THE ACOUSTICBRAINZ GENRE DATASET
MULTI-SOURCE, MULTI-LEVEL, MULTI-LABEL, AND LARGE-SCALE

Bogdanov, Dmitry; Porter, Alastair; Schreiber, Hendrik; Urbano, Julián; Oramas, Sergio

Publication date
2019

Document Version
Final published version

Published in
International Society for Music Information Retrieval Conference 2019

THE ACOUSTICBRAINZ GENRE DATASET: MULTI-SOURCE, MULTI-LEVEL, MULTI-LABEL, AND LARGE-SCALE

Dmitry Bogdanov[1] **Alastair Porter**[1] **Hendrik Schreiber**[2] **Julián Urbano**[3] **Sergio Oramas**[4]
[1] Music Technology Group, Universitat Pompeu Fabra, Spain
[2] tagtraum industries incorporated, USA
[3] Multimedia Computing Group, Delft University of Technology, Netherlands
[4] Pandora, USA
dmitry.bogdanov@upf.edu, alastair.porter@upf.edu, hs@tagtraum.com,
urbano.julian@gmail.com, soramas@pandora.com

ABSTRACT

This paper introduces the AcousticBrainz Genre Dataset, a large-scale collection of hierarchical multi-label genre annotations from different metadata sources. It allows researchers to explore how the same music pieces are annotated differently by different communities following their own genre taxonomies, and how this could be addressed by genre recognition systems. Genre labels for the dataset are sourced from both expert annotations and crowds, permitting comparisons between strict hierarchies and folksonomies. Music features are available via the AcousticBrainz database. To guide research, we suggest a concrete research task and provide a baseline as well as an evaluation method. This task may serve as an example of the development and validation of automatic annotation algorithms on complementary datasets with different taxonomies and coverage. With this dataset, we hope to contribute to developments in content-based music genre recognition as well as cross-disciplinary studies on genre metadata analysis.

1. INTRODUCTION

Content-based music genre recognition (MGR) is a popular task in Music Information Retrieval (MIR) research [27]. The goal is to build systems that can predict the genre or subgenre of unknown music recordings (tracks, songs) using music features automatically computed from audio of those recordings. Such research can be supported by recent developments in the context of the AcousticBrainz [1] project, which facilitates access to a large dataset of music features [21] and metadata [22]. AcousticBrainz is a community database containing music features extracted from over four million distinct audio files [2] uniquely identified by public MusicBrainz Identifiers (MBID) [3] and thus tied to rich textual metadata. Users who contribute to the project run software on their computers to process their personal music collections and submit features to the AcousticBrainz database. Based on these features, additional metadata not already included in MusicBrainz, like mood, tempo, key, and genres can be estimated from content-based features in the database.

To facilitate new research in MGR, we have curated four supplemental genre datasets mapped to recordings in AcousticBrainz and containing fine-grained, hierarchical genre annotations, derived from both crowdsourced labels and expert annotations. Each of the four datasets contains multiple labels featuring hundreds of subgenres covering in total over 2,086,000 recordings, which are connected to AcousticBrainz via MBIDs. We refer to the combination of the four datasets and the music features from AcousticBrainz as the AcousticBrainz Genre Dataset. The four main characteristics of this new dataset are:

- **Multi-source.** It allows us to explore how the same music can be annotated differently by communities who follow their own genre taxonomies, and how this can be addressed when developing and evaluating MGR systems. This is especially valuable, because it has been previously noted that the evaluation of MGR systems is difficult due to subjectivity in genre annotations, with little inter-annotator agreement [8]. We are not aware of any other dataset offering such a unique and comprehensive view on genres.
- **Multi-level.** We provide information about the hierarchy of genres and subgenres within each annotation source. Previous research typically used a small number of broad genre categories. According to Sturm's 2012 survey [26], the most popular public datasets for automatic genre recognition were GTZAN and ISMIR04 [7, 13, 28], with 10 and 6 genres, respectively. Only 3.7% of the surveyed systems used 25 or more la-

[1] https://acousticbrainz.org

[2] As of April 2019.

