ISMIR 2004 Audio Description Contest

Pedro Cano, Emilia Gómez, Fabien Gouyon, Perfecto Herrera,
Markus Koppenberger, Beesuan Ong, Xavier Serra,
Sebastian Streich, Nicolas Wack
Music Technology Group, Universitat Pompeu Fabra

MTG-TR-2006-02
April 6, 2006

Abstract: In this paper we report on the ISMIR 2004 Audio Description Contest. We first detail the contest organization, evaluation metrics, data and infrastructure. We then provide the details and results of each contest in turn. Published papers and algorithm source codes are given when originally available. We finally discuss some aspects of these contests and propose ways to organize future, improved, audio description contests.

ISMIR 2004 Audio Description Contest

Pedro Cano, Emilia Gómez, Fabien Gouyon, Perfecto Herrera, Markus Koppenberger,

Beesuan Ong, Xavier Serra, Sebastian Streich, Nicolas Wack

Universitat Pompeu Fabra, IUA, Music Technology Group

Ocata 1. 08003 Barcelona, Spain.

Abstract: In this paper we report on the ISMIR 2004 Audio Description Contest. We first detail the contest organization, evaluation metrics, data and infrastructure. We then provide the details and results of each contest in turn. Published papers and algorithm source codes are given when originally available. We finally discuss some aspects of these contests and propose ways to organize future, improved, audio description contests.

Music Information Retrieval (MIR) established itself in the last few years as a very active multidisciplinary research field. This is clearly shown in the constantly growing number and subjects of articles published in the Proceedings of the annual International Conference on Music Information Retrieval (ISMIR, the first established international scientific forum for researchers involved in MIR) and also in related conferences and scientific journals such as ACM Multimedia, IEEE ICME or Wedelmusic, to name a few. The standardization of world-wide low-latency networks, the extensive use of efficient search engines in everyday life, the continuously growing amount of multimedia information (on the web, in broadcast data streams or in personal and professional databases) and the rapid development of online music stores (as e.g. Apples iTunes, Walmart or MusicMatch) set great challenges to MIR researchers. Indeed, applications are manifold, from automated music analysis to personalized music recommendation, online music access, query-based retrieval (e.g. “by-humming”, “by-example”) and automatic playlist generation.

Among the vast number of disciplines and approaches to MIR (an overview of which can be found in (Downie 2003)), automatic description of audio signals in terms of musically-meaningful concepts plays an important role. As in other scientific endeavor, long-term improvements are

bounded to systematic evaluation of models. For instance, text retrieval techniques significantly improved over the year thanks to the TREC initiative (see trec.nist.org) and the standardization of databases and evaluation metrics greatly facilitated progress in the fields of Speech Recognition (Przybocki et al. 1998; Pearce et al. 2000), Machine Learning (Guyon et al. 2004) or Video Retrieval (see http://www-nlpir.nist.gov/projects/trecvid/). Systematic evaluations permit to measure but also to guide progresses in a specific field. Since a few years, the MIR community has recognized the necessity to conduct rigorous and comprehensive evaluations (Downie 2002; Downie 2004, Berenzweig et al. 2004). An Audio Description Contest took place during the 5th edition of the ISMIR in Barcelona, Spain, in October 2004. The goal of this Contest was to compare state-of-the-art audio algorithms and systems relevant for MIR. It represents the first world-wide competition on algorithms for audio description for MIR.

The original idea to organise such an event emerged from the research infrastructure in place at the Music Technology Group of the Pompeu Fabra University (who hosted ISMIR 2004) where around 50 researchers work on tasks related to musical audio analysis and synthesis (audio fingerprinting, singing voice synthesis, music content processing, etc., see www.iua.upf.es/mtg). A massive storage and computer cluster facility hosts a common repository of audio data and provides computing functionalities, thus permitting evaluation of developed algorithms (Cano et al. 2004). Several audio description tasks were proposed to the MIR community in advance and the contest organizers gave full support for other potential tasks that would emerge from the community. Participation was open and all aspects of the several contests (data, evaluation methods, etc.) were publicly discussed and agreed. Finally, a total of 20 participants (from 12 research laboratories) took part in one or several of the following tasks: Melody extraction, Tempo induction, Genre Identification, Artist Identification and Rhythm classification.

