A PLAN FOR SUSTAINABLE MIR EVALUATION

Brian McFee
Center for Data Science / MARL
New York University
brian.mcfee@nyu.edu

Eric J. Humphrey
Spotify, Ltd.
ejhumphrey@spotify.com

Julián Urbano
Music Technology Group
Universitat Pompeu Fabra
urbano.julian@gmail.com

ABSTRACT

The Music Information Retrieval Evaluation eXchange (MIREX) is a valuable community service, having established standard datasets, metrics, baselines, methodologies, and infrastructure for comparing MIR methods. While MIREX has managed to successfully maintain operations for over a decade, its long-term sustainability is at risk. The imposed constraint that input data cannot be made freely available to participants necessitates that all algorithms run on centralized computational resources, which are administered by a limited number of people. This incurs an approximately linear cost with the number of submissions, exacting significant tolls on both human and financial resources, such that the current paradigm becomes *less* tenable as participation increases. To alleviate the recurring costs of future evaluation campaigns, we propose a distributed, community-centric paradigm for system evaluation, built upon the principles of openness, transparency, reproducibility, and incremental evaluation. We argue that this proposal has the potential to reduce operating costs to sustainable levels. Moreover, the proposed paradigm would improve scalability, and eventually result in the release of large, open datasets for improving both MIR techniques and evaluation methods.

1. INTRODUCTION

Evaluation plays a central role in the development of music information retrieval (MIR) systems. In addition to empirical results reported in individual articles, community evaluation campaigns like MIREX [6] provide a mechanism to standardize methodology, datasets, and metrics to benchmark research systems. MIREX has earned a special role within the MIR community as the central forum for system benchmarking. However, the annual operating costs incurred by MIREX are unsustainable by the MIR community. Much of these costs derive from one-time expenditures — *e.g.*, the time spent getting a participant's algorithm to run — which primarily benefit individual participants, but not the MIR community at large. If we, as a community, are to continue hosting regular evaluation campaigns, we will soon require a more efficient and sustainable model.

The evaluation problem has lurked the MIR community for years. The MIReS Roadmap for Music Information ReSearch identified it as one of the main technical-scientific grand challenges in MIR research [20], and during the ISMIR 2012 conference a discussion panel was held to explicitly address this issue [17]. Previous research has discussed some limitations of MIREX-like evaluation and made general proposals to avoid them [21, 23], and other community-led platforms have been put forward to try to minimize them in practice, most notably MusiClef [14] and the MSD Challenge [11]. However, for different reasons, they have been unable to continue operating.

Reflecting upon the prior work, we propose in this article a sustainable, open framework for community-driven MIR evaluation campaigns. Our proposal is motivated by three complementary factors. First, we strive to reduce the cost of running and maintaining the evaluation framework. Second, we hope to improve transparency and openness wherever possible. Third, we plan to establish a sustainable framework that will produce open, public data sets consisting of both inputs and reference annotations. By directing the majority of resources toward the production of open data, the proposed framework will be of value to the greater MIR community in perpetuity, and benefits will not be limited to participants in a particular year's campaign.

We stress that this document describes not a fully implemented framework, but a specific *proposal* put forward by a group of authors dedicated to seeing it put into practice. Our goal in writing this document at this early stage is to solicit input from the MIR community before implementation details have been finalized. In collaboration with the community, we hope to develop a framework that benefits as many people as possible and requires minimal financial support for years to come.

2. MIREX

The Music Information Retrieval Evaluation eXchange is a framework for the community driven evaluation of MIR algorithms [6, 7]. The annual tradition of MIREX was established early in the lifetime of ISMIR, and drew significant inspiration from the well-established TREC framework [24]. Thanks in large part to the vision of MIR pioneers, the first official iteration took place at ISMIR 2005 after much preliminary work, including a trial run the year

