MELON PLAYLIST DATASET: A PUBLIC DATASET FOR AUDIO-BASED PLAYLIST GENERATION AND MUSIC TAGGING

Andres Ferraro[⋆] *Yuntae Kim*[†] *Soohyeon Lee*[†] *Biho Kim*[†] *Namjun Jo*[†] *Semi Lim*[†]
Suyon Lim[†] *Jungtaek Jang*[†] *Sehwan Kim*[†] *Xavier Serra*[⋆] *Dmitry Bogdanov*[⋆]

[⋆] Music Technology Group - Universitat Pompeu Fabra, Spain
[†] Kakao Corp, Korea

ABSTRACT

One of the main limitations in the field of audio signal processing is the lack of large public datasets with audio representations and high-quality annotations due to restrictions of copyrighted commercial music. We present Melon Playlist Dataset, a public dataset of mel-spectrograms for 649,091 tracks and 148,826 associated playlists annotated by 30,652 different tags. All the data is gathered from Melon, a popular Korean streaming service. The dataset is suitable for music information retrieval tasks, in particular, auto-tagging and automatic playlist continuation. Even though the latter can be addressed by collaborative filtering approaches, audio provides opportunities for research on track suggestions and building systems resistant to the cold-start problem, for which we provide a baseline. Moreover, the playlists and the annotations included in the Melon Playlist Dataset make it suitable for metric learning and representation learning.

***Index Terms*—** Datasets, music information retrieval, music playlists, auto-tagging, audio signal processing

1. INTRODUCTION

Open access to adequately large datasets is one of the main challenges in the field of audio signal processing and music information retrieval (MIR) due to the limitations of the copyrighted material. The lack of public datasets makes collaboration between researchers and reproducibility of academic studies more difficult, limiting developments in these fields.

In this work, we present a public dataset of information about 148,826 playlists collected by Kakao[1] from Melon,[2] the most popular music platform in Korean. This dataset also contains the mel-spectrogram representations of the audio for 649,091 tracks, covering the music consumed in Korea (i.e., mainly Korean pop, but also Western music). Thus, we provide a large-scale public dataset of playlists that includes audio information for commercial music directly accessible without the need to collect it from different external sources, which is the problem of other existing playlist datasets. The dataset can be accessed online prior registration.[3]

The playlists are collected from Melon users manually verified by moderators for providing quality public playlists. These users add metadata to the playlists, such as tags and title, which are also included in the dataset. The dataset was originally collected for the automatic playlist continuation (APC) and tag prediction challenge. Possible applications go beyond the scope of the original challenge, and the size of the dataset makes it suitable for deep learning approaches that require large amount of information. New methods can be applied for music, e.g., deep metric learning, representation learning, and semi-supervised learning.

The paper is structured as follows. We review related public datasets in Section 2 and describe the proposed dataset in Section 3. Section 4 highlights its main applications and shows an example task of automatic playlist continuation in a cold-start scenario. Section 5 concludes the paper.

2. RELATED WORK

Table 1 summarizes the existing datasets for the tasks of music auto-tagging and automatic playlist continuation.

MagnaTagATune [1] (MTAT) is commonly used for auto-tagging, but mainly for prototyping because of its small size. The Million Song Dataset [2] (MSD) contains audio features extracted for one million songs, it was expanded by the MIR community with additional metadata, including collaborative tags from Last.fm. It was previously possible to download 30-second audio previews for MSD through the 7digital service, but it is no longer accessible. Another limitation of this dataset is the noise in the tags [3].

To address the issue of open access to audio, the FMA [4] and MTG-Jamendo datasets [5] were proposed for auto-tagging, both containing audio under Creative Commons licenses. The former is based on poorly structured music archives with inconsistent annotations and low-quality recordings. The latter tries to address this issue, focusing on a free music collection maintained for a commercial use-case, thus

[1] `https://www.kakaocorp.com`
[2] `https://www.melon.com`
[3] `https://arena.kakao.com/melon_dataset`

Dataset	Tracks	Tags	Playlists	Audio (official)
MTAT	5,405	188	–	30 s previews
MSD	505,216	522,366	–	–
FMA	106,574	161	–	full CC tracks
MTG-J	55,609	195	–	full CC tracks
MPD	2,262,292	–	1,000,000	some previews through API
MPSD	1,993,607	–	74,996	–
Melon Music	649,091	30,652	148,826	20-50 s mel-spectrograms

Table 1. Public datasets for automatic playlists continuation and auto-tagging compared to Melon Playlist Dataset. CC stands for audio available under Creative Commons licenses.

