OpenMIC-2018: AN OPEN DATASET FOR MULTIPLE INSTRUMENT RECOGNITION

Eric J. Humphrey
Spotify
ejhumprey@spotify.com

Simon Durand
Spotify
durand@spotify.com

Brian McFee
New York University
brian.mcfee@nyu.edu

ABSTRACT

Identification of instruments in polyphonic recordings is a challenging, but fundamental problem in music information retrieval. While there has been significant progress in developing predictive models for this and related classification tasks, we as a community lack a common data-set which is large, freely available, diverse, and representative of naturally occurring recordings. This limits our ability to measure the efficacy of computational models.

This article describes the construction of a new, open data-set for multi-instrument recognition. The dataset contains 20,000 examples of Creative Commons-licensed music available on the Free Music Archive. Each example is a 10-second excerpt which has been partially labeled for the presence or absence of 20 instrument classes by annotators on a crowd-sourcing platform. We describe in detail how the instrument taxonomy was constructed, how the data-set was sampled and annotated, and compare its characteristics to similar, previous data-sets. Finally, we present experimental results and baseline model performance to motivate future work.

1. INTRODUCTION

Music information retrieval (MIR) applications often depend on statistical models and machine learning algorithms to relate audio content to semantically meaningful representations. The development and evaluation of these methods, in turn, depends on access to data, typically audio recordings which have been annotated for a particular task such as chord recognition or tag prediction. Ideally, the data we use to develop and evaluate models should be large, diverse, and open access, so that we as researchers and engineers can diagnose failure modes and propose improvements. However, because the vast majority of music is subject to copyright, this has historically been difficult to achieve. This has resulted in a proliferation of *de facto* standard data-sets which are small, biased, and not freely available, which ultimately impedes scientific progress.

To address this problem, McFee et al. [15] proposed an iterative evaluation framework for developing open access data-sets for MIR, with a specific focus on instrument recognition. While this proposal was apparently met with enthusiasm from the community, little progress has been made in the intervening time toward enacting the proposal. We hypothesize that this was primarily due to two factors: a lack of a conveniently accessible audio data, and the expense of creating the initial development set. Recently, two complementary data-sets have been published, which we combine here to resolve both of these issues: the Free Music Archive data-set [8], and AudioSet [11]. The result is a diverse, open access collection of 20,000 audio clips annotated for the presence of 20 distinct instrument categories, which we denote as *OpenMIC-2018*.

1.1 Our contributions

Our primary technical contribution is a new, open dataset for training and evaluating instrument recognition algorithms. This article describes in detail how the dataset was constructed by using a combination of model transfer from previous datasets and crowd-sourced annotation. Our goals in documenting the data construction process are two-fold. First, it provides transparency around the various decisions and compromises made in this specific dataset. Second, we describe technical issues and general solutions which may be of interest to future developers of music datasets.

1.2 Related work

Instrument recognition, either monophonic or polyphonic, is a long-standing problem in MIR, and many datasets for evaluating methods have been developed over the years. Table 1 lists some of the commonly used datasets, along with various descriptive attributes. Of specific interest are the size of the collections, the number of instrument classes, the duration of each example, the diversity of the collection (*e.g.*, genre or style), whether the examples are polyphonic, the number of instrument labels per example, and whether the data is open access.

Broadly speaking, existing datasets can be broken into two categories, according to whether samples contain notes played by isolated instruments (RWC [12], Good-sounds [1], or NSynth [10]), or recordings of instrument ensembles. Datasets of isolated instrument recordings are often easier to produce and annotate at large scale because long recordings spanning multiple notes can be segmented

Table 1. A qualitative comparison of different existing datasets for instrument identification.