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

Dataset	GTZAN [28]	Rosamerica [14]	FMA [9]	USPop [1]	KPop [17]	MSD [2]	RWC [12]	Ballroom [13]	ISMIR04 [7]	Acousticbrainz
Recordings	1,000	400	106,574	7,000	1,894	1,000,000	100	698	729	692,217–1,935,991
Genres	10	8	16	10	7	No[1]	10	9	6	15–31
Subgenres	—	—	161	—	—	—	33	—	—	265–745
Hierarchical	No	No	Yes	No	No	No	Yes	No	No	Yes
Multi-Label	No	No	Yes	No	No	No	No	No	No	Yes
Audio	Yes	Yes[2]	Yes	No	No	No[3]	Yes	Yes	Yes	No
Public ID	No	No	Yes	No	No	Yes	No	No	No	Yes

[1] While the original dataset only contains free-form tags and no explicit genre labels, there have been several attempts to map MSD-tracks to genres [10, 23, 24]. [2] Available upon request. [3] 7-Digital previews have been available.

Table 1: Popular genre recognition datasets, compared to the proposed AcousticBrainz Genre Dataset.

bels. In contrast, our dataset contains dozens of genres and hundreds of subgenres.

- **Multi-label.** Genre recognition is often treated as a single category classification problem, likely because existing datasets are often single-label (e.g., GTZAN [28] or Ballroom [13]; see Table 1). Yet, previous studies suggest that if there is a diversity of responses in terms of genre labels to any particular recording, the standard evaluation methodology that uses single genre category as ground truth is not adequate [8, 20]. Our data is intrinsically multi-label, which allows treating genre recognition as a multi-label classification problem.
- **Large-scale.** MIR research is often performed on small music collections. We provide a very large dataset with audio features for over two million recordings annotated with genres and subgenres. However, we only provide precomputed features, not audio.

Compared to popular MGR datasets (see Table 1), the AcousticBrainz Genre Dataset is unique in that it is the only one that has all of these characteristics, which opens up interesting research opportunities. The remainder of the paper is structured as follows. We describe the dataset in detail in Section 2. In Section 3, we report on how the data has already been used for a task held within MediaEval 2017–18 [3, 4]. Section 4 describes a baseline implementation, and finally Section 5 presents our conclusions.

2. DATASET

The AcousticBrainz Genre Dataset dataset consists of genre annotations (Section 2.1) and precomputed music features (Section 2.2), distributed in predefined splits (Section 2.3). All related information about the dataset including downloads, data format, and baselines is available online. [4]

2.1 Genre Annotations

We provide four datasets with genre and subgenre annotations extracted from different online metadata sources. Two sources feature expert annotations using a strict taxonomy, two others use free-form tags from users: [5]

- **AllMusic** [6] and **Discogs** [7] are based on editorial metadata databases maintained by music experts and enthusiasts. These sources contain explicit genre/subgenre annotations of music albums following predefined genre taxonomies. To build the datasets we assumed that the annotations for an album also correspond to all of the recordings it contains. AllMusic data has been previously used [23] to provide genre annotations for the Million Song Dataset [2], while Discogs has been recently proposed as an alternative source of genre metadata for MIR [5]. To retrieve annotations from these sources we used the artist, album name and year metadata associated with each recording in AcousticBrainz. AllMusic has no publicly available API, and therefore we used a scraper to parse HTML data directly from the website. For Discogs, its public API was used. Annotations in AllMusic contain up to three levels of hierarchy, which we simplified to two levels by taking the most generic and the most specific annotations.
- **Lastfm** [8] is based on a collaborative music tagging platform with large amounts of genre labels provided as folksonomy tags by its users for music recordings. **Tagtraum** [9] is similarly based on genre labels collected from users of the music tagging application beaTunes. [10] To retrieve labels from the Lastfm API and genre annotations from the Tagtraum database we queried them using used artist names and recording titles. We then automatically inferred a genre/subgenre taxonomy and annotations from these labels following the algorithm proposed in [24]. This procedure exploits the fact that co-occurrences for genres are usually asymmetrical. For example, while Alternative Rock almost always co-occurs with Rock, Rock does not necessarily co-occur with Alternative Rock. This lets us derive a hierarchy. We performed manual post-processing to consolidate spelling variations and to remove location and era names (e.g., "50s", "Canadian") or labels that were clearly not a genre (e.g., "awesomelyrics").