In the remainder of this paper we report on the ISMIR 2004 Audio Description Contest (see also ismir2004.ismir.net/ISMIR_Contest.html). We first detail the contest organization, evaluation metrics, data and infrastructure. We then provide the details and results of each contest in turn. Published papers and algorithm source codes are given when originally available. We finally discuss

some aspects of these contests and propose ways to organize future, improved, audio description contests.

Contest organization

Contests

Ten different contests were initially proposed: audio fingerprinting, music genre classification, musical instrument classification, artist identification/similarity, melody extraction, rhythm classification, tempo induction, key and chord extraction, music structure analysis and chorus detection.

After public discussions and demonstrations of interest, 5 tasks were finally selected: melody extraction (4 participants), artist identification (2 participants), rhythm classification (1 participant), music genre classification (5 participants) and tempo induction (6 participants).

UPF researchers, as organizers, did not take part in the competitions.

Evaluation metrics

A great deal of concern was dedicated to the design of general, unbiased evaluation functions as well as methods that assess the statistical significance of the different results. This was done through consultation with contest participants as they have the expertise for evaluating their particular problem. Participants contributed to metrics design at diverse degrees and all agreed with the final metrics.

Data and annotations

The contest organizers provided the data used for the contests (although participants were initially allowed, and encouraged, to contribute to the extension of the contest data). Depending on the specific contest, two types of (copyright-free) data was distributed to participants: training data and preparatory data. The former were provided to disambiguate target concepts and so that participants' models would fit relatively well the data they would be tested upon. The latter was provided in order

to compare whether algorithms yielded the same output when run in participants' labs and on the organizers' machines, and to check proper formatting of algorithm input and output.

Annotations were either put manually or were legacy metadata.

Depending on the specific contest, algorithms were also tested against a set of distortions, for instance equalization, MP3 or GSM encoding/decoding, resampling, noise addition, cell-phone effect and the like.

Test data was not given to participants before the final evaluation of the algorithms. After the results publication, great part of the data (as far as copyright licenses allowed) has been made public in order to stimulate further research (see the contest webpage). Metadata needed to univocally identify copyrighted content is now public on the contest webpage.

Infrastructure

Algorithm evaluations took place at UPF labs before the conference, and were communicated to participants before being made public during the ISMIR conference.

Depending on the participants, contest entries were made in different formats: source code (C++, Matlab), binaries (Windows, Linux), Matlab pre-parsed pseudo-code files (p-files, source code is not visible). The use of external frameworks, such as HTK or Weka was also granted.

Submissions were made via an email interface. Anonymous participation was allowed. Deletion or update of a submission was possible until the final submission deadline.

To avoid conflicts with non-disclosure agreements or the like, it was not mandatory for the participants to make their algorithms public.

Genre classification

The goal of the contest was to classify a set of given songs into their respective musical genres.

Although it is probably one of the most commonly used criteria for distinguishing songs, the concept of musical genre is quite controversial. Musical genres taxonomies as found in the music

industry are inconsistent (Pachet et al. 2000). Nevertheless, music consumers seem to agree at least on a rough level of music classification into a few broad Genres. Consider for example `Rock' and `Classical' music, people may disagree on a precise definition for each, but they would most likely not mix them up. Furthermore, Genre classification has been a very attractive topic to MIR researchers and the literature related to computational models for Genre classes is very rich (Tzanetakis et al. 2002). Finally, and most importantly, real needs for automatic genre classification given existing taxonomies do exist. Magnatune (see www.magnatune.com) is an example of label that could greatly benefit from systems that classify new songs according to their taxonomy.

Data Sets

Three distinct sets of songs were built for the contest: two of them (training and development set) were distributed to the participants for testing their algorithms, a third one (evaluation set) was kept undisclosed for the final evaluation. The training set was intended to train the models that would later be used for evaluation and the development set allowed the participants to test their models on a different set of songs than the training one. The evaluation set was similar in composition to the training and development set, so that final results could be anticipated.