Property	Count
Track-playlist relations	5,904,718
Unique tracks	649,091
Tag-playlist relations	516,405
Unique tags	30,652
Playlists	148,826
Playlist titles	121,485
Unique playlist titles	116,536
Artists	107,824
Albums	269,362
Genres	30

Table 2. Melon Playlist Dataset statistics.

containing better quality audio and annotations. Yet, their content is different from commercial music platforms.

Recently the Million Playlist Dataset [6] (MPD) was released by Spotify. This dataset contains information about one million playlists created by their U.S. users. However, it does not include the tracks' audio information. Even if it may be possible to download 30-second audio previews with the Spotify API, it is unclear if it is legal to redistribute them. Also, there can be inconsistencies when trying to download audio previews in the future (e.g., due to songs changing their identifier or restricted access to some of the previews in different countries). These limitations significantly affect the reproducibility and complicate the use of MPD for audio research.

The Million Playlists Songs Dataset [7] (MPSD) combines multiple smaller datasets (Art of The Mix [8], #nowplaying [9], and 30Music [10]). Similar to MPD, this dataset does not provide audio nor its representations for the songs. Since it contains playlists collected from different sources, there can be noise in the data due to song matching inconsistencies between multiple sources. Also, one of the source datasets, 30Music, was originally created for session-based recommendations instead of playlist continuation.

In this paper, we try to overcome the limitations of the existing datasets. Our main contribution is to provide a large research dataset of commercial music with quality playlist and tag information that includes audio representations suitable for audio-based approaches. Furthermore, our dataset is different because it represents music consumption in Korea instead of Western countries, bringing more cultural diversity in MIR research applied to music consumption platforms.

3. MELON PLAYLIST DATASET

All the data was originally collected from Melon for a playlist continuation challenge that took place on the Kakao Arena[4] platform between April and July 2020 with participation of 786 teams. The dataset consists of 649,091 tracks, represented by their mel-spectrograms, and 148,826 playlists with annotations by 30,652 different tags. The playlists were created and annotated by selected users recognized for the quality of their submissions. These users are named Melon DJs on the platform after Melon moderators verify them for the quality of the playlist metadata (titles, tags, and genres) they provide.

The mel-spectrograms were computed using Essentia[5] music audio analysis library [11] version *2.1b5.dev677* with the following settings: 16 KHz sample rate, frame and hop size of 512 and 256 samples, and Hann window function. The scripts for their computation are provided with the dataset.

To reduce distributable data size, we computed mel-spectrograms only for a segment of each song (20 to 50 seconds long, not adjacent to the start or the end of the songs). Furthermore, for copyright reasons, we used a reduced 48 mel-bands resolution, which did not negatively affect the performance of the auto-tagging approaches in our previous study [12], while having a significantly lower reconstructed audio quality. These decisions allow saving bandwidth and disk space required to transfer and store the dataset. The dataset is distributed in 40 files, 6 GB each, with a total download size of 240 GB.

The dataset also includes playlist and tracks metadata. Playlist metadata contains tags and titles submitted by playlist creators, the number of users who like the playlist, and the last modification date. Track metadata contains album, title, artists, release date, and genres. The statistics of the dataset are presented in Table 2.

Figure 1 shows the distribution of the tracks concerning their release year. Over 95% of the tracks in the dataset were published after the year 1990. Considering genre annotations, 25.45% tracks in the dataset belong to only Korean music genres, 38.44% tracks to non-Korean music genres, and 27.70% tracks to both Korean and non-Korean genres (8.39% tracks are annotated with music genres origin of which is unknown).