Each source's genre taxonomy varies in class space,

[4] https://mtg.github.io/acousticbrainz-genre-dataset

[5] 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.

[6] https://allmusic.com

[7] https://discogs.com

[8] https://last.fm

[9] http://www.tagtraum.com

[10] https://www.beatunes.com

Dataset	AllMusic	Discogs	Lastfm	Tagtraum
Type	Explicit	Explicit	Tags	Tags
Annotation level	Album	Album	Track	Track
Recordings	1,935,991	1,290,489	806,627	692,217
Release groups[11]	233,789	169,109	164,290	98,333
Genres	21	15	30	31
Subgenres	745	300	297	265
Genres/track	1.33	1.37	1.14	1.13
Subgenres/track	3.14	1.70	1.28	1.72

Table 2: Overview of the AcousticBrainz Genre Dataset. Data is split in 70/15/15% for training, validation and test.

specificity, and breadth, and has its own definitions for the classes (i.e., the same label may have different meanings in difference sources). 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, but subgenres are not always present.

Table 2 provides an overview of the entire AcousticBrainz Genre Dataset. The bottom rows show the size of the genre taxonomies in each source. Compared to the others, the AllMusic taxonomy comprises few genres, but is much richer in terms of subgenres. Conversely, the Tagtraum taxonomy has the most genres, but the least number of subgenres. Figure 1 shows the distributions of genres in all four sets, where we can appreciate clear biases towards pop, rock and electronic. [12] This bias seems less acute in the Discogs and Lastfm sets. Figure 2 shows how label counts are distributed in all four datasets. In terms of genres, most recordings are annotated with only one genre, with some having as many as 8 genres in AllMusic and Discogs. In terms of subgenres, most recordings in the simpler Tagtraum and Lastfm sets are annotated with 1 or 2 subgeners, but in the more complex AllMusic and Discogs sets we find 10 or more subgenre annotations for some recordings. We can see that the distribution in AllMusic is quite smooth, while in the other sets we see clear biases towards 1 genre and 1 or 2 subgenres. We did not aim to create a representative or unbiased dataset, instead collecting as much data as possible for recordings in AcousticBrainz. We understand that biases likely exist due to the coverage of MusicBrainz, AcousticBrainz, and the sources of genre information.

A more detailed picture of the complexity and similarity among datastets can be made in terms of entropy of the label distributions. In particular, we may compute the conditional entropy of a dataset X given another dataset Y:

$$H(X|Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(y)}\,, \qquad (1)$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the taxonomies of X and Y, respectively. Eqn (1) computes the amount of information needed to describe a recording in X given its labels in Y. For simplicity, we ignore the multi-label nature of the data and set $p(x)$ equal to the probability that a recording contains the label x, ignoring the other labels in the same recording. As a byproduct, this allows us to compute $H(X|X) \neq 0$, understood as the amount of information needed to fully describe a recording in X when some label in X is already known. To make this distinction explicit, let us refer to this as *conditional pseudo-entropy* $\tilde{H}$.

[11] https://musicbrainz.org/doc/Release_Group

[12] Details on the genre/subgenre taxonomies and their distributions are reported on the dataset website.

	Allmusic	Discogs	Lastfm	Tagtraum
Allmusic	59.6	39.6	28.9	33.3
Discogs	35.4	21.2	15.1	17.8
Lastfm	32.1	19.2	11.2	16.0
Tagtraum	29	17.7	11.6	10.6

(a) Genre and subgenre labels.