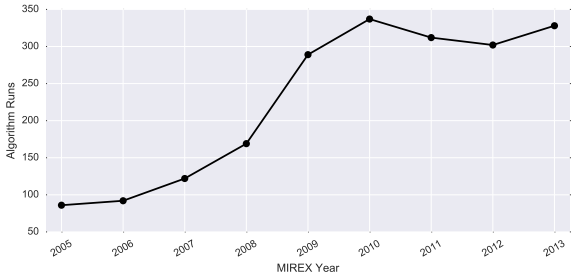

Figure 1. Number of algorithms run at MIREX over the course of almost a decade.

prior called the Audio Description Contest (ADC). The practicalities of MIREX are hosted by the IMIRSEL group at UIUC, and the organization has successfully earned multiple grants to jump-start the evaluation effort at ISMIR.

At a high level, MIREX operates in the following steps:

1. Identify some task of interest, *e.g.*, chord estimation.
2. Formulate the problem and evaluation metrics.
3. Construct (and annotate) a corpus of data.
4. Release a subset of the data for development purposes; retain the rest as private data for evaluation.
5. Invite participants to submit system implementations, which then are executed on private servers.
6. Evaluate predicted outputs against reference annotations or human judgments.
7. Repeat steps 5–6 annually. Intermittently revisit step 2 if needed, and steps 3 and 4 if possible.

Importantly, this approach differs from TREC-style evaluation by operating in an "algorithm-to-data" model, where facilitators oversee the application of code submissions to privately held data, rather than participants submitting predictions over a freely available dataset. The rationale for this decision is understandable. In contrast to other machine perception domains, such as natural language processing, speech recognition, or computer vision, intellectual property and copyright law imposes stiff penalties on the illegal distribution of recorded music. Due to a history of litigation from the recording industry, there is a pervasive sense of fear in the MIR community that sharing copyrighted audio data would invite crippling lawsuits [6].

However, experience with MIREX over the last decade has demonstrated that bringing algorithms to data entails fundamental limitations. First, as a matter of practicality, doing so incurs steep computational and financial costs that the community cannot hope to sustain indefinitely. Running hundreds of research systems demands significant computational resources, which must be either rented or purchased outright. More often than not, these systems are research prototypes which require substantial, manual intervention to operate correctly, and are seldom optimized for efficiency or ease-of-use. While task-dependent runtime limits are placed on algorithm execution (between 6 and 72 hours), MIREX requires months, if not years, of annual compute-time.

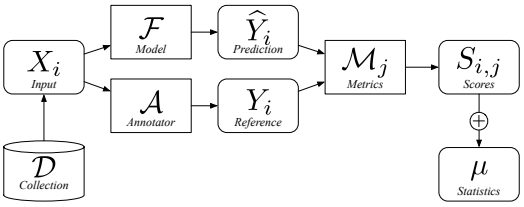

Figure 2. Diagram of the standard approach to the benchmarking of MIR systems.

The financial burden of computation can be negligible in comparison to the requisite human effort. As a point of reference, MIREX 2007 alone required "nearly 1000 person-hours" to supervise the execution of 122 algorithms from 40 teams [6]. As illustrated in Figure 1, the number of algorithm runs at MIREX has nearly tripled in the years since. [1] As a rough estimate, the last decade has likely consumed on the order of *10,000* person-hours just bringing algorithms to data. Not only is this rate unsustainable, but the combined operating costs only *increase* with participation. Said differently, the worst thing that could happen to MIREX in its current form is growth.

Operating costs aside, MIREX has indeed produced valuable insights into the development of MIR systems [6]. Unfortunately, many scientific endeavors are largely impeded or, at worst, wholly obfuscated in the current paradigm. To illustrate, consider the standard approach to benchmarking MIR systems as depicted in Figure 2. An input, X_i, is observed by an annotator, $\mathcal{A}$, who produces a reference output, Y_i. Similarly, a system $\mathcal{F}$ operates on the same input, and produces an estimate $\widehat{Y}_i$. Each of several performance metrics, $\mathcal{M}_j$, are applied to these two representations, yielding a number of performance scores, $S_{i,j}$ This process is repeated over a sample of n input-output pairs, $\{X_i, Y_i\} \in \mathcal{D}$, and the sample-wise scores are aggregated into summary statistics, μ, the reliability of which generally increases with $n = |\mathcal{D}|$.

