NOTES FROM THE ISMIR 2012 LATE-BREAKING SESSION ON EVALUATION IN MUSIC INFORMATION RETRIEVAL

Geoffroy Peeters
STMS IRCAM-CNRS-UPMC
geoffroy.peeters@ircam.fr

Julián Urbano
University Carlos III of Madrid
jurbano@inf.uc3m.es

Gareth J. F. Jones
Dublin City University
gjones@computing.dcu.ie

ABSTRACT

During the last day of the ISMIR 2012 conference there were two events related to Music IR Evaluation. A panel took place during the morning to discuss several issues concerning the various evaluation initiatives with the general audience at ISMIR. A late-breaking session during the afternoon kept the discussion alive between a group of researchers who wanted to dig deeper into these issues. This extended abstract reports the main topics covered during this short session and the general thoughts that came up.

1. PANEL SESSION

Since MIREX [1] first appeared in 2005, other MIR evaluation forums also started in the last couple of years, namely the Million Song Dataset Challenge [2] and MediaEval MusiCLEF [3]. Although these initiatives are all independent from ISMIR 2012 and are organized by different institutions and groups of individuals, a special panel session of the conference was dedicated to reporting on these initiatives and to reflect on evaluation methodologies in MIR. The aims of this panel were to discuss the methodologies currently used in MIR evaluations and compare them to the evaluation practices in other research fields. The following are the main topics covered during the panel session.

Methodology for task definition. What methodology should be used to define a task (bottom-up vs. top-down)? For which purpose should a task be evaluated: low-level tasks (process-oriented such as beat, chords) vs. full-system tasks (user-oriented such as music recommendation systems). Specific tasks that are part of large-scale international evaluations define de facto the specific topics that new contributors to the MIR field will work on. The methodology followed to define tasks is therefore of utmost importance.

Methodology for evaluation. How should a specific task be evaluated? Which data and which measures? What is the validity and reliability of the results obtained? Measures and data used in large-scale international evaluations define de facto the standards for the specific tasks. The methodology followed to define the data and the measures is therefore of utmost importance.

Data. How to get more data? How to deal with data availability (not only music collections, but also raw system outputs, judgments, annotations)? Should we explore low-cost evaluation methodologies? Currently, most MIR systems are concerned with audio-only or symbolic-only scenarios, but multi-modal systems (such as aggregating information from the audio-content, from lyrics content or web mining) should allow deciding also on the impact on final user application of each technology.

Methodology. What is the best methodology to drive improvements? What kind of evaluation framework (open vs. close)? What could be improved in previous evaluation initiatives? How can we make results reproducible? How can we make MIR evaluation sustainable along time?

2. LATE-BREAKING SESSION

There were 17 people participating to this late-breaking session. The discussion followed the lines of the Panel on Evaluation Initiatives that was held in the morning during ISMIR. The following are the main topics covered during the late-breaking session, comments from the attendees and our personal view regarding some points.

Methodology. There was a general consensus on the need to question the evaluation methodologies we currently follow with MIR tasks. As mentioned during the morning panel, some of the tasks evaluated did not start from a clear user need, and there were no community discussions on the development of the most appropriate test-sets, annotation and annotation procedures, evaluation measures, etc. In fact, tasks are often initiated thanks to graduate students who build a test-set and make it available to the rest of the community or for evaluation initiatives. We find this very problematic because there is a lack of documentation regarding these methodologies, making it very difficult to assess the reliability of the very evaluations and processes followed, so much that in some cases it is even impossible to carry out similar experiments in other labs. This is particularly problematic for newcomers, who are often faced with one of the MIR tasks but can get overwhelmed very easily due to the lack of clear and centralized documentation.

Data. The MIR community has gotten used to the fact that test-sets can not be made publicly available. This has been justified by the fact that music audio data are in many cases under copyright and by the fact that distributing the ground-truth once an evaluation performed would involve

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME
[2] http://www.kaggle.com/c/msdchallenge
[3] http://www.multimediaeval.org/mediaeval2012/newtasks/music2012/

being able to create a new test-sets for the next evaluation which remain very costly for our community. For this last reason, the same test-sets are often used over the successive experiments, and therefore their inaccessibility is supposed to refrain researchers from cheating or overfitting. However, in our view, this inaccessibility slows down improvement of systems. With only performance figures over the years, there is no way we can know why our systems failed or succeeded, which is the key for improvement. An example of this was given during the morning panel for the Beat Tracking task for which all systems seem to perform very badly for a specific song, but there is no way of knowing what this song is. As pointed out during the morning panel, the issue of the copyright related to the music could be solved by using copyright free music (such as Jamendo)