	Allmusic	Discogs	Lastfm	Tagtraum
Allmusic	1.94	2.40	1.62	1.49
Discogs	2.37	2.15	1.57	1.50
Lastfm	2.87	2.88	1.18	1.8
Tagtraum	2.09	2.00	1.17	0.67

(b) Only genre labels.

Table 3: Conditional pseudo-entropy $\tilde{H}(X|Y)$ between pairs of datasets, where X is the dataset in the row and Y the one in the column.

Table 3a shows the conditional pseudo-entropies when considering both genre and subgenre labels. As the diagonal shows, the AllMusic dataset is much more complex than the others, as anticipated by the high number of subgenres in the taxonomy and the smooth distribution shown in Figure 2. Interestingly, the Lastfm column shows that knowing Lastfm labels provides the most information when predicting labels in the other taxonomies, only surpassed by known labels in the target taxonomies (diagonals). Lastfm and Tagtraum are the most similar sets, with AllMusic and Discogs being the most dissimilar. This suggests that labels produced by different non-expert user communities and following no common guidelines, are more similar than those produced by different set of experts following different guidelines.

Table 3b shows similar results, but considering only the genre labels. The pseudo-entropies are orders of magnitude smaller because genres encode less information, and as a result relative differences among datasets are also smaller. Discogs is the most complex dataset because of its higher variability in the number of genres per recording (see rows in Figure 2), followed by AllMusic. This time, we see that Tagtraum provides the most information when predicting labels in another taxonomy. As before, the most similar sets are Lastfm and Tagtraum, and the most dissimilar are AllMusic and Discogs.

2.2 Music Features

We provide music features precomputed from audio for all music recordings. All features are taken from the

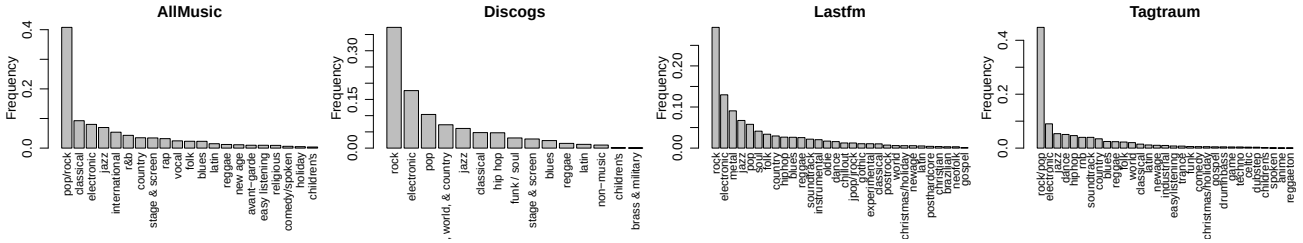


Figure 1: Distributions of genre labels.

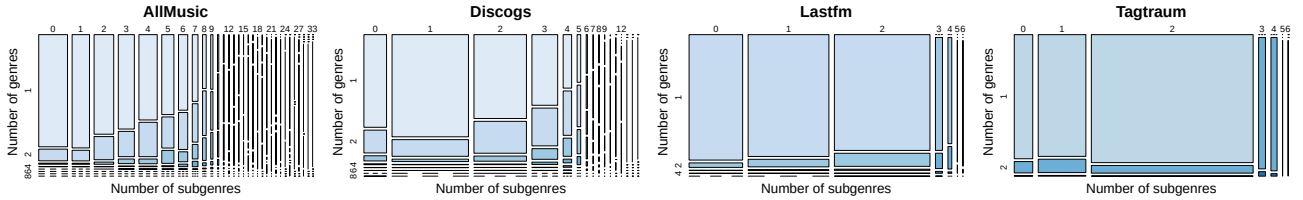


Figure 2: Distributions of label counts. Box heights represent the amount of recordings with the number of genre labels indicated in the row, and widths represent the amount of recordings with the number of subgenre labels in the column.