In the current MIREX formulation, a lack of transparency renders participants scientifically blind in a number of ways [23]. First, there is no direct access to the reference annotations Y_i, and, in most cases, no access to the input X_i either. Furthermore, $\widehat{Y}_i$ is of little use without X_i for context. This makes it exceedingly difficult to learn from the results of the evaluation. Without access to the underlying data, how can one diagnose the cause of an erroneous estimate, or discover avenues for improvement? Similarly, there is no way to gauge the distribution of the data or estimate any bias in the sampling, beyond what may be inferred from public development sets.

The behavior of the human annotators is also obscured, as are the instructions provided when the annotation was initially performed [16]. Consequently, the problem formulation itself is effectively hidden, and subject to drift over time. For the sake of visibility into the evaluation metrics $\mathcal{M}_j$, the original NEMA infrastructure is open

[1] `https://www.hathitrust.org/documents/mirex_htrc_same_rda.pdf`

source, and an ongoing community-led effort continues to standardize and improve upon these implementations [18]. Still, without access to the data, it is exceedingly difficult to perform *meta-evaluation*, like comparing new metrics on old data, without seeking privileged access to the historical records.

Finally, it is only fair to admit that large scale evaluation is a considerable undertaking with plenty of room for error and misinterpretations. Conducting these campaigns in the open makes it easier to detect and diagnose any missteps.

Due to the issues highlighted above, resources that could have been devoted to constructing and enlarging open data sets have instead been absorbed by irrecoverable operating costs. This is not only detrimental to system development, but also to evaluation, because it is critical to routinely refresh the evaluation data to reduce bias and prevent statistical over-fitting. Without a fresh source of annotated data, early concerns about the community eventually over-fitting to benchmark datasets are beginning to manifest.

In some cases, this is because the data used for evaluation exists in the wild: one submission in 2011 achieved nearly perfect scores in chord estimation due to being pre-trained on the test data. [2] In other cases, participants may mis-interpret the task guidelines, as evidenced by submissions of offline systems for tasks that are online, or by others that mistakenly use annotations from previous folds in a train-test task. Across the board, hidden evaluation data is slowly being over-fit by trial and error, as teams implicitly leverage the weak error signal across campaigns to "improve" systems. These results can be misleading due to the fixed but unknown biases in the data distribution, which become apparent when datasets are expanded — like the introduction of the Billboard dataset to the chord estimation task in 2012 [2] — or as further insight is gained, as in the disclosure of the meta-data in the music similarity corpus.

To make matters worse, there is no feasible plan in place to replenish the evaluation datasets currently used by MIREX, nor any long-term plans to replace that data when it inevitably becomes stale. MIREX has primarily relied upon the generosity of the community to both curate and donate data for the purposes of evaluation. This approach also struggles with transparency, making it susceptible to issues of methodology and problem formulation, and can hardly be relied upon as a regular resource. Data collection is a challenging, unavoidable reality that demands a viable plan going forward.

After a decade of MIREX, we have learned which techniques work and which do not. Most importantly, we have learned the importance of establishing a collective endeavor to periodically and systematically evaluate MIR systems. Unfortunately, we have also learned of the burdens entailed by such an initiative, and the limitations of its current form.

[2] http://nema.lis.illinois.edu/nema_out/mirex2011/results/ace/nmsd2.html

3. OPEN EVALUATION OF MIR SYSTEMS

Summarizing the previous section, MIREX suffers from three deficiencies that render the situation untenable: (i) the financial and labor costs cannot be sustained indefinitely, (ii) the lack of data transparency limits the scientific value of the endeavor, (iii) the lack of a strategy for obtaining and annotating new data.

Thus, to address these deficiencies, our proposed plan has three key differentiators from the MIREX model:

(i) Distributed computation reduces operating costs to scale favorably with increased participation.

(ii) Freely distributable audio facilitates reproducibility and benefits the entire community.