Playlists contain up to 200 tracks, with 41.46 tracks on average. The average of tags per playlist is 3.91 with a maxi-

[4] `https://arena.kakao.com/c/8`

[5] `https://essentia.upf.edu`

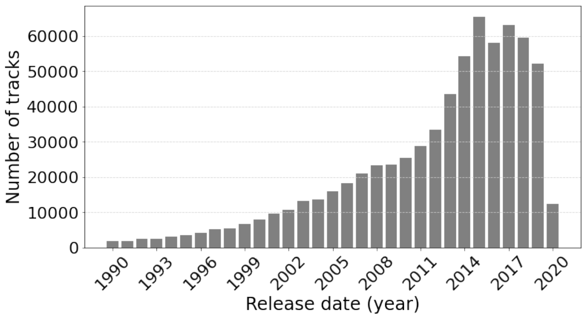

Fig. 1. The distribution of release year of all tracks.

Tracks	Tags	Title	Frequency
all	half	half	3860 (11.43%)
half	all	half	0 (0.00%)
half	half	all	13165 (39.00%)
all	all	half	2554 (7.56%)
all	half	all	2 (0.00%)
half	all	all	14168 (41.97%)
all	all	all	6 (0.01%)

Table 3. Number of playlists in test and validation sets for which the tracks, tags and title were hidden either entirely ("all") or for the half of the instances ("half").

mum of 11 tags. The number of different genres in a playlists on average is 6.31 with a maximum of 26. Figure 2 shows the distribution of number of tags, genres and tracks in the playlists.

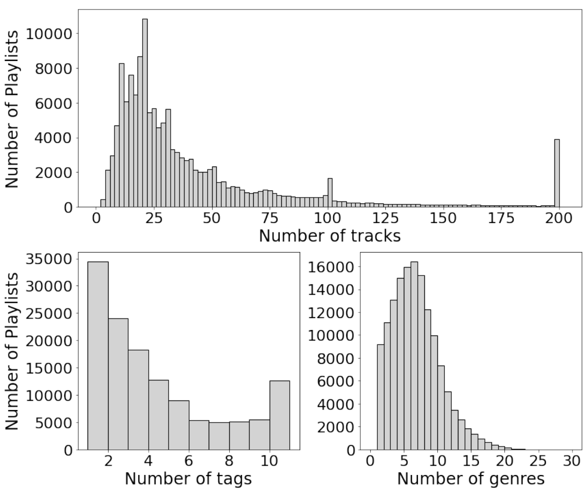

Fig. 2. Number of tracks, tags, and genres in playlists.

3.1. Kakao Arena challenge and the dataset split

In the context of the challenge we divided the playlists in three groups: 115,071 playlists (77.32%) in the train set, 23,015 playlists (15.46%) in the validation set, and 10,740 playlists (7.22%) in the test set. For the 33,755 playlists in validation and test sets, we considered either fully or partially hiding the tags, titles and tracks metadata. Table 3 shows the total number of playlists for each of these problem cases. The goal of the challenge was to predict the missing tracks and tags for the playlists in the test set.

Even though the challenge has finished, the Kakao Arena evaluation platform remains open for submissions of the predicted tracks and tags for the APC and auto-tagging tasks. In this way, it offers the possibility to the research community to benchmark new approaches in a standardized way using the test set with hidden tracks and tags.

4. AUTOMATIC PLAYLIST CONTINUATION

Melon Playlist Dataset offers many research possibilities. The most direct are playlists generation and auto-tagging for which it was originally created.

The task of APC consists on recommending a list of tracks to continue a given playlist. Many approaches had been proposed for this task including collaborative and content-based [13, 14]. Collaborative filtering approaches usually offer the best performance according to offline metrics in the task of track recommendations to users. Given that it is not possible to recommend items without any previous interaction with these approaches (the cold-start problem), in the last years deep learning approaches have been proposed to overcome this problem by predicting the collaborative representations from audio [15, 16]. Melon Playlist Dataset is the first public dataset to contain playlist information together with directly available audio information of the tracks on a large scale, allowing to experiment with such audio-based approaches.

In what follows, we provide an example of an audio-based APC approach, allowing us to expand a playlist with previously unseen tracks. We focus on underrepresented tracks in our evaluation, which is different from the Kakao Arena challenge, where the tracks in the test set had significantly more associated track-playlist interactions available for collaborative filtering. For this reason, and for reproducibility outside the Kakao Arena platform, we create an alternative split.