Collection	# Examples	# Instruments	Duration	Diverse	Polyphonic	Multi-label	Open
RWC [12]	3,544	50	scale				
Good-sounds [1]	6,548	12	note				✓
NSynth [10]	305,979	1,006	note				✓
MedleyDB [3]	122	80	song	✓	✓	✓	
MusicNet [19]	330	11	song		✓	✓	✓
IRMAS [4]	6,705	11	3s	✓	✓		
OpenMIC-2018	20,000	20	10s	✓	✓	✓	✓

to generate examples with a shared label. However, the acoustic properties of ensemble recordings differ significantly from those of isolated recordings, so models developed on single-instrument data often do not generalize to the polyphonic case. Conversely, ensemble recordings are typically difficult to precisely annotate, which results in either high-quality collections with a small number of distinct tracks (MedleyDB [3] or MusicNet [19]), or in collections with more tracks but with only partial annotations (such as IRMAS [4] with predominant instrument tags for short excerpts). An ideal dataset would be large, diverse, strongly annotated (including both positive and negative examples), and freely available, so that at each instant in any recording, full information about all active instruments is available. While existing datasets succeed on some of these criteria, none achieves all simultaneously.

1.3 The Free Music Archive

The Free Music Archive [1] (FMA) is a web-based repository of freely available music recordings. Recently, a snapshot of FMA has been released to the research community to facilitate content-based music analysis evaluation [8]. The FMA snapshot includes 106,574 tracks by some 16,341 artists, along with pre-computed features. Each track is annotated with both coarse (16 categories) and fine (161 categories) genre tags. Tracks are provided under a small variety of licenses, with the vast majority being Creative Commons [7]. This allows practitioners to archive and redistribute data (with some minor restrictions), which is fundamental to the practice of open and reproducible scientific research.

While previous authors have noted the particular genre biases present on FMA [8], it nonetheless provides a large pool of realistic musical content which could be used in research applications. Despite the specific quirks of the FMA collection, using it as a basis for large-scale MIR evaluation has several benefits. In addition to the obvious benefits of being open access, it also facilitates data revision and inclusion of new contributions from the community at large. This in turn makes it easier for corrections to be integrated, and the collection to grow over time and not become stale.

2. CONSTRUCTING OpenMIC-2018

In developing OpenMIC-2018, we took inspiration from ImageNet [9]. ImageNet was constructed by selecting and annotating natural images to represent categories (synonym sets, or *synsets*) drawn from the WordNet ontology [16], with a goal of having at least 500 positive examples for each category. Candidate images were selected by querying image search engines for each category term, and then labels were verified by crowd-sourced annotation. The label correction and verification step was critical at the time, due to the poor accuracy of image search engines when the dataset was constructed in 2009.

We follow a similar strategy here, with a few notable modifications. Rather than querying the Internet for candidate samples, we restrict attention to freely available content hosted on the Free Music Archive, and specifically those with explicit Creative Commons licensing. Additionally, instead of the WordNet ontology, we use the recently published AudioSet concept ontology [11], which itself derives from WordNet, but is adapted to acoustically meaningful concepts. Using existing AudioSet data, we construct a multi-instrument estimator and use this model to rank the unlabeled FMA data and provide candidates for annotation. The remainder of this section describes the entire process in detail, which is visualized in Figure 1.

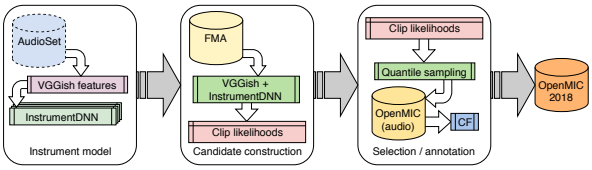

Figure 1. A multi-label instrument detector (*InstrumentDNN*, section 2.2) is trained on AudioSet data. The model is used to score each 10s clip in FMA by likelihood of each instrument (section 2.3). Clips are sorted into quantiles for each instrument, then sub-sampled and annotated by CrowdFlower workers (section 3).