AcousticBrainz database and were extracted from audio using Essentia, an open-source library for music audio analysis [6]. They include features characterizing overall loudness, dynamics, and spectral shape of the signal, rhythm descriptors (including beat positions and BPM value), and tonal information (including chroma features, keys and scales). [13] Only a statistical characterization of time frames is provided (bag of features), that is, no frame-level data is available. The features for each recording are provided in a JSON file. [14]

2.3 Training, Validation and Test Sets

We provide four training sets and four validation sets with all data publicly available, and four test sets with a hidden ground truth. The training and validation sets can be used for the evaluation of MGR systems (Section 3.3). The test sets do not include a publicly available ground truth and have anonymized MBIDs; they are reserved for future MGR challenges. Nevertheless, it is possible to run an evaluation on the test sets upon request. [15]

The datasets were created by a random split of the full data ensuring that:

- No recording appears in more than one set;
- No recordings in any set are from the same release groups present in other sets (e.g., albums, singles, EPs);
- The same genre and subgenre labels are present in all three sets for the same source;
- Genre and subgenre labels are represented by at least 40 and 20 recordings from 6 and 3 release groups in training and validation/test sets, respectively.

The approximate split ratios of the datasets are 70% for training, 15% for validation, and 15% for testing. Partitioning scripts are provided to create training-validation splits ensuring these characteristics in the data. The four ground truths partially overlap. The full intersection of all training 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 the formats are available online. Each recording in the training and validation sets is identified by an MBID, which can be used by researchers to gather related data. Importantly, our split avoids the "album effect" [11], which leads to a potential overestimation of the performance of a system when a test set contains recordings from the same albums as the training set. We don't filter for the artist effect, in order to preserve some low-count tags and to address the fact that artists can release albums with different broad genres. MusicBrainz artist IDs allow researchers to perform this filtering if desired. The training sets additionally include information about release groups of each recording, which may be useful for researchers in order to avoid this effect when developing their systems.

3. RESEARCH TASK

MGR systems typically attempt to predict a single label per recording. Given that the AcousticBrainz Genre Dataset features multiple hierarchical labels from different sources, we suggest the following two subtasks designed for the datasets introduced in Section 2.

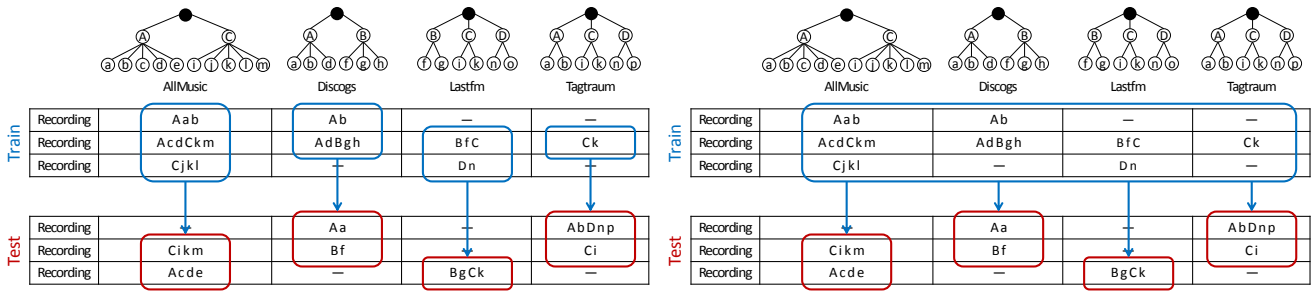
3.1 Subtask 1: Single-source Classification

This task, depicted in Figure 3a, explores conventional systems, each one trained on a single dataset. Researchers make 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

[13] More details are available online: `http://essentia.upf.edu/documentation/streaming_extractor_music.html`

[14] An example JSON file: `http://acousticbrainz.org/api/v1/6bb7e980-791c-44b5-9024-cc7c90bc8230/low-level?n=0`

[15] Please, contact the authors.



(a) Subtask 1: Single-source classification

(b) Subtask 2: Multi-source classification

Figure 3: Suggested tasks for the AcousticBrainz Genre Dataset.

other sources. This subtask can serve as a baseline for the multi-source classification task described below.