4.1. Method

We created a subset of Melon Playlist Dataset, discarding the playlists with less than 5 tracks. For each playlist we split its track-playlist interactions, using the tracks that appear at least in 10 playlists for our training set (*APC-train*) and the rest of the tracks (considered cold-start tracks) for testing (*APC-test*). The *APC-train* subset contains interactions for a total of 104,645 playlists and 81,219 tracks.

Similar to Van den Oord et al. [16], we train a Matrix Factorization (MF) model on the APC-train track-playlist matrix using WARP loss function [17] and optimizing the parameters on 10% of the training interactions.

The MF model outputs the latent factors of the tracks and playlists in APC-train, we train an audio model to predict these track factors from mel-spectrograms provided in the Melon Playlist Dataset. To this end, we split the tracks in APC-train into *APC-train-train* (90%) for training and *APC-train-val* (10%) for validation. We use a fully-convolutional neural network common for auto-tagging, based on VGGish architecture [18] and trained with Mean Squared Error (MSE) as a loss function. We observed reasonable approximation of the CF track factors by the audio model, with the MSE of 0.0098.

Once trained, we apply the model to predict latent factors for the cold-start tracks in APC-test and match those factors to the playlist factors [14] in APC-train to generate rankings of the best tracks to expand those playlists. We evaluate the top-10 and top-200 rankings using MAP and nDCG [19] and the rest of playlist-track interactions kept as ground truth in *APC-test* for the playlists.

Method	MAP@10	nDCG@10	MAP@200	nDCG@200
Random	0.0000	0.0001	0.0001	0.0010
Audio	0.0159	0.0395	0.0135	0.0516
CF	0.0165	0.0414	0.0148	0.0545

Table 4. Performance on APC-train-val.

4.2. Results

In all evaluations we compare the audio approach to the random baseline and the collaboration filtering approach used as our lower-bound and upper-bound baselines, respectively. Table 4 shows the performance on the validation set (APC-train-val). Comparing the performance of latent factors predicted from audio with the ones from the MF model itself, we see that the performance of both is very similar, which shows that the audio-based approach can be used to predict latent factors for unseen tracks.

For the collaborative filtering baseline on APC-test, we use all interactions in APC-train together with 70% of the interactions in the APC-test to train the MF model and the other 30% to evaluate. Some test tracks are discarded from evaluation due to this split. For consistency, we use the same set of test tracks for evaluation of the rest of the approaches.

Table 6 shows the overall performance using all considered tracks in APC-test for ranking. In addition, we independently evaluated three subsets of APC-test described in Table 5, generating separate ranking lists among the tracks with different popularity (or "cold-startness") level in the dataset. The results on these subsets are given as an additional reference, but they aren't directly comparable as the performance is measured on ranking lists of different track sets.

Test subset	Track in # playlist	Tracks	Playlists
APC-test-1	8-9	17,042	27,229
APC-test-2	5-8	46,069	35,910
APC-test-3	2-5	155,688	31,925

Table 5. Track frequency based subsets of the APC-test set.

	Method	MAP@10	nDCG@10	MAP@200	nDCG@200
APC-test	Random	0.0000	0.0000	0.0000	0.0002
	Audio	0.0007	0.0014	0.0010	0.0052
	CF	0.0802	0.1338	0.0581	0.1099
APC-test-1	Random	0.0001	0.0003	0.0003	0.0022
	Audio	0.0041	0.0065	0.0063	0.0267
	CF	0.0846	0.1200	0.0979	0.1923
APC-test-2	Random	0.0000	0.0000	0.0001	0.0009
	Audio	0.0022	0.0038	0.0032	0.0136
	CF	0.0490	0.0745	0.0582	0.1291
APC-test-3	Random	0.0000	0.0000	0.0000	0.0002
	Audio	0.0001	0.0001	0.0001	0.0002
	CF	0.0274	0.0416	0.0341	0.0756

Table 6. Performance on APC-test.

5. CONCLUSIONS

We presented Melon Playlist Dataset, the first public large-scale dataset of commercial music including the playlists, audio representation, and tags altogether, submitted by users verified for their quality annotations. Since the dataset reflects the music consumption in Korea, it offers novel opportunities to diversify MIR research.