2.1 AudioSet

AudioSet is a recently released concept ontology and human-annotated dataset derived from YouTube videos, with the goal of providing a testbed for identifying acoustic events [11]. The ontology consists of 632 classes, represented as a lattice-like graph, rather than hierarchical tree structure, *i.e.* one low-level class may have two distinct parents. The annotated dataset consists of at least 100 positive examples of 485 classes, distributed (non-uniformly) across nearly 1.8M video clips of 10 seconds (or less)

[1] http://freemusicarchive.org/

drawn from YouTube. Similar in spirit to the work presented here, AudioSet is motivated by a lack of large-scale annotated audio data for scientific research purposes.

While the AudioSet ontology includes musical instruments, the audio data does not match our requirements for an open music instrument sample. The collection is derived from YouTube videos, for which there are no guarantees on the legality of licensing, sharing, and archiving the content. Though abstract features are made available via a publicly available acoustic model, an inability to make the source content directly accessible has limited the value of other large collections, such as the Million Song Dataset [2]. Furthermore, the content is often quite different from musical performances, an important characteristic at the root of what makes this task both challenging and interesting: many of the positively labeled examples are solo performances, which makes it difficult to model and evaluate on realistic, highly correlated ensemble performances.

That said, AudioSet serves two important functions in this project. It is impractical to annotate the entire FMA collection of more than 100K recordings outright; however, it is *also* extremely unlikely that one could draw a random subsample with sufficient representation across a number of instruments. The occurrence of musical instruments is heavily biased by popularity, such as voice, guitar, or piano, and this is especially true in the Free Music Archive. Here, we leverage AudioSet to build a multi-instrument estimator that allows us to sub-sample and more efficiently use annotation resources. For better or worse, we also leverage the previous work in ontology construction, while circumventing the important, but difficult, challenge of selecting which instruments to consider: here, we are limited to only those with enough signal in AudioSet on which to build a baseline model.

We manually identify the classes that correspond to musical instruments, resulting in a set of more than 70 relevant classes. For the sake of coverage, they are merged into "instruments", *e.g.* "Acoustic Guitar", "Electric Guitar", and "Tapping (guitar technique)" become *guitar*, while "Cello" and "Violin" remain distinct. Note that this class resolution is intentionally approximate, as the long-term goals of this project include iteratively refining these concepts as acoustic models improve. We then filter the 1.8M clips in AudioSet to those containing these classes. Unsurprisingly, the distribution skews toward instruments common in Western popular music, such as guitar, violin, or drums, and we cut this list at 1500 examples. Additionally, we randomly draw 8000 non-musical examples as negative instances for building the instrument model described in section 2.2. In summary, the resulting instrument subset consists of 206K clips, totalling roughly 2M seconds (570 hours) of annotated content for 23 instruments. [2]

2.2 Multi-instrument modeling

AudioSet offers no licensing guarantees on the source content, and there is no approved mechanism for directly accessing the audio data. To make the dataset more generally useful, the developers of AudioSet have released both a pre-trained feature embedding model [13] based on the VGG architecture for object detection in images [18], [3] and its outputs over the original AudioSet audio signals. [4] This model, referred to as "VGGish", produces a 128-dimensional feature vector every 0.96 seconds with an equal window size, such that adjacent features capture non-overlapping context. VGGish features are ZCA-whitened and each coefficient is quantized to 8-bits to reduce the footprint of the dataset.

Using the sub-sampling process described above, we filtered the AudioSet features down to those clips relevant for the instrument ontology considered here. The data are conditionally partitioned by YouTube ID into training, validation, and test splits with a $3:1:1$ ratio. We randomly generate over 200 unique, fully connected deep network architectures and hyper-parameter configurations, spanning depth (1–8 layers), width (128 to 2048 units, by powers of 2), the application of dropout and batch normalization, different optimization algorithms (stochastic gradient descent, RMSProp, and Adam [14]), as well as various parameters for each operation. All models are trained for 50 epochs of the training data, and the parameter checkpoint with the highest macro-F_1 (class-averaged) score over the validation data is taken as the best model.

