TISMIR

Meléndez-Catalán, B., et al. (2019). Open Broadcast Media Audio from TV: A Dataset of TV Broadcast Audio with Relative Music Loudness Annotations. *Transactions of the International Society for Music Information Retrieval,* 2(1), pp. 43–51. DOI: https://doi.org/10.5334/tismir.29

DATASET

Open Broadcast Media Audio from TV: A Dataset of TV Broadcast Audio with Relative Music Loudness Annotations

Blai Meléndez-Catalán[*,†], Emilio Molina[†] and Emilia Gómez[*,‡]

Open Broadcast Media Audio from TV (OpenBMAT) is an open, annotated dataset for the task of music detection that contains over 27 hours of TV broadcast audio from 4 countries distributed over 1647 one-minute long excerpts. It is designed to encompass several essential features for any music detection dataset and is the first one to include annotations about the loudness of music in relation to other simultaneous non-music sounds. OpenBMAT has been cross-annotated by 3 annotators obtaining high inter-annotator agreement percentages, which allows us to validate the annotation methodology and ensure the annotations reliability. In this work, we first review the current publicly available music detection datasets and state OpenBMAT's contributions. After that, we detail its building process: the selection of the audio and the annotation methodology. Then, we analyze the produced annotations and validate their reliability. We continue with an experiment to highlight the value of these annotations and investigate the most challenging content in OpenBMAT. Finally, we describe the details about the format in which the dataset is presented and the platform where we have made it available. We believe OpenBMAT will contribute to major advancements of the research on music detection in real-life scenarios.

Keywords: dataset; music detection; music relative loudness; automatic classification; inter-annotator agreement; TV broadcast audio

1. Introduction

Music detection refers to the task of finding music segments in an audio file.[1] Thus, the minimum requirement for a dataset to be suitable for this task is to include annotations about the presence of music. However, we find the following two features to be essential to any music detection dataset that aims to provide a certain level of generalization: first, music should appear both isolated and mixed with other type of non-music sounds, because, otherwise, the dataset may not be representative of many real-life scenarios such as broadcast audio; and second, a significant number of the audio files included in the dataset should be multi-class, i.e., contain class changes so that it allows the evaluation of an algorithm's precision in detecting them.

The two main applications of music detection algorithms are (1) the automatic indexing and retrieving of auditory information based on its audio content, and (2) the monitoring of music for copyright management (Zhu et al., 2006; Seyerlehner et al., 2007; Izumitani et al., 2008; Giannakopoulos et al., 2008). Additionally, the detection of music can be applied as an intermediate step to improve the performance of algorithms designed for other purposes (Gfeller et al., 2017). In the current copyright management business model, broadcasters are taxed based on the percentage of music they broadcast. It is relevant to know whether this music is used in the foreground or the background as it is considered differently for the distribution of copyright royalties by some collective management organizations.[2] In this scenario, the music detection task falls short as we need to estimate the loudness of music in relation to other simultaneous non-music sounds, i.e., its relative loudness. We define *relative music loudness estimation* as the task of finding music segments in and audio file and classifying them into foreground or background music. In this paper, we use the concept of loudness as it was defined by Moore (2012): "that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud" (p. 133).

Currently, there is no dataset with annotations about the relative loudness of music, and the only publicly available dataset including the aforementioned two

* MTG, Universitat Pompeu Fabra, Roc Boronat, Barcelona, ES

† BMAT Licensing S.L., Bruniquer, Barcelona, ES

‡ Joint Research Centre, European Commission, ES

Corresponding author: Blai Meléndez-Catalán (blaimelcat@gmail.com)

features is the dataset published by Seyerlehner et al. (2007).[3] Unfortunately, despite containing both isolated music and music mixed with other type of sounds, its annotations do not reflect this information and specify only the presence of music. In the case of the dataset published by Scheirer and Slaney (1997),[4] which is the first open dataset that included annotations about the presence of music, the type of sounds that appear mixed with music are restricted to speech, and this is reflected in the chosen taxonomy: *Music*, *speech*, *simultaneous music and speech* and *other*. Other publicly available datasets that include music presence annotations are MUSAN[5] (Snyder et al., 2015) and GTZAN[6] datasets, but none of them include music mixed with other type of sounds and both of them consist of single-class instances, i.e., audio files annotated as a single segment of a single class. For more information about these datasets see Section 2.

In this paper, we present OpenBMAT, a dataset containing 27.4 hours of audio sampled from 8 different TV program types that have been broadcast in the most popular TV channels of 4 different countries: *France*, *Germany*, *Spain* and the *United Kingdom*. It consists of 1647 one-minute multi-class audio files that include music and non-music sounds both mixed and isolated. OpenBMAT has been cross-annotated by 3 annotators using a taxonomy that mixes the presence of music with its loudness with respect to other type of simultaneous non-music sounds. This taxonomy contains 6 classes (defined in Section 3.2.1): *Music*, *Foreground Music*, *Similar*, *Background Music*, *Low Background Music* and *No Music*. In **Table 1**, we show that, despite not being the longest dataset, OpenBMAT is the only one that brings together all the appropriate characteristics for the task of music detection and also for the estimation of the music's relative loudness.