3.2 Subtask 2: Multi-source Classification

This task (Figure 3b) explores the combination of several ground-truth sources to train, but still make predictions for each test set separately, again following the corresponding genre class spaces. These predictions may be produced by a single system for all datasets or by one system for each dataset. Researchers are free to make their own decisions about how to combine the training data from all sources.

3.3 Evaluation

The development of an appropriate methodology that models each subtask as a single experiment with a "source" factor and replicated observations, is an interesting point that we leave for future research. For simplicity, we follow traditional evaluation on each test dataset separately, as if they were four independent experiments. As for metrics, we propose ROC AUC, precision, recall and F-score at the label level for a system-oriented view, and also at the recording level for a user-oriented view. We do not use hierarchical metrics because the hierarchies in the Lastfm and Tagtraum datasets are not explicit. Instead, we compute metrics at different levels:

- Per recording: using all labels, only genre labels, or only subgenre labels
- Per label: using all recordings
- Per genre label: using all recordings
- Per subgenre label: using 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. We provide evaluation scripts for development purposes and two simple baselines:

- Random baseline reproduces the joint distribution of labels as found in the training sets.
- Popularity baseline always predicts the most popular genre in the training set.

In the context of the MediaEval 2017–18 task, [16] researchers were expected to create predictions for both subtasks, reporting whether they used the entire data available for development or only its parts for every submission. Overall, we received over 100 submissions from 7 research teams covering both subtasks.

[16] Task details and evaluation results are available online: `https://multimediaeval.github.io/2018-AcousticBrainz-Genre-Task`

4. BASELINE

In this section we present our baseline approach for the proposed MGR tasks. This baseline employs an oversimplistic deep learning architecture for the single-source task and a fusion approach that demonstrates the possibilities of merging different genre ground truth sources in the multi-source task. To this end, we explore how stacking deep feature embeddings obtained on different datasets can benefit MGR systems. We propose an early fusion approach, similar to the one proposed in [19] for multi-modal genre classification. The approach incorporates knowledge across datasets by stacking deep feature embeddings learned on each dataset individually and using those as an input to predict genres for each dataset.

4.1 Input Features

We use all available features provided for the challenge. As a pre-processing step, we apply one-hot encoding for a few categorical features related to tonality (`key_key`, `key_scale`, `chords_key`, and `chords_scale`) and standardize all features (zero mean, unit variance). In total, this amounts to 2669 input features.

4.2 Neural Network Architecture

A simple feedforward network (extractor network) is used to predict the probabilities of each genre given a track. The network consists of an input layer of 2669 units (the size of the feature vector for an input recording), followed by a hidden dense layer of 256 units with ReLu activation, and the output layer where the number of units coincides with the number of genres to be predicted in each dataset. Dropout of 0.5 is applied after the input and the hidden layer. As the targeted genre classification task is multi-label, the output layer uses sigmoid activations and is evaluated with a binary cross-entropy loss.

Mini-batches of 32 items are randomly sampled from the training data to compute the gradient. The Adam [15] optimizer is used to train the models, with the suggested

Subtask	AllMusic	Discogs	Lastfm	Tagtraum
Single-source	0.648	0.759	0.828	0.802
Multi-source	0.812	0.886	0.906	0.887

Table 4: ROC AUC on validation datasets.

default parameters. The networks are trained for a maximum of 100 epochs with early stopping on validation loss. Once trained, we extract the 256-dimensional vectors from the hidden layer for the training, validation, and test sets.

The model architecture is used to train a multi-label genre classifier on each of the four datasets. The models are trained on 80% of the training set and validated after each epoch using the other 20% using the provided split script with release group filtering. Predictions are computed for the validation and test sets.

4.3 Embedding Fusion Approach

One model per dataset is trained. These models serve for predictions in Subtask 1. For Subtask 2, the given models are used as feature extractors. All four models share the same input format, so input feature vectors from one dataset can be used as input to a model trained on other datasets. For each model we feed all tracks from the training, validation and test sets of each dataset, and obtain the activations of the hidden layer as a 256-dimensional feature embedding. Therefore, for each track in each dataset we obtain four different feature embeddings, coming from each of the four previously trained models.