Overall, we find that roughly 15% of the models behave with statistical equivalence, achieving a mean macro-F_1 score of 0.514 ($\sigma = 0.0095$) and a micro-F_1 (item-averaged) score of 0.656 ($\sigma = 0.0056$) on the test partition. The best configuration is determined to be a 7-layer network, with widths of $[1024, 512, 256, 1024, 256, 1024, 23]$, batch-normalization on the first four layers, and pointwise dropout applied to the inputs of the last five $[0.0, 0.0, 0.25, 0.125, 0.25, 0.25, 0.5]$. The winning model, which we refer to as *InstrumentDNN*, is trained with the Adam optimizer in Keras for 8 epochs, with a learning rate of 0.0001 and a β_1 of 0.99. For reproducibility, the training data and trained model are made publicly available in the source repository.

2.3 FMA clip sampling

The VGGish model is applied to each track in FMA, and the resulting ZCA features are processed by InstrumentDNN to produce time-varying instrument likelihoods. Full tracks are then divided into candidate clips by performing maximum-likelihood aggregation over 10 second windows with a 4 second hop size. To account for framing effects, the maximum likelihood of each instrument is taken over the middle 8 seconds, centered on the frame. This produces over 7M clip candidates.

We ultimately want an approximately balanced sample that has good positive representation of each instrument class. Therefore the candidate set is sub-sampled by the following process. First, we consider the median like-

[2] https://github.com/cosmir/open-mic-data

[3] https://github.com/tensorflow/models/tree/master/research/audioset

[4] https://research.google.com/audioset/download.html

lihood of each class over all candidates, and sort instruments in ascending order, as a proxy for class occurrence in the FMA. Then, proceeding from least to most likely instrument class, the clip candidate set is reordered by descending conditional class likelihood. We randomly selected N instances from the 99th percentile rank of that class, such that no two clips share a source track, *i.e.* sampled clips are recording-independent. All remaining clip candidates that also share a common recording with any sampled are discarded, and the process is repeated for the next instrument. For K instruments, the sampling process yields $N \times K$ clips from distinct tracks. Initially, we set $N = 1000, K = 23$, but manual inspection of the results revealed three classes that either InstrumentDNN cannot reliably detect, are poorly represented in FMA, or both: *harp*, *bagpipes*, and *harmonica*. We removed these classes, leaving $K = 20$ instruments and 20K clips.

3. CROWD-SOURCED ANNOTATION

At this stage, roughly 25M seconds of audio have been sub-sampled to 200K, a 10^5 reduction, while rebalancing for instrument occurrence. Strongly labeling a collection of this size is still cost-prohibitive, and we must be pragmatic with our annotation efforts. In tackling this challenge, one can think of annotation as a sparse, binary matrix completion problem where most of the values in the instrument occurrence matrix will be zero. Therefore annotation effort is best allocated by flattening this matrix into clip-instrument pairs, and prioritizing likely positives.

Framed this way, our most likely positives are identified by the clip selection process: each instrument has 1K potential positive examples that must be validated by human annotators. We would also like to obtain a number of strong negatives as well, and draw 500 instances per class that fall in the bottom 10^{th} likelihood percentile from the space of examples contained in OpenMIC-2018. Instrument-wise percentile thresholds are computed over the full space of clip candidates. In contrast to positive sampling, negative samples are drawn working from most to least likely instruments. This is because the most likely instrument categories will have the fewest potential strong negatives. Additionally, random sampling is constrained to draw no more than three strong negatives per clip, so as to distribute this information across the collection. Finally, to capture potential correlations and confusions, all additional likelihoods in the 99^{th} percentile rank of their respective instrument classes are added to the pool of binary questions for human annotators. This results in 33,250 potential positive and 10,000 potential negative binary estimates for human validation, which makes up roughly 10% of all possible clip-instrument judgements.