In Section 2, we describe the currently publicly available datasets that are suitable for the task of music detection. Then, in Section 3, we provide the details about the sampling process and the annotation methodology followed to create OpenBMAT. Section 4 is devoted to the analysis of the dataset's content as well as the validation of the annotations reliability. In Section 5, we perform a simple experiment where we evaluate a state-of-the-art algorithm using OpenBMAT to analyze the benefits of including agreement information in the annotations and the challenges that OpenBMAT poses for current methods. In Section 6, we specify the exact content and format of the annotations and the platform where it is available. Finally, Section 7 provides our main conclusions.

2. Related Work

Scheirer and Slaney (1997) published the first open dataset that included annotations about the presence of music. The taxonomy, however, was designed for the task of discriminating music and speech and is composed of four classes: *Music*, *Speech*, *Simultaneous Music and Speech* and *Other*. The dataset contains 245 manually annotated 15-seconds 16-bit monophonic wav audio files at a sampling rate of 22050 Hz digitally sampled from an FM tuner from the San Francisco Bay area. All audio files are single-class and they are separated in a training part of 184 files and a testing part of 61 files. The training part is divided in 60 files of isolated music, 60 files of simultaneous music and speech, 60 files of isolated speech and four files with other type of sounds. The testing part is divided in 20 files of isolated music with vocals, 21 files of isolated music without vocals and 20 files of isolated speech.

Seyerlehner et al. (2007) presented the first and only open dataset specific to music detection until now, i.e., with a taxonomy including only the classes *Music* and *No Music*. It has a total duration of nine hours unevenly distributed over 13 manually annotated mp3 audio files of duration between 6 and 90 minutes. The audio was extracted from TV programs of the Austrian National Broadcasting Corporation (ORF). The music content of each file ranges from 1.5% to 80%.

Tzanetakis[6] generated another dataset for the task of music and speech discrimination named GTZAN music/speech collection. It consists of 128 manually annotated 30-seconds 16-bit monophonic wav audio files at a sampling rate of 22050 Hz. These audio files are divided equally between the classes *Music* and *Speech*. Some of the present music genres are classical music, folk music, jazz, pop, rock or electronic music. Regarding the speech part, most of the content is in English, but other languages such as German, Chinese, Greek or Serbian appear too.

MUSAN (Snyder et al., 2015) is a dataset that contains: 60 hours of speech coming from Librivox[7] and archived US government files, which are unevenly distributed between 12 different languages; 42 hours of varied music styles such as Baroque, Classical, Romantic, Country, Hip-Hop, Jazz, etc.; and six hours of technical and non-technical noises including, for instance, sounds of the nature and the city. All files are single-class and belong to one of the following classes: *Music*, *Speech* and *Noise*. The dataset includes annotations of the musical genre in the case of

Table 1: Comparison between publicly available datasets.

Name/Author	Mixed music	Classes per inst.	Loudness	# instances	Duration (h)
Scheirer	Yes*, annotated	Single-class	No	245	1
Seyerlehner	Yes, not annotated	Multi-class	No	13	9
GTZAN	No	Single-class	No	128	1.1
MUSAN	No	Single-class	No	2016	108.9
OpenBMAT	Yes, annotated	Multi-class	Yes	1647	27.4

* Only with speech.

the music files and annotations of the speaker language and gender in the case of the speech files.

3. Building OpenBMAT

In this section, we follow the dataset description guide proposed by Peeters and Fort (2012) to provide the details of the OpenBMAT dataset building process. First, we describe the nature of its content and the sampling process that we have used to obtain the data; then, we detail the annotation methodology; and finally, we explain how the dataset is documented and where and how it is stored.

3.1. Raw corpus

OpenBMAT contains 27.4 hours of audio divided in 1647 one-minute audio files. Each of these audio files comes from a different recording that we have sampled from BMAT's[8] private database, which temporarily stores recordings from over 2000 TV channels that this company monitors as part of its business. We consider that having many short audio files allows the dataset to include a greater variety of contexts. Nevertheless, these audio files are long enough to be multi-class.

Some of the recordings in the database include tags about their content. Using these tags, we have forced the sampled audio files to cover a set of varied program types to ensure that the dataset is representative of several different broadcast contexts. The selected program types are: *Children*, *Documentary*, *Entertainment*, *Music*, *News*, *Series and Films*, *Sports* and *Talk*. Unfortunately, only the recordings from certain countries include these tags. This has constrained the dataset to audio broadcast by TV channels from *France*, *Germany*, *Spain* and the *United Kingdom*. We set a limit of 60 audio files for each country and program type, but for several combinations there were not enough audio files to reach that value. **Figure 1** shows the distribution of audio files by program type for each country. All the audio files in the dataset are 16-bit monophonic WAV files at a sampling rate of 22050 Hz and have been extracted from audio broadcast during 2017. Their loudness ranges roughly from –51 to –7 Loudness Units relative to Full Scale (LUFS). LUFS is a logarithmic measure of the perceived loudness in an audio file used in media broadcasting that was introduced in EBU R128.[9]