(iii) Incremental evaluation reduces bias by keeping test data fresh, and provides a feasible strategy for collecting new data.

3.1 Distributed computation

The primary difficulty in running an evaluation campaign is computing the outputs of all participating systems. This difficulty stems from two sources. First is the obvious computational complexity of running m submissions over n inputs. Second is the less obvious "human" complexity entailed in the task captains successfully executing the submitted programs in a foreign environment. While the computational issues can be ameliorated by running systems in parallel over multiple machines, the cost in human effort has no easy solution in the MIREX framework.

An alternative to this framework is exemplified by Kaggle. [3] Kaggle competitions are conducted with all input data publicly available, and participants submit only the outputs of their systems, *e.g.*, predictions made for each data point. This paradigm effectively resolves the difficulties listed above: both computation and human effort are distributed to the participants, rather than centralized at the evaluation site and task captains. This dramatically reduces the cost of maintenance and administration. However, we note a few potential challenges inherent to bringing data to the computation.

First, the input data must be made openly available to participants. This may increase the risk of bias if participants (unintentionally) tune their systems to the evaluation set. To mitigate bias, we propose the use of a large and diverse corpus of common tracks which are shared across *all tasks*, rather than a collection of small, task-specific datasets as is done in MIREX. For any given task, only a subset of the data need to be considered when comparing systems, and the evaluation set may be independently selected for each task. The knowledge of which items comprise a given evaluation subset would remain hidden from participants at submission time. This implies that each submission must span the entire corpus, reducing the feasibility of participants tuning their algorithms to a particular subset. Moreover, while this requirement increases computational overhead for the participant, it results in a large

[3] http://kaggle.com

collection of outputs for various methods on a common corpus, which is a valuable data source in its own right.

Second, distributed computation entails its own challenges with respect to transparency. While MIREX requires submissions to execute on a remote server beyond the participants' control, the scheme proposed here drops that requirement in favor of visibility into system inputs. Consequently, restrictions on the implementation (*e.g.*, running time) would become infeasible, and it may open the possibility of cheating by manual participant intervention. Using a large corpus with opaque evaluation sets will limit the practicality of this approach. [4] Obscuring which items belong to the evaluation subset comes at the expense of sample-wise measures, however, as doing so would reveal the partition. This is not inherently problematic if done following the completion of a campaign (instead of powering a continuous leader-board), but would require changing the evaluation set annually. These are minor concessions, and we argue that open data benefits the community at large, since its availability will serve a broader set of interests over time.

Finally, the proposed scheme assumes that multiple tasks operate on a common input form, *i.e.*, recorded audio. While the majority of current MIREX tasks do fit into this framework, at present it precludes those which require additional input data (score following, singing voice separation) or user interaction (grand challenges, query-by-humming/tapping). Our long-term plan is to devise ways of incorporating these tasks as well, while keeping with the principles outlined above. This is of course an open question for further research.

3.2 Open content

For the distributed computation model to be practical, we first need a source of diverse and freely distributable audio content. This is significantly easier to acquire now than when MIREX began, and in particular, a wealth of data can be obtained from services like Jamendo [5] and the Free Music Archive (FMA). [6] Both of these sites host a variety of music content under either public domain or Creative Commons (CC) licenses. [7] Since CC-licensed music can be freely redistributed (with attribution), it is (legally) possible to create and share persistent data archives.

The Jamendo and FMA collections are both large and diverse, and both can be linked to meta-data: Jamendo via DBTune to MusicBrainz [19] and FMA to Echo Nest/Spotify identifiers. Jamendo claims over 500,000 tracks charted under six major categories: *classical*, *electronic*, *hip-hop*, *jazz*, *pop*, and *rock*. FMA houses approximately 90,000 tracks which are charted under fifteen categories: *blues*, *classical*, *country*, *electronic*, *experimental*, *folk*, *hip-hop*, *instrumental*, *international*, *jazz*, *old-time/historic*, *pop*, *rock*, *soul/rnb*, and *spoken*. These categories should not necessarily be taken as ground truth annotations, but they reflect the tastes and priorities of their respective user communities. While there is undoubtedly a strong western bias in these corpora, the same can be said of MIREX's private data and the MIR field itself. However, using open content at least permits practitioners to quantify and possibly correct this bias.

