The MediaEval 2018 AcousticBrainz Genre Task: Content-based Music Genre Recognition from Multiple Sources

Dmitry Bogdanov[1], Alastair Porter[1], Julián Urbano[2], Hendrik Schreiber[3]
[1]Music Technology Group, Universitat Pompeu Fabra, Spain
[2]Multimedia Computing Group, Delft University of Technology, Netherlands
[3]tagtraum industries incorporated
dmitry.bogdanov@upf.edu,alastair.porter@upf.edu,urbano.julian@gmail.com,hs@tagtraum.com

ABSTRACT

This paper provides an overview of the AcousticBrainz Genre Task organized as part of the MediaEval 2018 Benchmarking Initiative for Multimedia Evaluation. The task is focused on content-based music genre recognition using genre annotations from multiple sources and large-scale music features data available in the AcousticBrainz database. The goal of our task is to explore how the same music pieces can be annotated differently by different communities following different genre taxonomies, and how this should be addressed by content-based genre recognition systems. We present the task challenges, the employed ground-truth information and datasets, and the evaluation methodology.

1 INTRODUCTION

Content-based music genre recognition is a popular task in Music Information Retrieval research [12]. The goal is to build systems able to predict genre and subgenre of unknown music recordings (tracks or songs) using music features of those recordings automatically computed from audio. Such research can be supported by our recent developments in the context of the AcousticBrainz [1] project, which facilitates access to large datasets of music features [9] and metadata [10]. AcousticBrainz is a community database containing music features extracted from audio files. Users who contribute to the project run software on their computers to process their personal audio collections and submit music features to the AcousticBrainz database. Based on these features, additional metadata including genres can then be mined for recordings in the database.

Since MediaEval 2017 [4], we have proposed a new genre recognition task using datasets based on AcousticBrainz. This task is different from a typical genre recognition task in the following ways:

- It allows us to explore how the same music can be annotated differently by different communities who follow different genre taxonomies, and how this can be addressed when developing and evaluating genre recognition systems.
- Genre recognition is often treated as a single category classification problem. Our data is intrinsically multi-label, so we propose to treat genre recognition as a multi-label classification problem.
- Previous research typically used a small number of broad genre categories. In contrast, we consider more specific genres and subgenres. Our data contains hundreds of subgenres.
- We provide information about the hierarchy of genres and subgenres within each annotation source. Systems can take advantage of this knowledge.
- MIR research is often performed on small music collections. We provide a very large dataset with two million recordings annotated with genres and subgenres. However, we only provide precomputed features, not audio.

MediaEval'18, 29-31 October 2018, France

2 TASK DESCRIPTION

The task invites participants to predict the genre and subgenre of unknown music recordings given automatically computed music features of those recordings. We provide four datasets of such music features taken from the AcousticBrainz database [9] together with four different ground truths created using four different music metadata websites as sources. Their genre taxonomies vary in class spaces, specificity and breadth. Each source has its own definition for its genre labels, i.e., the same label may carry a different meaning when used by another source. Most importantly, annotations in each source are multi-label: there may be multiple genre and subgenre annotations for the same music recording. It is guaranteed that each recording has at least one genre label, while subgenres are not always present.

Participants must train model(s) using the provided development sets and then predict genre and subgenre labels for the test sets. The task includes two subtasks:

- **Subtask 1: Single-source Classification**. This subtask explores conventional systems, each one trained on a single dataset. Participants submit predictions for the test set of each dataset separately, using their respective class spaces (genres and subgenres). These predictions will be produced by a separate system for each dataset, trained without any information from the other sources. This subtask will serve as a baseline for Subtask 2.
- **Subtask 2: Multi-source Classification**. This subtask explores the combination of several ground-truth sources to create a single classification system. We use the same four test sets. Participants submit predictions for each test set separately, again following each corresponding genre class space. These predictions may be produced by a single system for all datasets or by one system for each dataset. Participants are free to make their own decision about how to combine the training data from all sources.

3 DATA

3.1 Genre Annotations

We provide four datasets containing genre and subgenre annotations extracted from four different online metadata sources:

Table 1: Overview of the development datasets.

Dataset	AllMusic	Discogs	Lastfm	Tagtraum
Type	Explicit	Explicit	Tags	Tags
Annotation level	Release	Release	Track	Track
Recordings	1,353,213	904,944	566,710	486,740
Release groups	163,654	118,475	115,161	69,025
Genres	21	15	30	31
Subgenres	745	300	297	265
Genres/track	1.33	1.37	1.14	1.13
Subgenres/track	3.15	1.69	1.28	1.72

- **AllMusic** [2] and **Discogs** [6] are based on editorial metadata databases maintained by music experts and enthusiasts. These sources contain explicit genre/subgenre annotations of music releases (albums) following a predefined genre taxonomy. To build the datasets we assumed that release-level annotations correspond to all recordings in AcousticBrainz for that release.
- **Lastfm** [8] is based on a collaborative music tagging platform with large amounts of genre labels provided by its users for music recordings. **Tagtraum** [13] is similarly based on genre labels collected from users of the music tagging application beaTunes [3]. We have automatically inferred a genre/subgenre taxonomy and annotations from these labels following the algorithm proposed in [11] and a manual post-processing.

We provide information about genre/subgenre tree hierarchies for every ground truth.[1]

3.2 Music Features

We provide music features precomputed from audio for every music recording. All features are taken from the AcousticBrainz database and were extracted from audio using Essentia, an open-source library for music audio analysis [5]. The provided features are explained online.[2] Only statistical characterization of time frames is provided (bag of features), that is, no frame-level data is available.