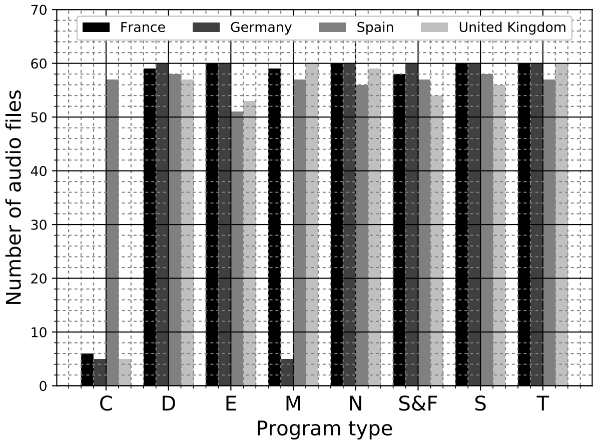


Figure 1: Distribution of audio files by program type and country. The program types are: children (C), documentary (D), entertainment (E), music (M), news (N), series & films (S&F), sports (S) and talk (T).

3.2. Annotation methodology

The annotation of the presence of music in an audio file can already lead to a certain level of disagreement between different annotators either because the music has such a low volume that it is hard to distinguish among other type of sounds or because the content of the audio file lies ambiguously between what can and cannot be considered music. Adding the annotation of the loudness of this music with respect to other simultaneous non-music sounds could drastically increase this disagreement. To be able to produce reliable annotations in this situation, it is important to create a well-defined taxonomy and a set of precise annotation steps.

3.2.1. Taxonomy description

We define a taxonomy formed by 6 classes. Each one applies to a different combination, in terms of loudness, of content that is considered music and content that is not. Classifying sounds as music and non-music is subjective and it is one of the tasks of the annotators. The definition of the classes is as follows:

- **Music**: isolated music.
- **Foreground Music**: mainly music with low-volume non-music in the background.
- **Similar**: music and non-music mixed at similar volumes.
- **Background Music**: mainly non-music with music in the background.
- **Low Background Music**: mainly non-music with music in the background at such a low volume that it is hard to hear.
- **No Music**: isolated non-music.

From our experience, we know that music detection algorithms struggle when trying to perceive music that has a very low volume. We have added the *Low Background Music* class to make the access to this kind of content easier to the users. The whole set of classes can be divided into two groups: *isolated* classes and *mixed* classes. *Music* and *No Music* are isolated classes because an audio segment of these classes can contain either music or non-music sounds but not a combination of both. The rest of the classes are mixed classes because segments annotated as such must contain both music and non-music sounds.

This taxonomy has two important characteristics: (1) all classes are defined to be mutually exclusive, i.e., there cannot be overlap between them; and (2) all audio can be annotated, i.e., all sound excerpts fall into one of the described classes.

3.2.2. Annotation process

The annotation mechanism is very simple: it consists in creating non-overlapping segments over the waveform of the audio files and assigning a class to each them. These segments must cover the entire duration of the audio file.

However, we are actually asking the annotators to describe the relative loudness of music, which has a continuous nature and fast variations, using a discreet set of classes. This can lead annotators to annotate a significantly different number of segments for a given time interval depending on the precision they are using to annotate. To limit the number of class changes that the annotators can annotate, we have established a minimum segment length.

This limitation can bias the annotation to a particular use case: while researchers may want to have very precise annotations to train algorithms that detect the slightest class change, in the industry paradigm it might not be optimal to have an algorithm that produces a large number of short segments for each audio file analyzed. For instance, when dealing with a large number of audio files daily, producing a high number of segments for each of them can lead to problems with the database. We have set this limit to 1 second in an attempt to consider both points of view. To ensure that the annotators follow this rule, we have defined a set of annotation steps:

1. Annotate all *No Music* segments that are longer than one second. From now on:
   (a) we consider that the rest of the audio file contains music either isolated or mixed with non-music sounds.
   (b) we refer to any non-annotated non-music sound as a *group*.
2. Merge groups iteratively if they fulfill two conditions:
   (a) they are separated by less than one second.
   (b) they have similar loudness in comparison with simultaneous music.
3. Annotate resulting groups that are longer than one second as one of the mixed classes.
4. If a group is shorter than one second
   (a) and it is separated by less than one second from another group, merge them. Repeat this operation until there are no more groups to merge. Once two groups merge they take the class that is majority.
   (b) and it is surrounded by *No Music*, annotate it as the appropriate class.
   (c) and it is surrounded by still not annotated audio, leave the group with no annotation.
5. Annotate any part of the audio that is not yet annotated at this point as *Music*.

3.2.3. Annotation reliability

OpenBMAT has been manually cross-annotated by three different annotators: two males and one female of ages comprised between 20 and 40 years and experience working with sound and/or music. Cross-annotating allows us to assess the reliability of the annotations produced (see Section 4.1). All three annotators have been trained to use the annotation tool described in Section 3.2.4 and have learned the annotation steps before starting the annotation of the dataset. Throughout the whole process they have been allowed to ask questions and we have regularly provided feedback. In average, they spent approximately 130 hours annotating the dataset. This means that the annotation rate was around 4.75 hours of annotation per hour of audio.