Given the four feature embeddings of each track, we apply the ℓ_2-norm to each of them and then stack them together into a single 1024-dimensional feature vector. We obtain new feature vectors for every track in the training, validation and test sets of each dataset. We use these feature vectors as input to a fusion network where the input layer is directly connected to the output layer. Dropout of 0.5 is applied after the input layer. The output layer is exactly the same as in the extractor network, where sigmoid activation and binary cross-entropy loss are applied. The fusion network is trained following the same methodology and partitions described for the extractor network. We train a fusion network per dataset, and obtain the genre probability predictions of the validation and test sets for Subtask 2.

4.4 Predictions Thresholding

The predictions made by each model are continuous, while the task requires binary prediction of genre labels. We apply a plug-in rule approach thresholding the prediction values to maximize the evaluation metrics. As an example, we decided to maximize the macro F-score, and applied thresholds individual for each genre label [18].

4.5 Results and Analysis

Full results and code for the baseline are available at the dataset website. Table 4 presents the ROC AUC metric on the validation sets. Table 5 presents the final results after applying thresholding. We can clearly see the benefit of models based on the embedding fusion approach compared to the models trained individually on each dataset. While the individual models (Subtask 1) are hardly usable, the combined models got a significant improvement in performance.

		Dataset			
		AllMusic	Discogs	Lastfm	Tagtraum
Single-source					
Per recording (all labels)	P	0.016	0.069	0.075	0.124
	R	0.579	0.538	0.446	0.507
	F	0.030	0.119	0.124	0.194
Per label (all labels)	P	0.023	0.076	0.074	0.097
	R	0.492	0.249	0.238	0.232
	F	0.032	0.095	0.095	0.115
Multi-source					
Per recording (all labels)	P	0.142	0.286	0.266	0.299
	R	0.475	0.545	0.476	0.513
	F	0.195	0.339	0.305	0.349
Per label (all labels)	P	0.065	0.108	0.115	0.127
	R	0.155	0.210	0.220	0.223
	F	0.074	0.122	0.133	0.140

Table 5: Precision, recall and F-scores on validation datasets produced by our baseline approach.

In our baseline, we focused on optimizing macro F-score, however choosing this metric for threshold optimization can have a negative effect on micro-averaged metrics. In the case of infrequent subgenre labels and an uninformative classifier, an optimal, but undesirable strategy may involve always predicting those labels [18]. Indeed, this was the case for the individual models, but the fusion models did not have this issue.

Overall, we may expect further improvements in performance by means of a more sophisticated network architecture (e.g., [16, 25]). The baseline is available online at the dataset webpage.

5. CONCLUSIONS

We have presented the AcousticBrainz Genre Dataset, a large-scale dataset of music features and hierarchical multi-label genre annotations from different sources. This is unique data for MIR research, as it allows researchers to explore how the same music pieces are annotated differently by different communities following their own genre taxonomies, and how this could be addressed by genre recognition systems. To this end, we have proposed a research task for building MGR systems based on music features available in the AcousticBrainz database and to explore how multiple sources of genre annotations can be combined by MGR systems. This task was already held within the MediaEval 2017–18 evaluation campaigns, and it may serve as an example of the development and validation of automatic annotation algorithms on complementary datasets with different taxonomies and coverage.

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.

6. REFERENCES

[1] Adam Berenzweig, Beth Logan, Daniel PW Ellis, and Brian Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, 28(2):63–76, 2004.

[2] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *International Society for Music Information Retrieval Conference*, 2011.

[3] Dmitry Bogdanov, Alastair Porter, Julián Urbano, and Hendrik Schreiber. The MediaEval 2017 AcousticBrainz Genre Task: Content-based Music Genre Recognition from Multiple Sources. In *MediaEval Benchmark Workshop*, 2017.