3.3 Development and Test Datasets

We provide four developement datasets, four validation datasets and four test datasets which will be used in the final evaluation of all submissions. The test datasets do not include any groundtruth and have been anonymized. The datasets were created by a random split of the full data ensuring that:

- no recordings appear in more than one of the above sets;
- no recordings in any set are from the same release groups (e.g., albums, singles, EPs) present in other sets;
- the same genre and subgenre labels are present in all sets for each ground truth;
- genre and subgenre labels are represented by at least 40 and 20 recordings from 6 and 3 release groups in development and validation/test sets, respectively.

The approximate split ratios of the datasets are 70% for training, 15% for validation, and 15% for testing. The validation dataset was previously used as the test set in the 2017 edition of the task and is now available for participants for validation as a reference for benchmarking across all current and future editions of the task[3].

Table 1 provides an overview of the resulting development sets. Details on the genre/subgenre taxonomy and their distribution in the development sets in terms of number of recordings and release groups are reported online.[4] Recordings are partially intersected (annotated by all four ground truths) in the development and test sets. The full intersection of all development sets contains 247,716 recording, while the intersection of the two largest sets, AllMusic and Discogs, contains 831,744 recordings.

All data are published in JSON and TSV formats; details about format are available online.[5] Each recording in the development sets is identified by a MusicBrainz ID (MBID)[6], which can be used by participants to gather related data. Importantly, our split allows to avoid the "album effect" [7], which leads to a potential overestimation of the performance of a system when a test set contains recordings from the same albums as the development set. The development sets additionally include information about release groups of each recording, which may be useful for participants in order to avoid this effect when developing their systems. Partitioning scripts were provided to create training-validation splits ensuring these characteristics in the data.

4 SUBMISSIONS AND EVALUATION

Participants are expected to submit predictions for both subtasks. We allow a maximum of five evaluation runs, each including both subtasks, and reporting whether they used the whole development dataset or only parts for every submission.

The evaluation is carried out for each dataset separately. We do not use hierarchical measures because the hierarchies in the Lastfm and Tagtraum datasets are not explicit. Instead, we compute precision, recall and F-score at different levels:

- Per recording, all labels.
- Per recording, only genre labels.
- Per recording, only subgenre labels.
- Per label, all recordings.
- Per genre label, all recordings.
- Per subgenre label, all recordings.

The ground truth does not necessarily contain subgenre annotations for some recordings, so we only considered recordings containing subgenres for the evaluation at the subgenre level. An example can be found online in the summaries of random baselines.[7] We also provided evaluation scripts for development purposes.

5 CONCLUSIONS

Bringing the AcousticBrainz Genre Task to MediaEval we hope to benefit from contributions and expertise of a broader machine learning and multimedia retrieval community. We refer to the MediaEval

[1] The resulting genre metadata is licensed under CC BY-NC-SA4.0 license, except for data extracted from the AllMusic database, which is released for non-commercial scientific research purposes only. Any publication of results based on the data extracts of the AllMusic database must cite AllMusic as the source of the data.

[2] http://essentia.upf.edu/documentation/streaming_extractor_music.html

[3] https://multimediaeval.github.io/2017-AcousticBrainz-Genre-Task/results/

[4] https://multimediaeval.github.io/2018-AcousticBrainz-Genre-Task/data_stats/

[5] https://multimediaeval.github.io/2018-AcousticBrainz-Genre-Task/data/

[6] https://musicbrainz.org/doc/MusicBrainz_Identifier

[7] https://multimediaeval.github.io/2018-AcousticBrainz-Genre-Task/baseline/

2018 proceedings for further details on the methods and results of teams participating in the task.

ACKNOWLEDGMENTS

We thank all contributors to AcousticBrainz. This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 688382 (AudioCommons) and 770376-2 (TROMPA), as well as the Ministry of Economy and Competitiveness of the Spanish Government (Reference: TIN2015-69935-P). We also thank tagtraum industries for providing the Tagtraum genre annotations.

REFERENCES

[1] AcousticBrainz. 2017. (2017). https://acousticbrainz.org
[2] AllMusic. 2017. (2017). https://allmusic.com
[3] beaTunes. 2017. (2017). https://www.beatunes.com
[4] Dmitry Bogdanov, Alastair Porter, Julián Urbano, and Hendrik Schreiber. 2017. The MediaEval 2017 AcousticBrainz Genre Task: Content-based Music Genre Recognition from Multiple Sources. In *MediaEval Benchmark Workshop*.
[5] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J.R. Zapata, and X. Serra. 2013. Essentia: An Audio Analysis Library for Music Information Retrieval. In *International Society for Music Information Retrieval (ISMIR'13) Conference*. Curitiba, Brazil, 493–498.
[6] Discogs. 2017. (2017). https://discogs.com
[7] A. Flexer and D. Schnitzer. 2009. Album and Artist Effects for Audio Similarity at the Scale of the Web. In *Sound and Music Computing Conference (SMC'09)*.
[8] Lastfm. 2017. (2017). https://last.fm
[9] A Porter, D Bogdanov, R Kaye, R Tsukanov, and X Serra. 2015. AcousticBrainz: a community platform for gathering music information obtained from audio. In *International Society for Music Information Retrieval (ISMIR'15) Conference*. Málaga, Spain, 786–792.
[10] A. Porter, D. Bogdanov, and X. Serra. 2016. Mining metadata from the web for AcousticBrainz. In *International workshop on Digital Libraries for Musicology (DLfM'16)*. ACM, 53–56.
[11] H. Schreiber. 2015. Improving genre annotations for the Million Song Dataset. In *International Society for Music Information Retrieval (ISMIR'15) Conference*. Málaga, Spain, 241–247.
[12] B. L. Sturm. 2014. The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval. *Journal of New Music Research* 43, 2 (2014), 147–172.
[13] Tagtraum. 2017. (2017). http://www.tagtraum.com