3.2.4. Annotation tool

The tool that we have used for the annotation of the dataset is BAT, an open-source, web-based tool for the manual annotation of events in audio files. As mentioned in Section 3.2.1, the selected taxonomy implies that classes cannot overlap and allows all audio to be annotated as one of its classes. BAT provides two functionalities that make annotation faster and easier in this situation: (1) it can be configured to avoid any overlap between regions; and (2) it includes shortcuts to expand a region to the limits of adjacent regions or the limits of the audio file being annotated. Additionally, the tool is exclusively oriented to the annotation of audio events, which allows for a simple and clear user interface, and it is connected to the database where all annotations are stored automatically. The audio files are annotated in chunks of 30 seconds and without zoom to maintain a fixed level of time precision during the annotation. We show a screenshot of this tool in **Figure 2**.

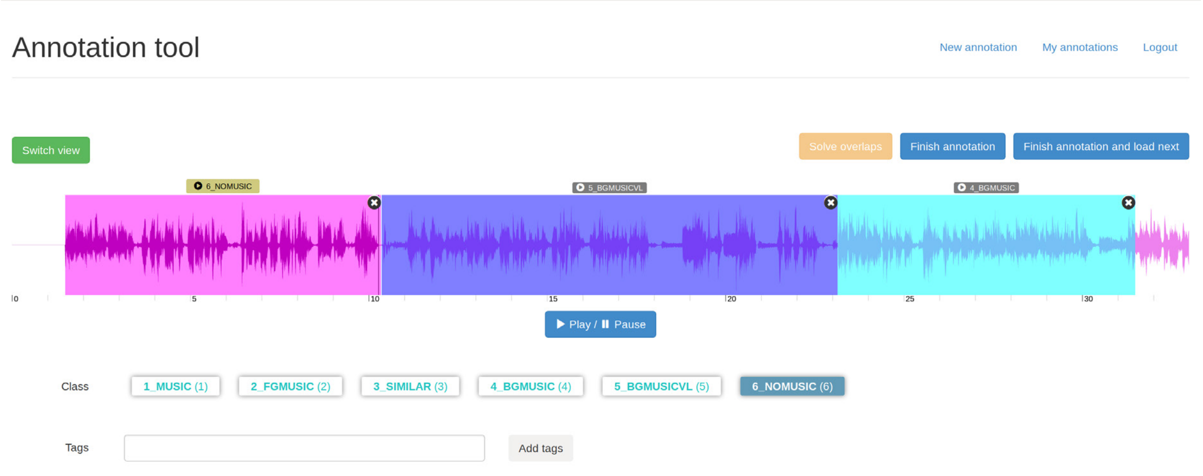


Figure 2: Screenshot of BAT, the annotation tool used for the annotation of OpenBMAT.

4. Analyzing OpenBMAT

In this section, we first validate the annotation methodology through the calculation and analysis of the agreement between annotators; and then, we provide statistics about the content of the annotations.

4.1. Annotation methodology validation

As explained in Section 3.2.3, we cross-annotated the dataset to allow for the assessment of the reliability of the annotations produced. We consider that obtaining reliable annotations validates the definition of the taxonomy and the annotation process, and ensures the usability of the dataset.

The information that we use for this assessment is the percentage of inter-annotator agreement between the three annotators in the annotated classes. For the rest of this paper, we will just use the term *agreement* when referring to inter-annotator agreement. Note that we have not removed any audio file from the dataset based on the agreement information.

We define two different levels of agreement: full agreement, which happens when all three annotators have annotated the same class; and partial agreement, which happens when at least two annotators have annotated the same class. We compute the percentage of full agreement ($\%FA_{af}$) and partial agreement ($\%PA_{af}$) in an audio file as the time during which the agreement level is reached (t_{FA} and t_{PA}, respectively) divided by the duration of the audio file (T_{af}) as shown in Eqn. 1 and Eqn. 2. To obtain the percentage of full agreement (%*FA*) and partial agreement (%*PA*) for the whole dataset, we compute the mean for all *N* audio files as shown in Eqn. 3 and Eqn. 4.

$$\%FA_{af} = \frac{t_{FA}}{T_{af}} \tag{1}$$

$$\%PA_{af} = \frac{t_{PA}}{T_{af}} \tag{2}$$

$$\%FA = \frac{1}{N}\sum_{n=1}^{N}\%FA_{af}(n) \tag{3}$$

$$\%PA = \frac{1}{N}\sum_{n=1}^{N}\%PA_{af}(n) \tag{4}$$

These values can be computed considering all the classes in the taxonomy, but also for the two mappings of these classes that we show in **Figure 3**. These mappings adapt the original taxonomy to the tasks of music detection (MD) and relative music loudness estimation (RMLE). The MD mapping unifies all classes under the *Music* class except for the *No Music* class, which is left unchanged. In the *R*MLE mapping, the *No Music* class also remains unchanged, the *Music* and *Foreground Music* classes are merged into the *Foreground Music* class and the remaining classes are mapped to the *Background Music* class.