Aside from western/non-western bias, there is also the potential for free/commercial bias. A common criticism of basing research on CC-licensed music is that the music is of substantially lower "quality" — which may refer to either artistry or production value, or both — than commercial music. This point is obviously valid for high-level tasks such as recommendation, which depend on a variety of cultural, semantic, and subjective factors beyond the raw acoustic content of a track. However, for the majority of MIREX tasks, in particular perceptual tasks like onset detection or source identification, this is a much more tenuous case. We do, however, note that FMA includes content by a variety of commercially successful artists, [8] but the vast majority of content in both sources is provided by relatively unknown artists, which makes it difficult to control for "quality".

To help users navigate the collections, both Jamendo and FMA provide popularity-based charts and community-based curation, in addition to the previously mentioned genre categories. Taken in combination, these features can be leveraged to pre-emptively filter the collections down to subsets of tracks which are either of interest to a large number of listeners, of interest to a small number of listeners with specific tastes, or representative of particular styles. This approach is similar in spirit to previous work using chart-based sampling, *e.g.* Billboard [2].

3.3 Incremental evaluation

In its first cycle of operation, the proposed framework requires a new, unannotated corpus of audio. Rather than fully annotating the corpus up front for each task, we plan to adopt an *incremental evaluation* strategy [4].

With incremental evaluation, the set of reference annotations need not be complete: some (most) input examples will not have corresponding annotations. Consequently, performance scores are estimated over a subset of annotated tracks, which may itself grow over time. Systems are ranked as usual, but with a degree of uncertainty inversely proportional to the number of available annotations.

Initially, uncertainty in the performance estimates will be maximal, due to a small number of available annotations. Subsequently, as reference annotations are collected and integrated to the evaluation system, they will be compared to the submissions' estimated annotations, and provide incrementally accurate and precise performance estimates. Prior research in both text and video IR has demonstrated that evaluation against incomplete reference data is feasible, even when only a small fraction of the annotations are known [3, 15, 25].

[4] Even in the unlikely event that a participant "cheats" by obtaining human-generated annotations, the results can be publicly redistributed as free training data, so the community ultimately wins out.

[5] http://jamendo.com

[6] http://freemusicarchive.org/

[7] https://creativecommons.org/licenses/

[8] https://en.wikipedia.org/wiki/Free_Music_Archive#Notable_artists

This raises the question: which inputs are worth annotating? Consider two systems that produce the same annotation for a fixed input example. Whether they are both right or wrong with respect to the reference annotation, there is no way to distinguish between them according to that example, so there is little value in seeking its annotation. Conversely, examples upon which multiple systems *disagree* are more likely to produce useful information for distinguishing among competing systems. Several recent studies have investigated the use of algorithm disagreement for this issue, be it for beat tracking [8], structural analysis [13, chapter 4], or chord recognition [9]. Others have studied alternative methods for music similarity [22], choosing for annotation the examples that will be most informative for differentiating between systems. In general, these methods allow us to minimize the required annotator effort by prioritizing the most informative examples. With many participating systems, and multiple complex performance metrics, prioritizing data for annotation is by no means straightforward. We hope that the community will assist in specifying these procedures for each task.

In the proposed framework, annotations may come from different sources, so it is imperative that we can trust or validate whatever information is submitted as reference data. Another line of work is thus the development and enforcement of standards, guidelines, and tools to collect and manage annotations. This will require developing web services for music annotation, as well as appropriate versioning and quality control mechanisms. Quality control is particularly important if annotations are collected via crowd-sourcing platforms like Amazon Mechanical Turk. [9]

3.4 Putting it all together

In contrast to the outline in Section 2, our proposed strategy would proceed as follows:

1. Identify and collect freely distributable audio.
2. Define or update tasks, *e.g.*, chord transcription.
3. Release a development set (with annotation), and the remaining unlabeled data for evaluation.
4. Collect *predictions* over the unlabeled data from participating systems.
5. Collect reference annotations, prioritized by disagreement among predictions and informativeness.
6. Estimate and summarize each submission's performance against the reference annotations.
7. Retire newly annotated evaluation data, adding it to the training set for the next campaign.
8. Go to step 3 and repeat. Revisit steps 1–2 if needed.