Having identified the questions worth asking, audio annotation presents unique design challenges around *how* to best ask these questions of humans. Unlike images, audio clips cannot be scanned in parallel by humans, and must be auditioned sequentially. This encourages annotation designs that ask several binary questions about the same example. Our first attempt to annotate OpenMIC-2018 took

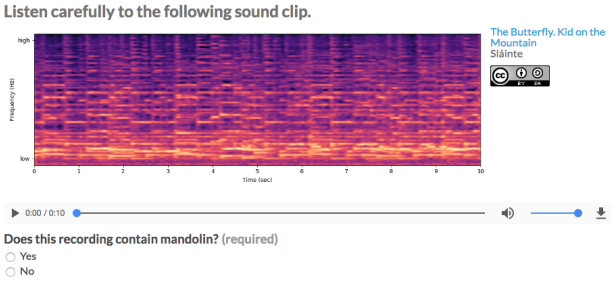

Figure 2. An example annotation task, showing the Mel-spectrogram visualization, playback, response field, and licensing meta-data.

this approach, but we found that annotators struggled with the increased burden of simultaneously judging multiple instrument tags. This resulted in poor agreement, unhappy annotators, and an increased level of effort and skill to complete. Our second attempt used 20 separate annotation tasks, one per instrument, and annotators were asked to determine the presence or absence of a specific instrument across multiple recordings.

Annotation was performed on the CrowdFlower [5] platform (CF). In contrast to Amazon Mechanical Turk, CF provides quality controls on sets of questions, collectively called a "job". A single contributor can provide at most 50 responses (or 10% of the job, whichever is larger), and a question is finalized when annotators reach a set agreement level and number of responses.

Additionally, CF makes it easy to include control questions for which an answer is already known. These are used to "quiz" contributors before they can perform any (paid) work on a job, and remove contributors whose accuracy drops below a threshold, *e.g.* 70%. It is important that control questions use clear, unambiguous examples. While these can be easily identified for popular classes, it is difficult in the rare classes, notably *mandolin* and *clarinet*. For these classes, control questions were generated by ranking clips according to the margin between the target instrument's likelihood and the maximum over other instruments for that clip, which gives preference toward clips where the target instrument was both present and prominent.

Each question is a single judgement of an instrument's presence or absence for a given audio clip. As shown in Figure 2, we use a radio button interface for the judgement, provide audio playback in the browser, and additionally display an approximately aligned Mel-spectrogram to facilitate the task, inspired by previous audio annotation research [5]. Finally, we are legally obligated to display track title, artist, and license information, which may provide coincidental information about a given track.

4. OpenMIC-2018 ANALYSIS

We collected over 230K judgements from more than 2,500 unique contributors across the 20 instrument classes. Figure 3 summarizes the resulting annotation distributions for

[5] `http://crowdflower.com`

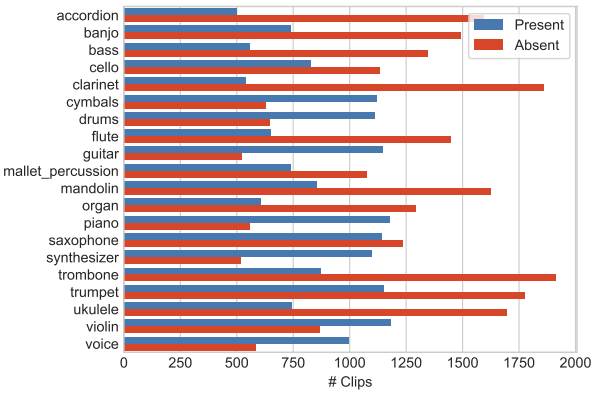

Figure 3. Statistics of crowd-sourced annotation for each instrument in OpenMIC-2018.