All the songs were courtesy from the Magnatune website where they are available under a Creative Commons (see creativecommons.org) license for "non-commercial use." The ground truth corresponds to the following genres (as found on Magnatune's main webpage): Classical, Electronic, Jazz/Blues, Metal/Punk, Rock/Pop and World. No attempt was made to subsequently change potentially arguable or misleading song tags.

Each of the sets contained 729 tracks, with a quantity of tracks in each genre roughly proportional to the ones we can find on the Magnatune website. The training set had the following distribution:

Genre	**Number of tracks**
Classical	320

Genre	Number of tracks
Electronic	115
Jazz/Blues	26
Metal/Punk	45
Rock/Pop	101
World	122

Each set had approximately the same distribution and no song from a set could be found in another one. Data and legacy metadata can be found on the contest web page.

Framework

Participants had to submit their system in three parts in order to comply with the framework:

The *DescriptorExtractor*: given an audio file (WAV format, 22.05KHz, mono) it extracts its features and write them in a file.

The *TrainModel*: given a list of feature files extracted by DescriptorExtractor, and their associated classes, it learns the concepts and stores them in a file for the EvaluateModel.

The *EvaluateModel*: given a model and a test instance in the form of feature file, it predicts the corresponding genre.

Evaluation Metrics

The criteria for evaluation were the total number of correctly identified songs, and the percentage of correctly identified songs normalized with respect to the probability of each class (which had distributions proportional to the ones on the Magnatune website).

We also used the McNemar test (Gillick et al. 1989) to assess the statistical significance of the results.

Participants

Five teams participated: Thomas Lidy and Andreas Rauber from the Vienna University of Technology (Lidy et al. 2003), Dan Ellis and Brian Whitman respectively from Columbia University and the MIT, Kris West of the University of East Anglia (West et al. 2004), Elias Pampalk from the Austrian Research Institute for Artificial Intelligence (Pampalk 2004), and George Tzanetakis form the University of Victoria (file ismir_genre.tar.bz2 in the marsyas-0.2 distribution, available as GPL under the sourceforge.net/ web)

Results

Elias Pampalk scored 84.1% correct identifications (78.8% normalized), Kris West scored 78.9% (69.9% normalized), George Tzanetakis scored 71.3% (58.7% normalized), Thomas Lidy and Andreas Rauber scored 70.6% (58.7% normalized) and Dan Ellis and Brian Whitman scored 63.3% (51.3% normalized).

Artist identification

The artist identification contest was very similar to the genre classification contest. The goal was, given a song, to identify which artist (performer) was performing it. In this case, we refer to artist as the singer, if alone, or the band who performed the song. We do not take into account the writer or the author of the songs.

The same framework and evaluation metrics were used as for the Genre classification contest. This time, however, the number of classes (artists) was much more important, and the number of songs in each class much smaller.

Data sets

Three sets were compiled as well, two of them given to the participants, and the third one

undisclosed.

The composition of the sets was the following: the training set was composed of 105 artists, with exactly 7 songs for each one of them, for a total of 735 songs. The development and evaluation set were composed of the same 105 artists, with exactly 3 songs for each one of them. Due to technical problems, we ran the contest using two different subsets of the original set. The first one was composed of 30 artists, the second one of 40 artists, each of them keeping the original number of tracks per artist. Data and legacy metadata can be found on the contest web page.

Evaluation metrics

The metrics corresponds to the percentage of correct artist identifications.

Participants

Two entries were made for this contest. Thomas Lidy and Andreas Rauber from Vienna University of Technology submitted an algorithm (Lidy et al. 2003). Dan Ellis and Brian Whitman respectively from Columbia University and the MIT also submitted one.

Results

On the 30-artist set, Thomas Lidy and Andreas Rauber scored 28% correct identifications, Dan Ellis and Brian Whitman scored 34%.

On the 40-artist set, both groups scored 24% correct identifications.

Melody extraction

The purpose of this contest was to evaluate and compare state-of-the-art algorithms for melody detection within polyphonic audio. This evaluation was made for audio recordings containing singing voice and solo instruments. It was centered in evaluating different approaches for the automatic extraction of the predominant pitch and the melody.