Steps 1–3 essentially constitute the startup cost, and are unavoidable for tasks which lack existing, open data sets. However, from the perspective of administration, only steps 3 and 5 require significant human intervention (*i.e.*, annotation), and both steps directly result in public goods. In this way, the proposed system will be significantly more efficient and cost-effective than MIREX.

[9] `https://www.mturk.com/`

4. DISCUSSION

In this section, we enumerate the goals and implementation of the proposed framework.

4.1 Timing: why now?

The challenges described in this document, which our proposed strategy aims to address, are not news to the community. The operating costs of MIREX became apparent early in its lifetime, and concerns about its sustainability loom large among researchers. That said, little has been done to resolve the situation in the intervening years. This raises an obvious question: why will things change *now*?

In many ways, the approach taken by MIREX made perfect sense in the early 2000's. However, recent infrastructural advances, coupled with growth and maturation of the MIR community, have introduced new possibilities. First and foremost, creative commons music is now ample, bringing the dream of sharing data within grasp. Cloud computing is cheap and ubiquitous, and can dramatically reduce the administrative costs of evaluation infrastructure and persistent data storage. Improvements in web infrastructure have also resolved many of the challenges of large-scale data distribution. Finally, browser support for interactive applications enables the development of web-based annotation tools, which significantly reduces the barrier to entry for annotators.

More broadly speaking, the community has matured significantly since MIREX began in 2005. Open source development and reproducible methods are now commonplace, but we remain hindered by the lack of open data for evaluation. Only by developing frameworks for open evaluation and data collection, can we further develop as a scientific discipline.

4.2 Implementation details

Effectively deploying the proposed framework will require two things: infrastructure development, and hosting. On the infrastructure side, we can leverage several existing open source projects: `mir_eval` for scoring [18], JAMS for data format standardization [10], and CodaLab for running the evaluation campaign. [10] The last remaining software component is a platform for collecting annotations. In addition to traditional desktop software, browser-based annotation tools would facilitate distributed data collection, and a simple content management system could collect annotations as they are completed.

As for hosting, since the burden of executing arbitrary (submitted) code is removed, the remaining software components can reside on either a private university server, or, more likely, cloud-based hosting such as Amazon Web Services. [11] Similarly, the audio data can be distributed via BitTorrent to participants, or hosted (at some minor cost) for traditional download.

[10] `http://codalab.org/`
[11] `https://aws.amazon.com/`

4.3 Data and evaluation integrity

Allowing participants to submit predictions, rather than software, may raise questions about integrity: how can we verify the process which generated the predictions? Ultimately, we must rely on participants to be honest in describing their methods. Although it is *possible* for a participant to manually annotate each track and achieve high scores, we hope that the scale of the dataset will make this approach fundamentally impractical. Additionally, in keeping with the spirit of open science and reproducible methods, we will encourage participants to link to a repository containing their software implementation, which can be used to independently verify the predictions.

When it comes to data integrity, we acknowledge that music is unique in that its perception is impacted by a plethora of cultural and experiential factors. In particular, CC-licensed music may lie outside of the gamut of commonly studied music, and differences in compositional style, instrumentation, or production, may lead to difficulties in validation and generalization. While this is unlikely to impact low-level tasks such as onset detection or instrument identification, more abstract tasks, such as chord estimation or structural analysis, may be more sensitive to selection bias. Relying on curation and chart popularity as a pre-filter in data selection may help to mitigate these effects. After collecting an initial set of annotated CC-licensed music, it will be possible to quantitatively measure the differences in system performance compared to existing corpora of commercial music.