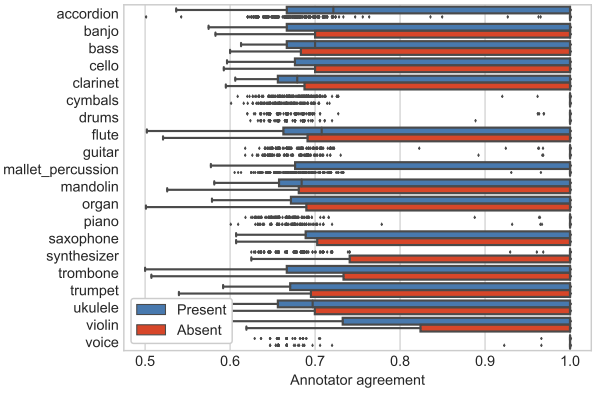

Figure 4. Annotator agreement for each instrument.

each instrument. For each instrument class, the number of confirmed positive and negative clips are plotted separately. Each class has at least 500 confirmed positives, and at least 1500 confirmed positive or negative. Although not every clip is tagged for every instrument, the abundance of strong negative labels facilitates supervised learning and strong evaluation. Figure 4 summarizes the inter-annotator agreements for each instrument's presence or absence. Some instruments produce more agreement for absence than presence (*accordion*, *violin*), while the reverse is true for others (*synthesizer*). Overall, we observed a high amount of agreement across all instruments.

Figure 5 compares InstrumentDNN's predicted likelihoods to the annotations for three instrument classes. InstrumentDNN produces a wide range of likelihood values on *mandolin* (fig. 5, left), indicating that the 99th percentile likelihood is well below the threshold for positive detection. This is likely due to a combination of model calibration errors and poor representation in AudioSet. However, the sampling strategy still produced a large number of validated positive examples. For more common classes, such as *cymbals* (fig. 5, center), there is a clearer distinction between the positive and negative selections. For the most common classes, such as *voice* (fig. 5, right), the vast majority of positive selections are validated by the annotators as positive, and conversely for the negative selections.

To measure the diversity of the annotated subset, fig. 6 compares the distribution of genres over both the sample and the background population of FMA. While both distributions exhibit non-uniform genre distributions, the sample is fairly representative of FMA. The instrument-based sampling does introduce some systematic bias, increasing representation of styles with distinctive instrumentation, such as *classical* or *jazz*. This effect can be observed directly in fig. 7, which shows the number of clips in each genre that are positively labeled for each instrument. For example, the majority of *organ* and *piano* examples are tagged as *classical*, while *synthesizer* is drawn primarily from *electronic* and *experimental*.

4.1 Experiment: baseline modeling

To estimate the expected performance of standard methods on OpenMIC-2018, we conducted a set of baseline experiments. We trained independent binary classifiers for each instrument. We report the accuracy of each of those models on 100 splits of the data, randomly selecting 500 test instances, and splitting the resulting training set in 3 folds for hyper-parameter selection. As input representation, we use the mean and standard deviation of VGGish features over the clip's duration.

We tested several baseline models, and for simplicity report only the best performing one: a random forest (RF) classifier. The hyper-parameter search is done on the number of trees ($\{10, 100, 1000\}$) and on the maximum depth of the tree ($\{2, 4, 8\}$). We also report the bias point of each instrument category, and the performance of InstrumentDNN. This last comparison point gives us a measure of how much information is gained by the crowd-sourced labels. This experiment is done with the scikit-learn [17], and the code to reproduce will be made available.

The results are shown in fig. 8. We see an overall gain in accuracy of more than 10 percent point (pp) compared to both the bias points and InstrumentDNN. The performance difference can partly be explained by the difference in training distributions between RF and InstrumentDNN, and because a strong signal can be learned from the dataset. The RF model performance is also more consistent across instruments with only a 20 pp difference between the worst and best instrument accuracy, compared to a 34 pp difference for InstrumentDNN. The gain compared to InstrumentDNN is therefore larger on the more difficult instruments, such as saxophone, mandolin and ukulele. In that case the crowd-sourced judgments might provide more value and help build a robust system.

