The MediaEval 2017 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,j.urbano@tudelft.nl,hs@tagtraum.com

ABSTRACT

This paper provides an overview of the AcousticBrainz Genre Task organized as part of the MediaEval 2017 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 within the Music Information Retrieval (MIR) community [6]. The goal is to build systems that are able to predict genre and subgenre of unknown music recordings (tracks or songs) using music features of those recordings automatically computed from audio. Although this task is commonly addressed in the MIR research, we realize that it is often oversimplified while it should be considered in more complexity resembling real-world use cases. Such research can be supported by our recent developments in the context of the AcousticBrainz[1] project facilitating access to large datasets of music features [3] and metadata [4]. AcousticBrainz is a community database containing MIR features extracted from audio files. Users who contribute to the project run software on their computers to analyze their personal audio collections and submit the analysis to the AcousticBrainz database. Additional metadata including genres can then be mined for recordings in the database.

We propose a new genre recognition task using datasets based on AcousticBrainz for MediaEval 2017. This task is different from a typical genre recognition task in the following:

- There are different genre taxonomies and people may not always agree on the meaning of genres. Genres labels are probably subjective categories. We want to explore how the same music can be annotated differently by different communities following different genre taxonomies, and how this should be addressed by genre recognition systems.
- Typically research is done on a small number of broad genre categories. In contrast, we propose to consider more specific genres and subgenres and our data contains hundreds of subgenres.
- Genre recognition is often treated as a single category classification problem. Our genre data is intrinsically multi-label and so we propose to treat genre recognition as a multi-label classification problem.
- Typically research is done on small music collections. Instead, we provide a very large dataset counting two million recordings annotated by genres and subgenres. The downside is that we are not able to provide audio, but only precomputed bags of features.
- Finally, we provide information about the hierarchy of genres and subgenres within each genre annotation source. Systems can take advantage of this knowledge.

2 TASK DESCRIPTION

The task invites participants to predict genre and subgenre of unknown music recordings (tracks) given automatically computed features of those recordings. We provide a training set of such music features taken from the AcousticBrainz database [3] together with four different ground truths of genre and subgenre labels. These ground truths were created using as a source four different music metadata websites. Their genre taxonomies vary in class spaces, specificity and breadth. Each source has its own definition for its genre labels meaning that these labels may be different between sources. 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 this data and then generate predictions of genre and subgenre labels for a test set following genre taxonomy of each ground truth. The task includes two subtasks:

- **Subtask 1: Single-source Classification**. This subtask explores conventional systems, each one trained on a single dataset. Participants will submit predictions for the test set of each dataset separately, following 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 how to combine several ground-truth sources to create a classification system. We will use the same four test sets. Participants will 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

[1]https://acousticbrainz.org

MediaEval'17, 13-15 September 2017, Dublin, Ireland

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: [2]

- **AllMusic**[3] and **Discogs**[4] 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 namespace and taxonomy. We propagated release-level annotations to recordings (tracks) in AcousticBrainz present on those releases to build the datasets.
- **Lastfm**[5] and **Tagtraum**[6] are based on collaborative music tagging platforms with large amounts of genre labels provided by their users for music recordings (tracks). We have automatically inferred a genre/subgenre taxonomy and annotations from these labels following the algorithm proposed in [5] and a manual post-processing.

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

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 [1]. The provided features are explained online. [7] 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

In total we provide four development and four test datasets associated with the four genre ground truths. They were created by a random split of the full data ensuring that:

- no recordings from test sets are also present in any of the development sets;
- no recordings from test sets are from the same release groups (e.g., albums, singles, EPs) as the development sets;
- the same genre and subgenre labels are present in both development and test 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 test sets, respectively.

The approximate split ratio is 70% to 15% of the total number of recordings (another 15% were reserved for further evaluation purposes). Table 1 provides an overview of the resulting development sets. Genre/subgenre taxonomy and their distribution in the development sets in terms of number of recordings and release groups are reported online.[8] Recordings are partially intersected in all four development sets as well as in the test sets. The full intersection of all development sets (recordings annotated by all four ground truths) contains 247716 recordings while the intersection of the two largest sets, AllMusic and Discogs, contains 831744 recordings.

Table 1: Overview of the development datasets

Dataset	AllMusic	Discogs	Lastfm	Tagtraum
Type	Explicit	Explicit	Tags	Tags
Annotation level	Release	Release	Track	Track
Tracks (recordings)	1353213	904944	566710	486740
Release groups	163654	118475	115161	69025
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

The details on the format of all distributed data are available online.[9] Each recording (a track or song) in the development sets is identified with a MusicBrainz identifier (MBID).[10] Importantly, our split allows to avoid the "album effect" [2] which consists in a potential overestimation of the performance of a system when a test set contains recordings from the same albums as the training 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.

4 SUBMISSIONS AND EVALUATION

Participants are expected to submit predictions for both subtasks. If they only want to work on the first subtask, they should submit the same predictions for the second subtask. We allow only five evaluation runs (each run includes both subtasks). Participants should report whether they used the whole development dataset or only its part for every submission.

The evaluation is carried out for each dataset separately. In particular, we compute precision, recall and F-score as follows:

- 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. Therefore, only recordings containing subgenres are considered for the evaluation on the subgenre level. An example can be found online in the summaries of random baselines. [11] We provide evaluation scripts for development purposes.[12]

[2] 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.
[3] https://allmusic.com
[4] https://discogs.com
[5] https://last.fm
[6] https://tagtraum.com
[7] http://essentia.upf.edu/documentation/streaming_extractor_music.html
[8] https://multimediaeval.github.io/2017-AcousticBrainz-Genre-Task/data_stats/
[9] https://multimediaeval.github.io/2017-AcousticBrainz-Genre-Task/data/
[10] https://musicbrainz.org/doc/MusicBrainz_Identifier
[11] https://multimediaeval.github.io/2017-AcousticBrainz-Genre-Task/baseline/
[12] Accessible to registered participants.

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 2017 proceedings for further details on the methods and results of teams participating in the task.

ACKNOWLEDGMENTS

We thank all contributors to AcousticBrainz and Tagtraum. This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688382 (AudioCommons).

REFERENCES

[1] 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*. 493–498.

[2] 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)*.

[3] Alastair Porter, Dmitry Bogdanov, Robert Kaye, Roman Tsukanov, and Xavier Serra. 2015. AcousticBrainz: a community platform for gathering music information obtained from audio. In *International Society for Music Information Retrieval (ISMIR'15) Conference*.

[4] 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.

[5] H. Schreiber. 2015. Improving genre annotations for the Million Song Dataset. In *International Society for Music Information Retrieval Conference (ISMIR'15)*.

[6] Bob L Sturm. 2013. Classification accuracy is not enough: On the evaluation of music genre recognition systems. *Journal of Intelligent Information Systems* (2013).