Model. The data issue begs the question of the evaluation model. Due to the privacy of datasets, the MIR community has also gotten used to an "algorithm-to-data" model in which participants submit their algorithm to an entity that runs and evaluates all systems for all tasks. In MIREX, this role is currently played by the IMIRSEL for the most part, although this year some tasks were decentralized and ran elsewhere. In our view, decentralization needs to go one step further, and try to follow a "data-to-algorithm" model as much as possible, where participants can run their systems on a publicly available data set, and then submit their raw output to a third party that scores the systems. This point is especially important because the current model places a very heavy burden on IMIRSEL in terms of workload and infrastructure.

Participation. MIREX is currently attracting fewer people than in the past (crisis effect?). One example is the Audio Cover Song Identification task, which had been very successful in the past but had no participants this year. However, there were some posters from MSD Challenge participants that actually tackled cover detection. This indicates that, despite the lack of participation in MIREX, there is still some interest in the task, so the question that followed is: why not let participants lead, organize and run the task by themselves? The next two points follow from this question.

Task leaders. During the session, a potential solution (proved to work efficiently in evaluation initiatives outside the MIR community) to the above-mentioned issues related to Methodology, Data, Model and Participation was discussed: decentralization of the evaluations through the creation of task-communities and task-leaders.

A task leader is a person who creates and animates a group of people interested in evaluating a task, defines the methodology for evaluation (finding or creating a suitable annotated MIR corpus, query set and relevance data as appropriate for the task, and selects or defines the performance evaluation metrics, animates discussions on the results obtained for a specific task). The task leader or coordinator essentially takes ownership of the running of the task, ensuring for example that instructions and data is made available according to an arranged schedule, answers questions from participants, and analyzes and collates submitted results. There should not be an evaluation task for which nobody leads the task, since definition of a task often requires considerable work and it is important that someone leads to establish consensus of the right way to structure and evaluate the task. Running a task that is poorly defined is dangerous considering the consequence of deriving conclusions from an ill defined evaluation. Even if the task is suitably defined, if there is no leader it is likely that it may not keep to its schedule or activities will be overlooked or not completed properly.

Some people questioned the personal value of being a task leader. It seems that people are afraid of the amount of extra work this involves, especially if it requires to create a new dataset from scratch. In that line, it was proposed to slightly increase the ISMIR registration fee and spend some funds every year for the incremental development of new datasets. The involvement of a tasks community can often help to solve this issue too. If someone is sufficiently passionate about the research questions involved in a task, they will often commit the effort to design the task and develop the required dataset, since this helps them to develop their own research; but also to encourage others to become involved and develop a community of like-minded researchers interested in this and related topics [4]. In fact, the majority of attendees were willing to volunteer as task leaders next year.

Facilitating the gathering and exchange of knowledge. Many people were also concerned with the traditional poster session held during the last ISMIR day, where participants show their approach for the various tasks and the various evaluation initiatives. The general feeling is that one poster session is not nearly enough for people to discuss results and task design, and wondered whether there should be one day solely devoted to this. For the time being, it was proposed to move this evaluation session to the very beginning of the conference, so researchers have the chance to discuss during the following days and exchange thoughts on task design.

3. CONCLUSION

The general outcome gathered from these two sessions is that the ISMIR community is really concerned about how we evaluate our systems. Individuals are often frustrated because the current evaluation practices we follow do not fully allow us to work and improve as much as we wish. However, despite acknowledging this, there seem to be some reluctance to get involved in a task-force devoted to improve our situation. On the other hand, various people vigorously showed their willingness to be part of such an endeavor, lead the evaluation of our tasks and commit to improve them. Our impression is therefore that the ISMIR community should encourage this type of research and the coordination of efforts for the common good.

4. ACKNOWLEDGMENTS

We gratefully thank the MIReS project (Roadmap for Music Information ReSearch) funded by EU-FP7-ICT-2011.1.5- 287711 for having funded and organized the Panel session on Evaluation Initiatives in MIR, the ISMIR organizers for hosting it, as well as all attendees in both the panel and the late-breaking sessions for participating in these interesting discussions.

[4] This is the case of the Audio Melody Extraction Annotation Initiative, which started from the discussions at ISMIR 2012. http://ameannotationinitiative.wikispaces.com/