5. CONCLUSION

OpenMIC-2018 should prove to be useful for developing and evaluating instrument detection models. We note that the dataset is not "complete" in that not every clip has been annotated for the presence or absence of every instrument. While this is true for every instrument dataset—if one considers instruments outside its vocabulary—it is usually not

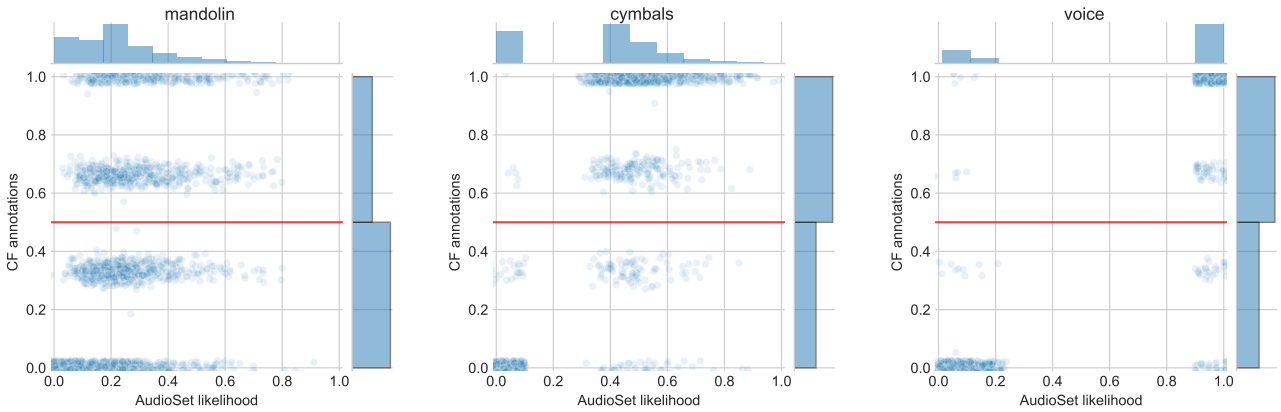

Figure 5. The distribution of the initial model likelihood compared to the crowd-sourced annotations for three instruments. Each dot represents a clip, the horizontal line indicates the majority vote threshold, and the marginal distributions of model likelihood and crowd agreement are shown as bar plots. Data have been randomly perturbed for clarity of visualization.

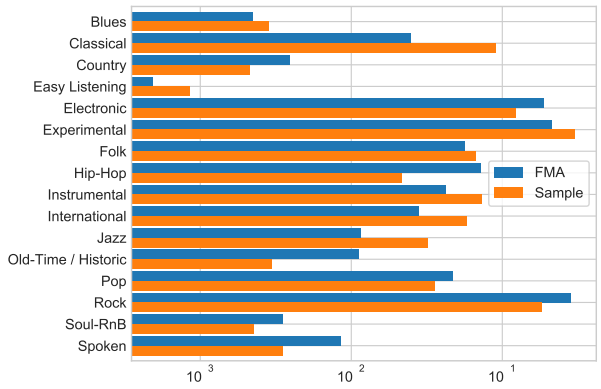

Figure 6. The distribution of (top) genres over the selected sample clips, and the background population in FMA.

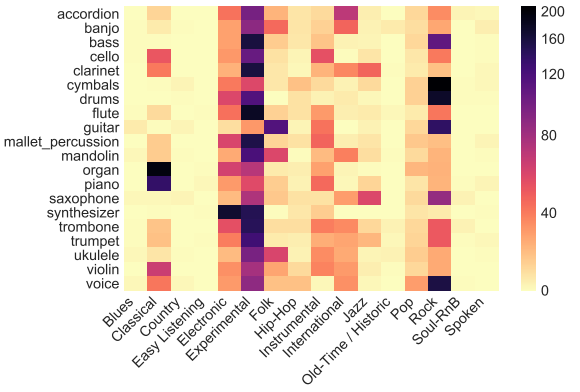

Figure 7. The distribution of (top) genres over the positive candidate sets for each instrument.

taken into consideration as part of the dataset design.