4.4 Collecting annotations

Statistically valid evaluation requires a continuous source of freshly annotated data. At present, we see three potential sources to satisfy this need.

First is the traditional route of raising funds and paying expert annotators. This option incurs both a direct financial cost and various hidden human costs, but is also the most likely to produce high-quality data, and may in fact be the only viable path for certain tasks. In the grand scheme of things, however, the financial burden may not be so severe. As a point of reference, MedleyDB, a well-curated dataset of over 100 multi-track recordings for a number of MIR tasks, cost approximately $12 per track to annotate ($1240 total) [1]. [12] The ISMIR Society maintains a membership of roughly 250 individuals each year: a $5 increase in membership dues would cover the annotation cost of a new dataset like MedleyDB annually. Grants or industrial partnerships could also subsidize annotation costs.

Second, for tasks that require minimal annotator training, we can leverage either crowd-sourcing platforms, *e.g.*, Mechanical Turk, or seek out enthusiastic communities interested in voluntary *citizen science* for music. Websites such as Genius [13] (6M monthly active users) and Ultimate Guitar [14] (3M monthly active users) demonstrate the existence of these communities for lyrics and guitar tablature. [15] As witnessed by the success of eBird [16] or AcousticBrainz, [17] motivated people with the right tools can play a substantial role in scientific endeavors.

Finally, if no funding can be found to annotate data, we may solicit annotations directly from participants and the ISMIR community at large. This approach has been effectively used to collect judgements for the audio and symbolic music similarity tasks.

In each case, we will institute a ratification system so that annotations are independently validated before inclusion in the evaluation set. As mentioned in section 4.2, web-based annotation tools will enable volunteer contribution, which can supplement paid annotations.

5. NEXT STEPS: A TRIAL RUN

We conclude by advocating a large-scale instrument identification task for ISMIR2017. In this task, the presence of active instruments (under a pre-defined taxonomy) are estimated over an entire recording. The taxonomy may be readily adapted from WordNet [12], and refined by the community, starting at this year's ISMIR conference. There are a number of strong motivations for pursuing instrument identification. It is an important, classic problem in MIR, but is currently absent from MIREX. Researchers typically explore the topic with disparate datasets, and the problem remains unsolved to an unknown degree. Compared with other music perception tasks, annotation requires a simple interface, *e.g.*, "check all that apply", and the task definition itself is relatively unambiguous: a particular instrument is either present in the mix or not. Instrument occurrence is largely independent of popularity, which results in a fairly minimal bias due to the use of CC-licensed music. Finally, computer vision found fantastic success with a similar endeavor, known as ImageNet [5], in which algorithms detect objects in natural images.

The following steps are necessary to realize this goal within the coming year: (i) establish a robust instrument taxonomy; (ii) acquire a large sample of audio content; (iii) build a web-based annotation tool and storage system; (iv) construct a development set; (v) implement or collect a few simple algorithms to prioritize content for initial annotation; (vi) perform data collection, through some combination of paid annotation, crowd-sourcing, and community support; and finally, deploy an evaluation server and leader-board to accept and score submissions.

Each of these components, while requiring some engineering and organizational efforts, are achievable goals with the help of the ISMIR community.

Acknowledgments. B.M. is supported by the Moore-Sloan Data Science Environment at NYU. J.U. is supported by the Spanish Government: JdC postdoctoral fellowship, and projects TIN2015-70816-R and MDM-2015-0502.

[12] Figures provided via personal communication with the authors.
[13] http://genius.com/
[14] http://www.ultimate-guitar.com/
[15] Data gathered from http://compete.com, March 2016
[16] http://ebird.org/
[17] http://acousticbrainz.org/

6. REFERENCES

[1] R.M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J.P. Bello. Medleydb: A multitrack dataset for annotation-intensive mir research. In *International Society for Music Information Retrieval Conference*, ISMIR, pages 155–160, 2014.

[2] J.A. Burgoyne, J. Wild, and I. Fujinaga. An expert ground truth set for audio chord recognition and music analysis. In *International Society for Music Information Retrieval Conference*, ISMIR, 2011.