Table 2 shows the %FA and %PA when considering all classes as well as when applying both mappings. It also presents these agreement percentage values for each pair of annotators, i.e., the percentage of pair-wise agreement (%PW). We observe that when considering all classes, there is already a %PA of 96.75%. This percentage increases to 99.79% with the RMLE mapping. We also observe that the %FA when considering all classes is 68.18%. **Figure 4** reveals that most of this agreement comes from isolated classes – especially the *No Music* class – as in all mixed classes there is a higher percentage of partial agreement than full agreement. The %FA increases to 89.1% for the RMLE mapping and to 94.78% for the MD mapping. **Figure 5** provides insight on the distribution of full agreement among audio files. It shows the percentage of audio files with a $\%FA_{af}$ over a certain value when using the RMLE mapping. We observe that over 35% of the audio files have a $\%FA_{af}$ higher than 99% and that almost 90% of the audio files have a $\%FA_{af}$ higher 70%.

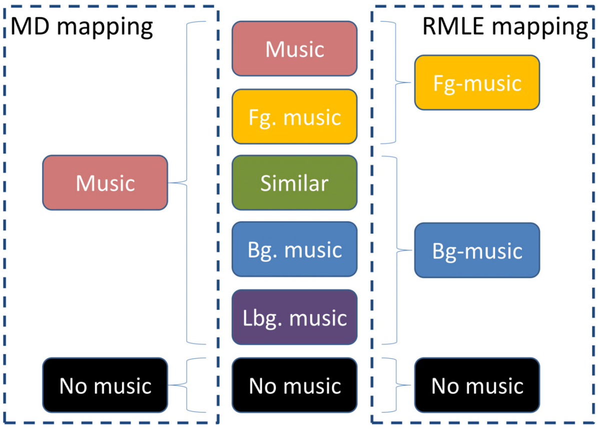


Figure 3: (Left) MD mapping: mapping to compute the agreement for the music detection task. (Right) RMLE mapping: mapping that includes information about the relative loudness of music.

Table 2: Percentages of full, partial and pair-wise (PW) agreement (Agr) for the whole dataset. These values have been computed for the complete taxonomy and both mappings.

Agreement level	No mapping Agr (%)	MD mapping Agr (%)	RMLE mapping Agr (%)
%FA	68.18	94.78	89.1
%PA	96.75	100	99.79
%PW (annotators 1 & 2)	77.46	96.22	91.7
%PW (annotators 2 & 3)	76.97	96.78	92.78
%PW (annotators 1 & 3)	78.66	96.55	93.52

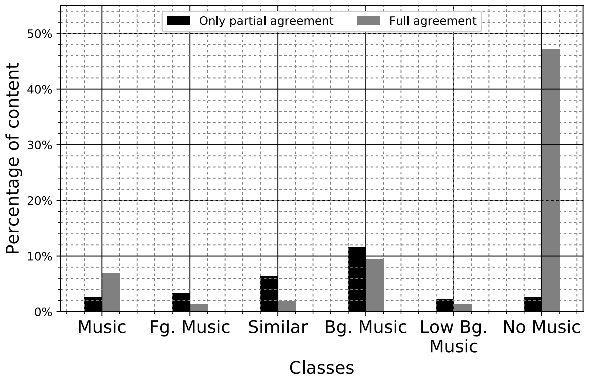


Figure 4: Percentage of the content of OpenBMAT by class and agreement level.

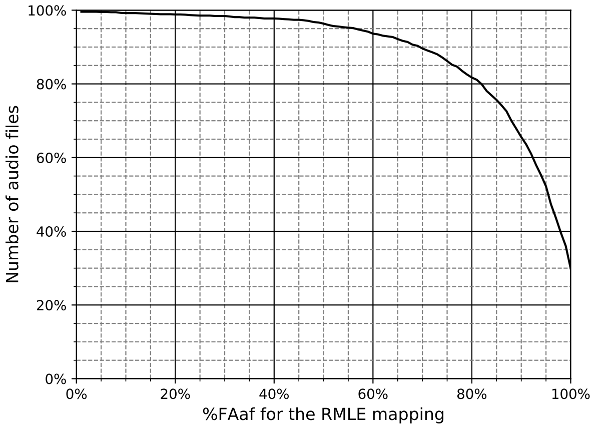
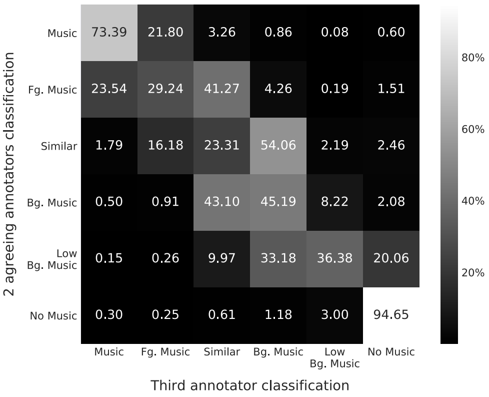


Figure 5: Percentage of audio files accumulated over a certain $\%FA_{af}$ value using the RMLE mapping.

