groups need to be comparable. Important aspects of these are datasets that can be easily shared among researchers, together with a set of defined tasks and splits. The FMA enables researchers to test algorithms on a large-scale collection, closer to real-world environments. Even though it is still two orders of magnitude behind commercial services who have access to tens of millions of tracks, [10] it is of the same scale as the largest image dataset which opened the door to dramatic performance improvements for many tasks in computer vision. By providing audio, we do not limit the benchmarking to pre-computed features and allow scientists to develop and test new feature sets, learn features, or learn mappings directly from the audio. For now, music classification, and MGR in particular, is the most straightforward use case for FMA. The inclusion of a genre hierarchy makes it specially interesting, as it offers possibilities rarely found in alternative collections.

In addition to the proposed usage and many others people will find, future work on the dataset itself should focus on (i) validating the ground truth by measuring agreement by independent annotators and (ii) obtaining additional metadata and labels. If the community finds interest in the dataset and validate its use, that can be achieved by scraping the website for information not available through the API, cross-referencing with other resources, or crowdsourcing (with e.g. Mechanical Turk or CrowdFlower).

In a post about the dataset, Cheyenne Hohman, the Director at the Archive, wrote that "by embracing the ... philosophy of Creative Commons, artists are not only making their music available for the public to listen to, but also for educational and research applications". Let's hope for a future where sharing is first and researchers feed open platforms with algorithms while they feed us with data.

6. ACKNOWLEDGMENTS

We want to thank the team supporting the Free Music Archive as well as all the contributing artists and curators for the fantastic content they made available. We want to thank the anonymous ISMIR reviewers for their thorough reviews and many constructive comments which have improved the quality of this work. We are grateful to

[10] 37M Echonest, 30M Spotify, 45M last.fm, 45M 7digital, 26M iTunes

SWITCH and EPFL for hosting the dataset within the context of the SCALE-UP project, funded in part by the swissuniversities SUC P-2 program. Xavier Bresson is supported by NRF Fellowship NRFF2017-10.

7. REFERENCES

[1] A Berenzweig, B Logan, D PW Ellis, and B Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music J.*, 2004.

[2] T Bertin-Mahieux, D Eck, and M Mandel. Automatic tagging of audio: The state-of-the-art. *Machine Audition: Principles, Algorithms and Systems*, 2010.

[3] T Bertin-Mahieux, D PW Ellis, B Whitman, and P Lamere. The million song dataset. In *ISMIR*, 2011.

[4] J Deng, W Dong, R Socher, LJ Li, K Li, and L Fei-Fei. Imagenet: A large-scale hierarchical image database. *Computer Vision and Pattern Recognition*, 2009.

[5] S Dieleman and B Schrauwen. End-to-end learning for music audio. In *ICASSP*, 2014.

[6] J Downie, A Ehmann, M Bay, and M Jones. The music information retrieval evaluation exchange: Some observations and insights. *Advances in Music Information Retrieval*, 2010.

[7] D PW Ellis. Classifying music audio with timbral and chroma features. In *ISMIR*, 2007.

[8] A Flexer. A closer look on artist filters for musical genre classification. In *ISMIR*, 2007.

[9] Z Fu, G Lu, K M Ting, and D Zhang. A survey of audio-based music classification and annotation. *IEEE Trans. on Multimedia*, 2011.

[10] J F Gemmeke, D PW Ellis, D Freedman, A Jansen, W Lawrence, R C Moore, M Plakal, and M Ritter. Audio set: An ontology and human-labeled dartaset for audio events. In *ICASSP*, 2017.

[11] M Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Trans. on Audio, Speech, and Language Processing*, 2006.

[12] M Goto, H Hashiguchi, T Nishimura, and R Oka. Rwc music database: Popular, classical and jazz music databases. In *ISMIR*, 2002.

[13] F Gouyon, S Dixon, E Pampalk, and G Widmer. Evaluating rhythmic descriptors for musical genre classification. In *Proc. of the AES 25th Int. Conf.*, 2004.

[14] C Harte, M Sandler, and M Gasser. Detecting harmonic change in musical audio. In *In Proc. of Audio and Music Computing for Multimedia Workshop*, 2006.

[15] H Homburg, I Mierswa, B Möller, K Morik, and M Wurst. A benchmark dataset for audio classification and clustering. In *ISMIR*, 2005.