Melody plays a major role for music description. Sometimes it happens that we are looking for a song and we remember neither its title nor its author, but only its chorus's main melody. There are

many aspects to consider when defining and representing melody, and many approaches to extract melodic features from audio, as reviewed in (Gómez et al. 2003).

Fundamental frequency is the main low-level descriptor to be considered when describing melody. Due to the significance of pitch detection for speech and music analysis, a lot of research has been made in this field. It is commonly agreed that pitch is a key feature defining melody. Although there has been much research devoted to pitch estimation (as reviewed in (Gómez et al. 2003; Klapuri 2004), it is still an unsolved problem even for monophonic signals. When dealing with polyphonic audio, there are two issues to handle: the identification of the voice that defines the melody within the polyphony, and the estimation of the fundamental frequency of this predominant voice. State-of-the-art multipitch estimators operate reasonably accurately for clean signals, frame-level error rates progressively increasing with polyphony. However, the performance decreases significantly in the presence of noise, and the number of concurrent voices is often underestimated.

Melody is not just defined by fundamental frequency information. It has been commonly represented as a succession of pitched notes. This melodic representation accounts for rhythmic information as inherently linked to melody. Once the fundamental frequency has been estimated, we need to delimitate the note boundaries, in order to extract a sequence of notes and the descriptors associated to note segments.

Data Sets

The material used for the evaluation was selected from a variety of styles. The goal was to identify the algorithm that works best as a general purpose melody extractor. So we compiled a set of musical excerpts that would face the algorithms with different types of difficulties. We chose a total of 20 musical excerpts with around 20 seconds of duration each, from the following categories: MIDI synthesized polyphonic sound with a predominant voice (4 items), saxophone melodic phrases plus background music (4 items), synthesized singing voice plus background music (4 items), pop music with singing voice (4 items), male opera singing with orchestra (2 items) and female opera singing with orchestra (2 items).

Part of the evaluation material was available to the participants in advance, together with its annotations. For reasons of consistency we combined half of the items from each category in this tuning set. All sound files were provided as 16 bit, single channel PCM data at 44.1kHz sampling rate. Data and legacy metadata can be found on the contest web page.

Evaluation metrics

Three evaluation metrics were used:

Fundamental frequency

Metrics 1 consisted on a frame-based comparison of estimated F0 and reference F0 on logarithmic scale. The reference was obtained by analyzing the isolated leading voice followed by some manual corrections. A value of 0 Hz was assigned to unpitched frames. The concordance was measured as the average absolute difference with a threshold of 1 semitone (= 100 cents) for the maximal error. Each frame contributed to the final result with the same weight.

Fundamental frequency regardless octave errors

Metrics 2 was basically the same as metrics 1. The distinction was that before computing the absolute difference, the values for F0 were mapped into the range of one octave. Octave errors, which are a very common problem in F0 estimation, were disregarded this way.

It should be stated that these two metrics operate in a domain which is still close to the signal and not yet as abstract as a transcribed melody. In other words, this is only an intermediate step towards a real extraction of the melody from polyphonic recordings.

Melodic similarity

Metrics 3 was the edit distance between the estimated melody and the correct melody. The correct melody was obtained by manual score alignment. Compared to the other two metrics the abstraction level here is clearly higher, because note segmentation is required. Especially for sung melodies this is non-trivial, because of vibratos and strongly varying pitches.

The edit distance metric calculates the cost for transformation of one melody to another one. Different weights can be assigned to different transformation operations (insertion, deletion or shifting). More information on the edit distance and its implementation is found in (Gratchen et al. 2002).

List of participants

Four approaches were evaluated in this contest: algorithm 1 by Rui Pedro Paiva from the University of Coimbra (Paiva et al. 2004), algorithm 2 by Sven Tappert from Berlin Technical University (Batke et al. 2004), algorithm 3 by Graham Poliner and Dan Ellis from Columbia University and algorithm 4 by Juan P. Bello from Queen Mary University of London.