Figure 6 presents the percentage of the content with full (diagonal) or partial agreement for each class divided by the classification of the third annotator. From this figure we extract the 2 main sources of partial agreement when considering all the classes in the taxonomy. The most common source is for one of the annotators to select an adjacent class in terms of loudness. This affects all classes to a different extent except for the *No Music* class. The second source of partial agreement appears when one of the annotators is unable to detect the music due to its low volume and annotates *No Music* instead of *Low Background Music*.

When no annotator coincides in the annotation of time interval, we consider that there is disagreement. Considering all classes in the taxonomy, there are 2 main sources of disagreement: the first one takes place when annotators disagree between *No Music* and any of the other classes due to a different interpretation of what music is. Examples of this can be background tones, musical sound effects such as a church bell or a ringing phone or even experimental music or music of uncommon styles. The second source of disagreement appears due to a different interpretation of what components of the audio belong to the music. This happens, for instance, when the audience claps following the beats of the music or when there are soft noises overlapping with the music that some annotators may find irrelevant. This can lead to strong differences in the annotation and it represents a significant percentage of the parts of the dataset with disagreement.



	Music	Fg. Music	Similar	Bg. Music	Low Bg. Music	No Music
Music	73.39	21.80	3.26	0.86	0.08	0.60
Fg. Music	23.54	29.24	41.27	4.26	0.19	1.51
Similar	1.79	16.18	23.31	54.06	2.19	2.46
Bg. Music	0.50	0.91	43.10	45.19	8.22	2.08
Low Bg. Music	0.15	0.26	9.97	33.18	36.38	20.06
No Music	0.30	0.25	0.61	1.18	3.00	94.65

Figure 6: (Rows) Class annotated by 2 annotators. (Columns) Class annotated by the third annotator. (Values) Percentage of the content with full or partial agreement for each class divided by the classification of the third annotator.

4.2. Content distribution

After the cross-annotation process, we obtain three different annotations of the same content. **Table 3** shows the content distribution as annotated by each annotator for the complete taxonomy and both mappings. We observe that the percentages are similar for all annotators: first, all annotators have considered that around 50% of the dataset is *No Music*, which implies that the other 50% contains music that appears either isolated or mixed with non-music sounds. This means that OpenBMAT is balanced in terms of music and non-music content. Second, the part of the dataset containing music is approximately distributed with a 30%/70% proportion between the RMLE mapping *Foreground Music* and *Background Music* classes, respectively. Also, if we divide the dataset in terms of isolated and mixed classes as explained in Section 3.2.1, the average proportion for all annotators is around 59% and 41%, respectively.

5. Experiment: Testing with OpenBMAT

We have carried out a simple experiment with two goals in mind: (1) to find out the content in OpenBMAT that is most challenging for music detection algorithms; and (2) to demonstrate the potential of having agreement information when assessing algorithm performance.

To achieve these goals, we have evaluated a state-of-the-art music detection algorithm with OpenBMAT. The selected algorithm is the winner of 2018 MIREX[10] competition for the task of music detection (Meléndez-Catalán, 2018). MIREX is an international annual evaluation campaign for MIR algorithms, coupled to the ISMIR[11] conference. A previous version of OpenBMAT was used to evaluate this algorithm in the aforementioned competition. However, both the taxonomy and the annotation method used to create the ground truth were significantly different, that is why both evaluations are not comparable.

We name the selected algorithm MMG. It is based on a convolutional neural network that estimates the proportion of the loudness that corresponds to the music content at each frame. By thresholding the estimated loudness, we can classify the frame either into the MD or the RMLE mapping classes, i.e., into *Music* and *No Music* or into *Foreground Music*, *Background Music* and *No Music*.

5.1. Challenges in OpenBMAT

We have divided the estimation errors in three groups: (1) those related to the ambiguity of what is music and what is not; (2) those related to the heterogeneity of music and non-music sounds; and (3) those related to the music's volume. The first error type includes, mainly, non-music sounds with certain musical features, such as a recognizable pitch or rhythm, that are classified as *Music*. Examples of this type of error are sounds like a church bell, a ringing phone or background noises with one or more prominent frequencies. Regarding the second error type, the algorithm tends to make more errors when analyzing music with a prominent voice part, such as a capella singing or opera, or music with a lot of percussion. The third type of error is the most frequent one and happens due to the incapacity of the algorithm to detect music with a very low loudness in comparison with the simultaneous non-music sounds. This error analysis shows that OpenBMAT includes audio that is challenging for state-of-the-art methods, and thus, will be useful to evaluate future algorithms for this task.

Table 4 presents the evaluation results. To compute the values in this table, we have run the algorithm against the three individual annotations and computed the mean of the statistics. We observe an accuracy lower than 90% for both evaluations, which indicates that there is still room for improvement for future algorithms. Note that the evaluation statistics using the MD mapping are significantly better than using the RMLE mapping. This happens because the division of the MD mapping *Music* class into the RMLE mapping *Foreground Music* and *Background Music* classes introduces new errors.