[16] E J Humphrey, Juan P Bello, and Y LeCun. Moving beyond feature design: Deep architectures and automatic feature learning in music informatics. In *ISMIR*, 2012.

[17] DN Jiang, L Lu, HJ Zhang, JH Tao, and LH Cai. Music type classification by spectral contrast feature. In *IEEE Int. Conf. on Multimedia and Expo*, 2002.

[18] A Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009.

[19] A Krizhevsky, I Sutskever, and G E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[20] E Law, K West, M I Mandel, M Bay, and J S Downie. Evaluation of algorithms using games: The case of music tagging. In *ISMIR*, 2009.

[21] Y LeCun, Y Bengio, and G Hinton. Deep learning. *Nature*, 2015.

[22] Y LeCun, L Bottou, Y Bengio, and P Haffner. Gradient-based learning applied to document recognition. In *Proc. of the IEEE*, 1998.

[23] U Marchand and G Peeters. The extended ballroom dataset. 2016.

[24] M Mauch, C Cannam, M Davies, S Dixon, C Harte, S Kolozali, D Tidhar, and M Sandler. Omras2 metadata project 2009. In *ISMIR*, 2009.

[25] B McFee et al. librosa 0.5.0, 2017.

[26] B McFee, E J Humphrey, and J Urbano. A plan for sustainable mir evaluation. In *ISMIR*, 2016.

[27] C McKay and I Fujinaga. Musical genre classification: Is it worth pursuing and how can it be improved? In *ISMIR*, 2006.

[28] C McKay, D McEnnis, and I Fujinaga. A large publicly accassible prototype audio database for music research. In *ISMIR*, 2006.

[29] E Pampalk, A Flexer, and G Widmer. Improvements of audio-based music similarity and genre classificaton. In *ISMIR*, 2005.

[30] E Pampalk, A Flexer, G Widmer, et al. Improvements of audio-based music similarity and genre classificaton. In *ISMIR*, 2005.

[31] F Pedregosa et al. Scikit-learn: Machine learning in python. *J. of Machine Learning Research*, 2011.

[32] A Porter, D Bogdanov, R Kaye, R Tsukanov, and X Serra. Acousticbrainz: a community platform for gathering music information obtained from audio. In *ISMIR*, 2015.

[33] L R Rabiner and BH Juang. *Fundamentals of speech recognition*. 1993.

[34] O Russakovsky et al. Imagenet large scale visual recognition challenge. *Int. J. of Computer Vision*, 2015.

[35] N Scaringella, G Zoia, and D Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 2006.

[36] M Schedl, N Orio, C Liem, and G Peeters. A professionally annotated and enriched multimodal data set on popular music. In *Proc. of the ACM Multimedia Systems Conference*, 2013.

[37] A Schindler, R Mayer, and A Rauber. Facilitating comprehensive benchmarking experiments on the million song dataset. In *ISMIR*, 2012.

[38] R Scholz, G Ramalho, and G Cabral. Cross task study on mirex recent results: An index for evolution measurement and some stagnation hypotheses. In *ISMIR*, 2016.

[39] H Schreiber. Improving genre annotations for the million song dataset. In *ISMIR*, 2015.

[40] K Seyerlehner, G Widmer, and P Knees. Frame level audio similarity - a codebook approach. In *Proc. of the Int. Conf. on Digital Audio Effects*, 2008.

[41] K Seyerlehner, G Widmer, and T Pohle. Fusing block-level features for music similarity estimation. In *Proc. of the Int. Conf. on Digital Audio Effects*, 2010.

[42] C N Silla Jr, A L Koerich, and C AA Kaestner. The latin music database. In *ISMIR*, 2008.

[43] B L Sturm. The state of the art ten years after a state of the art: Future research in music information retrieval. *J. of New Music Research*, 2014.

[44] D Tingle, Y E Kim, and D Turnbull. Exploring automatic music annotation with acoustically-objective tags. In *Proc. of the Int. Conf. on Multimedia Information Retrieval*, 2010.

[45] D Turnbull, L Barrington, D Torres, and G Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Trans. on Audio, Speech, and Language Processing*, 2008.

[46] G Tzanetakis and P Cook. Musical genre classification of audio signals. *IEEE Trans. on Speech and Audio Processing*, 2002.