Results

The results of the evaluation according to metrics 1 and 2 were as follows: Rui Pedro Paiva 64.74% and 65.20%, Sven Tappert 42.19% and 55.88%, Graham Poliner 56.14% and 57.14%, and Juan P. Bello 50.85% and 57.70%. The values reflect the average correctness of predominant pitch estimation over the entire dataset (i.e. tuning and test set). A monophonic pitch tracker developed in the context of the SMSTools (implemented at the UPF) was used as a baseline and yielded 32.75% and 42.23% (Gómez et al. 2003).

For metrics 3 only two of the participants submitted algorithms with the required output. Rui Pedro Paiva's algorithm achieved an average edit distance of 8.63, Juan P. Bello's algorithm reached 14.12 on average. Again, results were computed on the entire dataset.

Evaluation results show that the approach that performs the best for any of the metrics is the approach 1 by Rui Pedro Paiva.

Computation time

We also computed an estimation of the computation time for each of the algorithms. This gives an idea of the performance of the different methods, although it was not taken into account for identifying the winner of the contest. Algorithms were computed in two machines: Windows PC

Pentium 1.2 GHz, 1 Gb RAM and Linux PC Pentium 2 GHz, 500 Mb RAM. Results are presented in the following table:

ParticipantID	1	2	3	4
Operating system	Windows	Linux (MATLAB)	Linux	Linux (MATLAB)
Average time per audio excerpt (in seconds)	3346,67	60,00	470,00	82,50

This estimation shows that the fastest algorithms were algorithms 2 and 4. Algorithm 1 is by far the slowest.

Tempo induction

The goal of this contest was to evaluate state-of-the-art algorithms in the task of computing the basic tempo: the rate of musical beats in time, expressed in BPM. Much effort in the computer music community has been dedicated to this task (see (Gouyon et al. 2005) for a review). However, little attention has been dedicated to computational model evaluations. Early models usually did not present quantitative evaluation of the proposed models, and only recently have researchers begun to report on the performance of their systems, but they meet with several difficulties: the lack of publicly available annotated data sets and the lack of agreed evaluation metrics.

Data Sets

No training data was provided. However, 7 preparatory instances were given.

The test data consisted of 3199 tempo-annotated instances in 3 data sets. The instances range from 2 to 30 seconds, and from 24 BPM to 242 BPM (note however that only 15 excerpts have a tempo less than 50 BPM). They all have approximately constant tempi, and the format is the same for all: mono, linear PCM, 44100 Hz sampling frequency, 16 bit resolution. The total duration of the test set is approximately 45140 s (i.e. around 12 h 36 min). This data was not available to participants before the competition.

The first data set is made up of 2036 “Loops.” They are short drum, bass or synthesizer excerpts produced to include in DJ sessions, or for home recording needs. They all last a few bars and the total duration of this set is around 15170 s. The tempos range between 60 and 215 BPM. The genres are Electronic, Rock, House, Ambient and Techno.

The second data set is made up of 698 typical excerpts of Ballroom dance music downloaded from the web (see www.ballroomdancers.com). The durations are around 30 s and the total duration is around 20940 s. The tempos range between 60 and 224 BPM. The genres are Cha Cha, Jive, Quickstep, Rumba, Samba, Tango, Viennese Waltz and Slow Waltz.

The third data set is made up of 465 song excerpts whose durations are around 20 s. The total duration is around 9300 s. The tempos range between 24 and 242 BPM. The genres are Rock, Classical, Electronic, Latin, Samba, Jazz, Afro-beat, Flamenco, Balkan and Greek.

Part of the data and legacy metadata can be found on the contest web page.

Evaluation metrics

Two evaluation metrics were used for this contest. Accuracy 1 was computed as the percentage of tempo estimates within 4% (the precision window) of the ground-truth tempo. Accuracy 2 was computed as the percentage of tempo estimates within 4% of either the ground-truth tempo, or half, double, three times, or one third of the ground-truth tempo.

