WELL-DEFINED TASKS AND GOOD DATASETS FOR MIR
ISMIR 2013 LATE-BREAK

Eric Battenberg
Gracenote, Inc.
ebattenberg@gracenote.com

ABSTRACT

The late-breaking session on evaluation and datasets was well attended with many strong opinions presented. After significant debate, there were two main directions agreed upon by many of the attendees at the session. The first was the importance of providing a central web resource to help organize researchers in the their evaluation efforts. The second was the possibility of appointing a "data chair" to lead the creation and maintenance of such a resource.

1. INTRODUCTION

This is a writeup of the ISMIR 2013 late-breaking session on "Well-Defined Tasks and Good Datasets for Large-Scale MIR". This session was heavily attended, and many of the opinions and concerns expressed were quite passionate, indicating the great importance of this topic to the community as a whole. There was even remote participation via the online document used for note taking during the session.

Although this session was proposed with the hope of addressing the evaluation and data needs of deep-learning-based and large-scale MIR, there were a wide variety of concerns and suggestions proffered that seemed to span all aspects of MIR evaluation. This writeup cannot hope to list every point made by every attendee, so we will focus on the more fully formed suggestions that encompass the major themes of the session.

2. PROVIDE A CENTRAL RESOURCE

Many attendees agreed that maintaining a central list of MIR datasets and tasks would be of great benefit to the community. Such a resource could provide hosting for datasets, along with documentation and versioning of the data. The site could present evaluation details for each task, provide evaluation scripts in multiple languages, and/or host an automatic online evaluation service using a hidden test set.

Whatever information is presented on this web resource, it is most important that it is easy to access (i.e., people are aware of its location on the web) and that its data is persistent.

The following sections describe additional services that could be included on an MIR central web resource.

3. STANDARDIZED EVALUATION PACKAGE

One suggestion for a way to simplify and standardize MIR evaluation is to provide everything needed to evaluate a specific task (evaluation scripts, data, etc.) in a single package that can be cloned via something like GitHub or downloaded via BitTorrent. The package could serve as a completely automated solution with built-in error checking and documentation.

4. TASK SCOREBOARD

The idea to keep a "task scoreboard" was discussed at the session, but it has also been discussed in the past on the Music-IR email list. Each evaluation task and dataset should have its own scoreboard that gives baseline performance and a ranking of current state-of-the-art performance numbers along with links to associated implementations and write-ups. (Similar to that done for the MNIST dataset [1]).

A task scoreboard would make it very easy for researchers to determine the current state-of-the-art performance for each task and to compare how various techniques perform on the same task. It would also be interesting to keep a history of these performance rankings, so that it is easy to investigate which approaches dominated in the past.

5. CONTRIBUTION OF ISMIR

During the session, it was asked what ISMIR's role in this effort should be. While the MIREX initiative has been greatly helpful for organizing yearly evaluation competitions, it has remained somewhat isolated from the portion of the MIR community that doesn't participate in the competition. Many of the attendees agreed that ISMIR should focus significant effort on creating and maintaining the central web resource mentioned in Section 2.

To help spearhead this effort, a popular suggestion was to appoint a "data chair" who is allocated funding to maintain the above-mentioned central resource. Other responsibilities of the data chair could be to curate datasets of sufficient size and quality for the various tasks and to maintain the standardized evaluation packages and scripts.

[1] http://yann.lecun.com/exdb/mnist/

At the end of the session, there was widespread agreement that a central MIR evaluation and dataset resource should be created, and that the maintenance of such a resource should be led by a data chair.