5.2. Using agreement information

The concept of agreement between annotators is closely related to the existence or not of an actual ground truth. If all annotators coincide in the annotation of an audio segment, chances are that the segment has a clear ground truth according to the given taxonomy. If the content of the segment is too ambiguous for three humans to agree, then it might be that an actual ground truth does not exist, and thus, the segment might not be suitable to train an algorithm or evaluate its performance. It is in the hands of potential users of OpenBMAT to decide if they want to train or test their algorithms using only the content with full agreement or, otherwise, include segments with partial or no agreement. These ambiguous segments might generate a glass-ceiling effect for the algorithm's accuracy when used for testing.

Figure 7 shows the distribution of OpenBMAT's audio files by $\%FA_{af}$ in the horizontal axis and the accuracy that

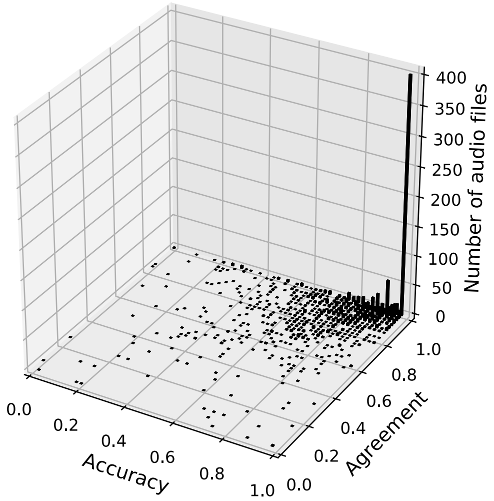


Figure 7: Audio file distribution by full agreement using the RMLE mapping and the accuracy achieved by MMG when evaluated against the annotations of one of the annotators.

Table 3: Columns 2 to 4: percentage of all the audio annotated by each annotator as each of the classes of the RMLE mapping. Columns 5 and 6: percentage of all the audio annotated by each annotator as *Music* or *No Music* (isolated) or as any of the other 4 classes (mixed).

Annotator	Fg. Music (%)	Bg. Music (%)	No Music (%)	Isolated (%)	Mixed (%)
Annotator 1	16.6	34.45	48.94	60.09	39.91
Annotator 2	12.7	37.28	50.02	57.84	42.16
Annotator 3	15	34.66	50.34	59.28	40.72

Table 4: Performance of MMG on the OpenBMAT dataset using the MD and RMLE mappings. We report overall accuracy (Acc), and Precision (P) and Recall (R) for each mapped class. In this table, Music stands both for *Music*, in the case of MD mapping, and *Foreground Music*, in the case of RMLE mapping.

Mapping	Acc.	Music P	Music R	Bg. Music P	Bg. Music R	No Music P	No Music R
MD	88.95	91.99	85.45	–	–	86.29	92.48
RMLE	82.71	77.64	69.96	78.51	76.09	86.8	91.33

MMG achieved on the annotations of one annotator in the vertical axis. Both agreement and accuracy values are computed using the RMLE mapping classes. We observe that by adding the agreement axis, we can distinguish between errors in audio files with a clear ground truth and errors in audio files with an ambiguous ground truth. Potential users will probably find more beneficial to focus in the first type of errors in order to improve their algorithms. Using the intervals in the agreement metadata (see Section 6), the users can know if the errors correspond to a segment with full, partial or no agreement. Therefore, having agreement information can help us better judge the severity of an algorithm's error and have more insight during its evaluation.

6. Metadata and Storage

With metadata we refer to all information related to the audio files in OpenBMAT. There are two sources of metadata per audio file: the annotations and the agreement. The annotations include the three original annotations as well as their MD and RMLE mappings. In total, there are nine annotations per audio file. The agreement metadata contains the $\%FA_{af}$ and the $\%PA_{af}$ of each audio file for the original annotations and both mappings, and the set of intervals with either full or partial agreement. These intervals give potential users the possibility to train or test their algorithms using, for instance, only the subset with full agreement. All this metadata is provided in JSON format. The metadata also includes the annotations in two additional formats: (1) as exported from BAT (one CSV file for each annotator) and (2) as separate TSV files for each audio file, annotator and taxonomy.

Apart from the metadata and the audio, OpenBMAT also contains:

- A predefined split in 10 subsets to allow for K-fold cross-validation. We have randomly assigned each audio file to a subset and the resulting subsets preserve the original proportion of the RMLE mapping classes.
- A python module with utilities such as (1) the generation of the annotations in JSON and TSV formats from BAT annotations, (2) the generation of the full and partial agreement audio subsets and (3) the possibility to load the annotations using the open evaluation library used in the DCASE challenge.[12] The DCASE challenge is an international competition similar to MIREX, but exclusive to scene and event analysis methods.
- A README file including a general description of the dataset and all the details about its structure and contents.

We will provide the dataset through a request form to anyone interested in using it for research purposes. We have created an entry in Zenodo, with the link to the request form, to enable version control giving each version a different DOI.

7. Conclusions

Being aware of the current industry needs, we have designed OpenBMAT: the first music detection dataset to include annotations about the loudness of music in relation to simultaneous non-music sounds. OpenBMAT is an open dataset, and in comparison with the other publicly available music detection datasets, it is the only one that encompasses two essential features: (1) it contains a significant amount of multi-class audio files, and (2) the music appears both isolated and mixed with different types of non-music sounds. All these characteristics make OpenBMAT the most complete open dataset available for the task of music detection and the estimation of its relative loudness.