More generally, previous datasets have not typically been designed with a plan for future correction, revision, and expansion. We are explicitly planning to expand and revise the dataset over time, either by additional crowd-sourcing, semi-supervised learning [6], or incremental evaluation [15]. OpenMIC-2018 will be placed under version control, archived, and each revision will receive a unique document object identifier (DOI) via Zenodo. [6]

In addition to supporting corrections and expanded coverage, we anticipate expanding the vocabulary beyond the initial 20 classes, both in breadth of instrument classes, and in depth to provide refinements of classes, such as *alto saxophone* and *tenor saxophone* rather than *saxophone*. Similarly, future work could re-use much of the framework developed here to annotate the same collection for a variety of qualities beyond instrumentation, and facilitate the development of integrated multi-task models.

Acknowledgments. B.M. is supported by the Moore-Sloan Data Science Environment at NYU.

[6] `http://about.zenodo.org/`

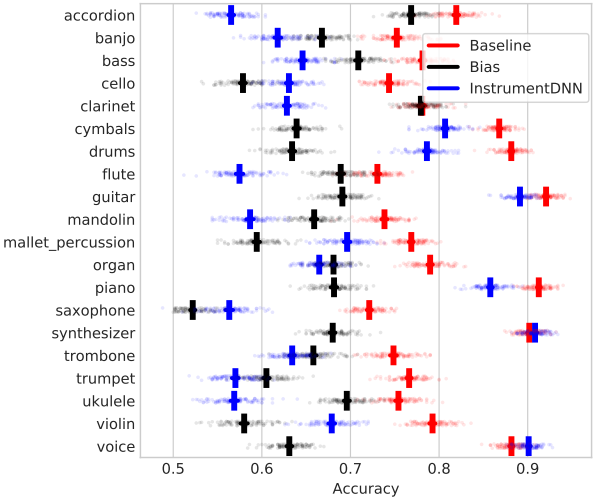

Figure 8. Accuracy of the Random Forest baseline (red), InstrumentDNN (blue), and the dataset bias (black). The RandomForest was trained on OpenMIC-2018, while InstrumentDNN was trained on AudioSet.

6. REFERENCES

[1] Giuseppe Bandiera, Oriol Romani Picas, Hiroshi Tokuda, Wataru Hariya, Koji Oishi, and Xavier Serra. Good-sounds. org: A framework to explore goodness in instrumental sounds. In *Proceedings of the 17th International Society for Music Information Retrieval Conference*, pages 414–419, 2016.

[2] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 591–596, 2011.

[3] Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *ISMIR*, volume 14, pages 155–160, 2014.

[4] Juan J Bosch, Jordi Janer, Ferdinand Fuhrmann, and Perfecto Herrera. A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals. In *ISMIR*, pages 559–564, 2012.

[5] Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, E Law, J Bello, and O Nov. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proceedings of the ACM on Human-Computer Interaction*, 1(1), 2017.

[6] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

[7] Creative Commons. About the licenses. 2015. https://creativecommons.org/about/.

[8] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, pages 316–323, 2017.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[10] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders. 2017.

[11] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 776–780. IEEE, 2017.

[12] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Music genre database and musical instrument sound database. 2003.

[13] Aren Jansen, Jort F Gemmeke, Daniel PW Ellis, Xiaofeng Liu, Wade Lawrence, and Dylan Freedman. Large-scale audio event discovery in one million youtube videos. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 786–790. IEEE, 2017.

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

[15] B. McFee, E.J. Humphrey, and J. Urbano. A plan for sustainable MIR evaluation. In *17th International Society for Music Information Retrieval Conference*, ISMIR, 2016.

[16] George A Miller. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2016.

[19] John Thickstun, Zaid Harchaoui, and Sham Kakade. Learning features of music from scratch. In *International Conference on Learning Representations*, 2017.