[4] Dmitry Bogdanov, Alastair Porter, Julián Urbano, and Hendrik Schreiber. The MediaEval 2018 AcousticBrainz Genre Task: Content-based Music Genre Recognition from Multiple Sources. In *MediaEval Benchmark Workshop*, 2018.

[5] Dmitry Bogdanov and Xavier Serra. Quantifying music trends and facts using editorial metadata from the discogs database. In *18th International Society for Music Information Retrieval Conference*, 2017.

[6] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, Jose R. Zapata, and Xavier Serra. Essentia: An audio analysis library for music information retrieval. In *International Society for Music Information Retrieval Conference*, 2013.

[7] Pedro Cano, Emilia Gómez, Fabien Gouyon, Perfecto Herrera Boyer, Markus Koppenberger, Bee Suan Ong, Xavier Serra, Sebastian Streich, and Nicolas Wack. ISMIR 2004 audio description contest. 2006.

[8] Alastair JD Craft, Geraint A. Wiggins, and Tim Crawford. How Many Beans Make Five? The Consensus Problem in Music-Genre Classification and a New Evaluation Method for Single-Genre Categorisation Systems. In *International Society for Music Information Retrieval Conference*, 2007.

[9] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *International Society for Music Information Retrieval Conference*, 2017.

[10] Sander Dieleman, Philémon Brakel, and Benjamin Schrauwen. Audio-based music classification with a pretrained convolutional network. In *International Society for Music Information Retrieval Conference*, 2011.

[11] Arthur Flexer and Dominik Schnitzer. Album and artist effects for audio similarity at the scale of the Web. In *Sound and Music Computing Conference*, 2009.

[12] Masataka Goto. Development of the RWC music database. In *Proceedings of the International Congress on Acoustics (ICA)*, 2004.

[13] Fabien Gouyon, Anssi P. Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle, and Pedro Cano. An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1832–1844, 2006.

[14] Enric Guaus i Termens. *Audio content processing for automatic music genre classification: descriptors, databases, and classifiers*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2009.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Khaled Koutini, Alina Imenina, Matthias Dorfer, Alexander Rudolf Gruber, and Markus Schedl. MediaEval 2017 AcousticBrainz Genre Task: Multilayer Perceptron Approach. In *MediaEval 2017 Workshop*, Dublin, Ireland, 2017.

[17] Jin Ha Lee, Kahyun Choi, Xiao Hu, and JH Downie. K-pop genres: A cross-cultural exploration. In *International Society for Music Information Retrieval Conference*, 2013.

[18] Zachary C Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. Optimal thresholding of classifiers to maximize f1 measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 225–239. Springer, 2014.

[19] Sergio Oramas, Francesco Barbieri, Oriol Nieto, and Xavier Serra. Multimodal deep learning for music genre classification. *Transactions of the International Society for Music Information Retrieval*, 1(1), 2018.

[20] Sergio Oramas, Oriol Nieto, Francesco Barbieri, and Xavier Serra. Multi-label music genre classification from audio, text, and images using deep features. *International Society for Music Information Retrieval Conference*, 2017.

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

[22] Alastair Porter, Dmitry Bogdanov, and Xavier Serra. Mining metadata from the web for AcousticBrainz. In *International workshop on Digital Libraries for Musicology*, 2016.

[23] Alexander Schindler, Rudolf Mayer, and Andreas Rauber. Facilitating comprehensive benchmarking experiments on the million song dataset. In *International Society for Music Information Retrieval Conference*, 2012.

[24] Hendrik Schreiber. Improving genre annotations for the Million Song Dataset. In *International Society for Music Information Retrieval Conference*, 2015.

[25] Hendrik Schreiber. MediaEval 2018 AcousticBrainz genre task: A CNN baseline relying on Mel-Features. In *Proceedings of the MediaEval 2018 Multimedia Benchmark Workshop*, Sophia Antipolis, France, 10 2018.

[26] Bob L. Sturm. A survey of evaluation in music genre recognition. In *International Workshop on Adaptive Multimedia Retrieval*, 2012.

[27] Bob L. Sturm. 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):147–172, 2014.

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