To be able to assess the annotations reliability, three annotators have cross-annotated OpenBMAT. The analysis of these annotations produces a full agreement of 94.78% when using the MD mapping and 89.1% when using the RMLE mapping. When considering all classes, the 3 annotators completely disagree only in 3.25% of the content. We have included agreement information for each audio file in OpenBMAT, specifically, the $\%FA_{af}$ and the $\%PA_{af}$ of each audio file and the segments with full and partial agreement. This can be useful for training as well as for testing of algorithms, as we have proved through the evaluation of a state-of-the-art algorithm. The error analysis of this evaluation has also revealed that OpenBMAT contains audio that is challenging to current state-of-the-art music detection algorithms.

The metadata in OpenBMAT can be improved by adding more annotators to increase the reliability of the annotations even more. It can also be expanded through the creation of annotations using taxonomies related to other tasks or by including more information of each audio file such as its quality, its sound level, etc. The current and future annotations should be continuously reviewed to correct possible mistakes. That is why it is crucial to put OpenBMAT under version control. We have uploaded it to Zenodo privately, and we will provide it upon request for research purposes.

Notes

[1] https://www.music-ir.org/mirex/wiki/2018:Music_and/or_Speech_Detection.

[2] https://createurs-editeurs.sacem.fr/brochures-documents/regles-de-repartition-2017.

[3] www.seyerlehner.info/download/music_detection_dataset_dafx_07.zip.

[4] https://labrosa.ee.columbia.edu/sounds/musp/scheislan.html.

[5] http://www.openslr.org/17/.

[6] opihi.cs.uvic.ca/sound/music_speech.tar.gz.

[7] https://librivox.org.

[8] https://www.bmat.com/.

[9] https://tech.ebu.ch/docs/r/r128.pdf.

[10] https://www.music-ir.org/mirex/wiki/MIREX_HOME.

[11] https://www.ismir.net/.

[12] https://github.com/DCASE-REPO/dcase_util.

Acknowledgements

We thank the annotators for the effort they invested in the annotation of OpenBMAT. We also thank Alex Ciurana from BMAT with whom we have discussed issues essential to the construction of this dataset. We thank Brisa Burriel for her advice on the best way to make OpenBMAT available to the research community and Sonia Espi for providing the means to do so. Finally, we thank the Catalan Industrial Doctorates Plan for funding this research.

Competing Interests

Emilia Gomez is a co-Editor-in-Chief of Transactions of the International Society for Music Information Retrieval. She was removed completely from all editorial processing. There are no other competing interests to declare.

References

Gfeller, B., Guo, R., Kilgour, K., Kumar, S., Lyon, J., Odell, J., Ritter, M., Roblek, D., Sharifi, M., Velimirović, M., et al. (2017). Now playing: Continuous low-power music recognition. In *NIPS 2017 Workshop: Machine Learning on the Phone (NIPS).*

Giannakopoulos, T., Pikrakis, A., & **Theodoridis, S.** (2008). Music tracking in audio streams from movies. In *Proceedings of the IEEE 10th Workshop on Multimedia Signal Processing (MMSP)*, pages 950–955. DOI: https://doi.org/10.1109/MMSP.2008.4665211

Izumitani, T., Mukai, R., & **Kashino, K.** (2008). A background music detection method based on robust feature extraction. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13–16.

Meléndez-Catalán, B. (2018). Music and/or Speech Detection MIREX 2018 Submission. Music Information Retrieval Evaluation eX-change.

Moore, B. C. (2012). *An introduction to the psychology of hearing*. Brill.

Peeters, G., & **Fort, K.** (2012). Towards a (Better) Definition of the Description of Annotated Mir Corpora. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 25–30.

Scheirer, E., & **Slaney, M.** (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1331–1334. DOI: https://doi.org/10.1109/ICASSP.1997.596192

Seyerlehner, K., Pohle, T., Schedl, M., & **Widmer, G.** (2007). Automatic music detection in television productions. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07).*

Snyder, D., Chen, G., & **Povey, D.** (2015). MUSAN: A Music, Speech, and Noise Corpus. arXiv:1510.08484v1.

Zhu, Y., Sun, Q., & **Rahardja, S.** (2006). Detecting musical sounds in broadcast audio based on pitch tuning analysis. In *IEEE International Conference on Multimedia and Expo*, pages 13–16. DOI: https://doi.org/10.1109/ICME.2006.262502

How to cite this article: Meléndez-Catalán, B., Molina, E., & Gómez, E. (2019). Open Broadcast Media Audio from TV: A Dataset of TV Broadcast Audio with Relative Music Loudness Annotations. *Transactions of the International Society for Music Information Retrieval,* 2(1), pp. 43–51. DOI: https://doi.org/10.5334/tismir.29

Submitted: 15 January 2019 **Accepted:** 24 June 2019 **Published:** 12 August 2019

]u[*Transactions of the International Society for Music Information Retrieval* is a peer-reviewed open access journal published by Ubiquity Press. OPEN ACCESS