Participants

12 algorithms entered the contest, 11 were submitted by 6 different research teams, and one open-source algorithm (GPL-licensed) was downloaded from the web. The algorithms are listed in alphabetical order: AlonsoACF and AlonsoSP, submitted by Miguel Alonso from the ENST in Paris (see Alonso et al. 2004); DixonACF (Dixon et al. 2003), DixonI and DixonT (Dixon 2001) submitted by Simon Dixon from the ÖFAI in Vienna; Klapuri, submitted by Anssi Klapuri from the Tampere University of Technology (Klapuri et al. 2005); Scheirer (source code of Eric Scheirer’s algorithm (Scheirer 1998) downloaded from the following web: sound.media.mit.edu/~eds/beat/tapping.tar.gz);

TzanetakisH, TzanetakisMM and TzanetakisMS, submitted by George Tzanetakis from Victoria University (Tzanetakis et al. 2002, marsyas-0.2 under www.sourceforge.net/projects/marsyas); Uhle, submitted by Christian Uhle from Fraunhofer Institute for Digital Media Technology (Uhle et al. 2004); and finally Anonymous, submitted by an anonymous participant.

Results

With respect to Accuracy 1, computed on the whole data set of 3199 instances, Klapuri reached 67.29% accuracy, Uhle 51.61%, Anonymous 45.26%, DixonACF 38.82%, Scheirer 37.85%, AlonsoSP 36.29%, DixonI 31.76%, TzanetakisMS 31.22%, TzanetakisMM 30.76%, AlonsoACF 27.78%, DixonT 26.56% and TzanetakisH 25.22%.

With respect to Accuracy 2, also computed on the whole data set, Klapuri reached 85.01% accuracy, DixonACF 82.3%, Anonymous 81.21%, Uhle 76.11%, DixonT 74.3%, DixonI 73.6%, AlonsoSP 69.77%, Scheirer 68.08%, AlonsoACF 57.89%, TzanetakisMS 55.51%, TzanetakisH 54.67% and TzanetakisMS 50.73%.

More details are given on the contest webpage.

Rhythm classification

The goal was to compare algorithms for automatic classification of 8 rhythm classes (Samba, Slow Waltz, Viennese Waltz, Tango, Cha Cha, Rumba, Jive, Quickstep) from audio data.

Evaluation material

This is the same data as the second data set used for the tempo induction contest (698 30-s instances corresponding to the 8 rhythmic classes, .wav format, 44100 Hz, 16 bits, mono), but divided into a training set and a test set. Participants were given a list of 488 training instances in order to train their systems (see www.iua.upf.es/mtg/ismir2004/contest/rhythmContest/TrainingDataURLs.txt). These files could be downloaded from the web (see www.ballroomdancers.com/Music/style.asp).

A test dataset of 210 instances (same format, coming from the same source) was used to evaluate the

submitted algorithm (see www.iua.upf.es/mtg/ismir2004/contest/rhythmContest/TestData.txt)

Data and legacy metadata can be found on the contest web page.

Evaluation metrics

The evaluation metric was the percentage of correctly classified instances.

Participants

A single algorithm has been submitted to the rhythm classification contest, it was submitted by Thomas Lidy and Andreas Rauber from the Vienna University of Technology and Andreas Pesenhofer from eCommerce Competence Center in Vienna, Austria (Lidy et al. 2003).

Results

The algorithm accuracy was of 82%.

Similar accuracy figures were reported on the same data (but with different evaluation procedures) in (Dixon et al., 2004).

Discussion

A public panel was organized during the ISMIR conference to make the above results public and to stimulate a discussion between contest participants, organizers and conference attendees regarding the past contest and the general issue of large-scale systematic evaluation. A clear conclusion emerged: public competitions will greatly facilitate progress in automatic audio description research (as well as other parts of MIR). ISMIR 2005 (and probably many, if not all, future editions of the conference) will also feature an evaluation of algorithms: the Music Information Retrieval Evaluation eXchange (MIREX, see ismir2005.ismir.net). Future evaluations will cover an even broader range of tasks than those described in this paper, including symbolic data description and retrieval. In the following we describe several issues that rose during the organization of this year's contest and were discussed during the panel.