The dataset has various applications. As an example, we considered automatic playlist continuation in a cold-start scenario and trained a baseline model to predict the latent factors of collaborative filtering from mel-spectrograms. All the code to reproduce this experiment, including the generation of dataset splits, is available online.[6]

Our dataset's main limitation is that it provides mel-spectrograms instead of audio, making it impossible to apply methods based on other audio representations (e.g., raw waveforms). Nevertheless, the provided mel-spectrograms are suitable for the tasks of auto-tagging and automatic playlist continuation, which are the main focus of the proposed dataset. They offer a good trade-off considering the common limitations of re-using copyrighted commercial music in the field of MIR and audio signal processing. Besides, due to the large scale of the dataset, the reduced audio representations lower its distributable size, facilitating transfer and storage.

[6]https://github.com/andrebola/icassp2021

6. REFERENCES

[1] Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie, "Evaluation of algorithms using games: The case of music tagging," in *Proc. of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, 2009, pp. 387–392.

[2] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere, "The million song dataset," in *Proc. of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[3] Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark Sandler, "The effects of noisy labels on deep convolutional neural networks for music tagging," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 139–149, 2018.

[4] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson, "Fma: A dataset for music analysis," in *Proc. 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017, number CONF.

[5] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra, "The mtg-jamendo dataset for automatic music tagging," in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States, 2019.

[6] Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani, "Recsys challenge 2018: Automatic music playlist continuation," in *Proc. of the 12th ACM Conference on Recommender Systems*, 2018, p. 527–528.

[7] Felipe Falcao and Daniel Mélo, "The million playlists songs dataset: a descriptive study over multiple sources of user-curated playlists," in *16th Brazilian Symposium on Computer Music*, 2017.

[8] Brian McFee and Gert RG Lanckriet, "Hypergraph models of playlist dialects," in *Proc. 13th International Society for Music Information Retrieval Conference (ISMIR)*. Citeseer, 2012, vol. 12, pp. 343–348.

[9] Martin Pichl, Eva Zangerle, and Günther Specht, "Towards a context-aware music recommendation approach: What is hidden in the playlist name?," in *IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2015, pp. 1360–1365.

[10] Roberto Turrin, Massimo Quadrana, Andrea Condorelli, Roberto Pagano, and Paolo Cremonesi, "30music listening and playlists dataset," *Poster Proc. ACM Conference on Recommender Systems*, vol. 15, 2015.

[11] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Herrera Boyer, Oscar Mayor, Gerard Roma Trepat, Justin Salamon, José Ricardo Zapata González, Xavier Serra, et al., "Essentia: An audio analysis library for music information retrieval," in *Proc. of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, 2013.

[12] Andres Ferraro, Dmitry Bogdanov, Xavier Serra, Jay Ho Jeon, and Jason Yoon, "How low can you go? reducing frequency and time resolution in current cnn architectures for music auto-tagging," in *Proc. 28th European Signal Processing Conference (EUSIPCO)*, 2020.

[13] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi, "Current challenges and visions in music recommender systems research," *International Journal of Multimedia Information Retrieval*, vol. 7, no. 2, pp. 95–116, 2018.

[14] Hamed Zamani, Markus Schedl, Paul Lamere, and Ching-Wei Chen, "An analysis of approaches taken in the acm recsys challenge 2018 for automatic music playlist continuation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 5, pp. 1–21, 2019.

[15] Dawen Liang, Minshu Zhan, and Daniel PW Ellis, "Content-aware collaborative music recommendation using pre-trained neural networks.," in *Proc. of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, 2015, pp. 295–301.

[16] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen, "Deep content-based music recommendation," in *Advances in neural information processing systems*, 2013, pp. 2643–2651.

[17] Jason Weston, Samy Bengio, and Nicolas Usunier, "Wsabie: scaling up to large vocabulary image annotation," in *Proc. of the 22nd international joint conference on Artificial Intelligence-Volume Volume Three*, 2011, pp. 2764–2770.

[18] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho, "Convolutional recurrent neural networks for music classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2392–2396.

[19] Francesco Ricci, Lior Rokach, and Bracha Shapira, "Introduction to recommender systems handbook," in *Recommender systems handbook*, pp. 1–35. Springer, 2011.