[3] B. Carterette. Robust Test Collections for Retrieval Evaluation. In *SIGIR*, pages 55–62, 2007.

[4] B. Carterette and J. Allan. Incremental test collections. In *ACM International conference on Information and Knowledge Management*, pages 680–687. ACM, 2005.

[5] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, CVPR, pages 248–255. IEEE, 2009.

[6] J.S. Downie. The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[7] J.S. Downie, A.F. Ehmann, M. Bay, and M.C. Jones. The music information retrieval evaluation exchange: Some observations and insights. In *Advances in music information retrieval*, pages 93–115. Springer, 2010.

[8] A. Holzapfel, M.E.P. Davies, J.R. Zapata, J.L. Oliveira, and F. Gouyon. Selective sampling for beat tracking evaluation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(9):2539–2548, 2012.

[9] E.J. Humphrey and J.P. Bello. Four timely insights on automatic chord estimation. In *International Society for Music Information Retrieval Conference*, ISMIR, pages 673–679, 2015.

[10] E.J. Humphrey, J. Salamon, O. Nieto, J. Forsyth, R.M. Bittner, and J.P. Bello. JAMS: A JSON annotated music specification for reproducible MIR research. In *ISMIR*, pages 591–596, 2014.

[11] B. McFee, T. Bertin-Mahieux, D. Ellis, and G. Lanckriet. The million song dataset challenge. In *4th International Workshop on Advances in Music Information Research*, AdMIRe, April 2012.

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

[13] O. Nieto. *Discovering structure in music: Automatic approaches and perceptual evaluations*. PhD thesis, New York University, 2015.

[14] N. Orio, D. Rizo, R. Miotto, M. Schedl, N. Montecchio, and O. Lartillot. Musiclef: a benchmark activity in multimodal music information retrieval. In *International Society for Music Information Retrieval Conference*, ISMIR, pages 603–608, 2011.

[15] P. Over, J. Fiscus, G. Sanders, D. Joy, M. Michel, G. Awad, A.F. Smeaton, W. Kraaij, and G. Quenot. TRECVID 2014–An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics. In *TREC Video Retrieval Evaluation Conference*, 2014.

[16] G. Peeters and K. Fort. Towards a (Better) Definition of the Description of Annotated MIR Corpora. In *International Society for Music Information Retrieval Conference*, pages 25–30, 2012.

[17] G. Peeters, J. Urbano, and G.J.F. Jones. Notes from the ISMIR 2012 Late-Breaking Session on Evaluation in Music Information Retrieval. In *International Society for Music Information Retrieval Conference*, 2012.

[18] C. Raffel, B. McFee, E.J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D.P.W. Ellis. mir_eval: A transparent implementation of common mir metrics. In *International Society for Music Information Retrieval Conference*, ISMIR, 2014.

[19] Y. Raimond and M.B. Sandler. A web of musical information. In *ISMIR*, pages 263–268, 2008.

[20] X. Serra, M. Magas, E. Benetos, M. Chudy, S. Dixon, A. Flexer, E. Gómez, F. Gouyon, P. Herrera, S. Jorda, O. Paytuvi, G. Peeters, J. Schlüter, H. Vinet, and G. Widmer. Roadmap for Music Information ReSearch, 2013.

[21] B.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.

[22] J. Urbano and M. Schedl. Minimal Test Collections for Low-Cost Evaluation of Audio Music Similarity and Retrieval Systems. *International Journal of Multimedia Information Retrieval*, 2(1):59–70, 2013.

[23] J. Urbano, M. Schedl, and X. Serra. Evaluation in Music Information Retrieval. *Journal of Intelligent Information Systems*, 41(3):345–369, 2013.

[24] E.M. Voorhees and D.K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[25] E. Yilmaz, E. Kanoulas, and J.A. Aslam. A Simple and Efficient Sampling Method for Estimating AP and NDCG. In *SIGIR*, pages 603–610, 2008.