Encouraging participation

Even if the participation of 12 research groups made the contest successful, it would be better if future evaluations would attract more researchers, ideally the majority of researchers working in MIR. It seems that performance uncertainty and fear of failure acted as barriers to participation. Hopefully, these factors should only play a role for the first time. By repeating evaluation tasks over public data, potential participants can test their systems and get an idea of how their approaches compare with the state-of-the-art, and then they have the option of not participating if their results are clearly inferior. (They may of course still participate even with worse results; negative results are as valuable to science as good ones.)

Contest organizers could also set up a "validation" data set in addition to training and test data (Guyon et al. 2004). The content of this data set would be very similar to that of the final test set, at the difference with the training set, it would not be provided to participants. This data would serve to evaluate algorithms just before the submission deadline (e.g. one or two weeks before), so that all potential participants can compare their algorithms to others' and proceed to final fine-tuning or even withdraw from participation.

Higher participation rates would certainly be achieved by providing more time for preparation and submission. This year's contest schedule seemed too tight to many researchers.

Another big incentive to participation can be availability of ground-truth data. Setting up test data could be the outcome of the contributions of all contest participants. This data could subsequently be made public with some delay (e.g. one or two years). This way, each participant would contribute only a small part and would benefit from much more "fresh new" data before many other researchers in the community.

Modular evaluations

Even if it is useful to know which system performs best on a specific task, it is also clear that much more insights would be gained by knowing which parts of this system are especially valuable. It would therefore be of greater interest to evaluate different system modules rather than whole

systems taken as black boxes.

A solution could be to motivate participants to submit several systems, with small, but conceptually relevant, variations in some submodules.

Data and metadata gathering

The Creative Commons licensing schemes (CC, see creativecommons.org/) have made things easier for an activity such as the one we have reported here. Under CC non-commercial license it is possible to grant access to musical files for research purposes and, included in the license, there is the possibility to copy and redistribute the files to other partners. Several music distributors (see creativecommons.org/audio/) offer their music under CC and therefore they are potential catalyzers of our research. As we have started benefiting from that, we should start helping them to improve their business with the help of our technologies.

Another valuable effort from inside the MIR/MDL community is the RCW database (see staff.aist.go.jp/m.goto/RWC-MDB/ and Goto 2004). This database is a copyright-cleared music collection compiled specifically for research purposes and including audio and midi files spanning across several music genres. It has some manufacturing costs associated to its acquisition, but it is free of royalty payments and researchers can freely use it for publications and presentations.

Acknowledgements

Many researchers in the MIR community should be thanked for helping this contest happen. Special thanks go to Dan Ellis for his contribution to the organization and for gathering conclusions and ideas that emerged during the panel. We also would like to thank the panelists: Juan Bello, Stephen Downie, Dan Ellis, Marc Leman, Elias Pampalk and George Tzanetakis. All contest participants are also warmly thanked. Additional thanks go to Maarten Grachten for his contribution on the algorithm for computing melodic similarity, and people from the MTG for their contributions to the evaluation material (Jordi Bonada, Lars Fabig, Alex Loscos and Oscar Mayor). We also would like to thank Kris West and Stephen Cox for handing us out the Matlab code for computing the results of the McNemar

test.

References

Alonso, M., David, B., Richard, G., 2004, “Tempo and Beat Estimation of Musical Signals”, Proc. ISMIR 2004, pages 158-163

Batke, J., Eisenberg, G., Weishaupt, P., Sikora., T., 2004. “A query by humming system using MPEG-7 descriptors”, Proc. of the 116th AES Convention, 2004

Berenzweig, A., Logan, B., Ellis, D., Whitman, B., 2004 “A large-scale evaluation of acoustic and subjective music-similarity measures” Computer Music Journal 28(2):63-76, 2004

Cano, P. Koppenberger, M. Ferradans, S. Martinez, A. Gouyon, F. Sandvold, V. Tarasov, V., and Wack, N. 2004. 'MTG-DB: A Repository for Music Audio Processing' Proceedings of 4th International Conference on Web Delivering of Music Barcelona, Spain

Cano, P. Batlle, E. Kalker, T. and Haitsma, J. 2002. 'A Review of Algorithms for Audio Fingerprinting' Proceedings of 2002 International Workshop on Multimedia Signal Processing St. Thomas, Virgin Islands

Dixon, S., 2001, “Automatic Extraction of Tempo and Beat From Expressive Performances”, Journal of New Music Research 30(1):39-58

Dixon, S. Gouyon, F. Widmer, G. 2004. ′Towards Characterisation of Music via Rhythmic Patterns′ Proceedings of Fifth International Conference on Music Information Retrieval Barcelona

Dixon, S., Pampalk, E., Widmer, G., “Classification of Dance Music by Periodicity Patterns”, Proc. ISMIR 2003, pages 159-165

Downie, J. (Editor) 2002 “The MIR/MDL evaluation project white paper collection” 2nd edition, Proc. ISMIR 2002

Downie, J. 2003 “Music Information Retrieval” Annual Review of Information Science and Technology 37:295-340

Downie, J. 2004 “The scientific evaluation of music information retrieval systems: Foundations and future” Computer Music Journal 28(2):12-33.

Gillick L., Cox, S. “Some statistical issues in the comparison of speech recognition algorithms” Proc. IEEE Conference on Acoustics, Speech and Signal Processing, 1989, pp.532–535.

Gómez, E., Klapuri, A., and Meudic, B., 2003. Melody Description and Extraction in the Context of Music Content Processing. Journal of New Music Research Vol.32.1

Goto, M., 2004,“Development of the RWC Music Database”, Proc. 18th International Congress on Acoustics (ICA 2004), pp.I-553-556, 2004

Gouyon, F., and Dixon, S., “A review of automatic rhythm description systems,” Computer Music Journal, vol. 29, no. 1, pp. 34–54, 2005.

Guyon, I., Gunn, S., Ben Hur, A., Dror, G. 2004. Result Analysis of the NIPS 2003 Feature Selection Challenge. Proc. NIPS.

Grachten, M., Arcos, J. Ll., and López de Mántaras., R., 2002. A Comparison of Different Approaches to Melodic Similarity. ICMAI02.

http://www.iiia.csic.es/~maarten/articles/MelSim.pdf

ISMIR2004 Melody Extraction Contest Definition Page http://ismir2004.ismir.net/melody_contest/results.html

Klapuri, A., 2004 “Signal Processing methods for the automatic transcription of music”, PhD Thesis, Tampere University of Technology.

Klapuri, A., Eronen, A., and Astola, J., “Analysis of the meter of acoustic musical signals,” IEEE Trans. Speech and Audio Processing, 2005.

Lidy, T., Rauber, A., 2003, “Genre-oriented Organization of Music Collections using the SOMeJB System: An Analysis of Rhythm Patterns and Other Features”,

Proc. DELOS Workshop on Multimedia Contents in Digital Libraries, 2003

Pachet, F., Cazaly, D., “A taxonomy of musical genres” 2000. Proc. Of Content-Based Multimedia Information Access Conference

Paiva, R., Mendes, T., Cardoso, A., 2004. “A Methodology for Detection of Melody in Polyphonic Musical Signals”, Proc. of the 116th AES Convention, 2004

Pampalk, E., 2004 “A Matlab Toolbox to Compute Music Similarity from Audio”, Proc. ISMIR, pages 254-257, 2004

Pearce, D., Hirsch, H., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proc. ICSLP

Przybocki, M., Martin, A. 1989, “NIST speaker recognition evaluations”, Proc. International Conference on Language Resources and Evaluations, pp.331-335, 1989

Scheirer, E., 1998, “Tempo and Beat Analysis of Acoustic Musical Signals”, J. Acoust. Soc. Am. 103(1):588-601

Tzanetakis, G., Cook, P. “Musical genre classification of audio signals” IEEE Trans. Speech and Audio Processing, 10(5):293–302, 2002.

Uhle, C., Rohden, J., Cremer, M., Herre, J., 2004, “Low Complexity Musical Meter Estimation from Polyphonic Music”, Proc. AES 25th International Conference, pages 63-68

West, K., Cox, S., 2004, “Features and Classifiers for the automatic classification of musical audio signals”, Proc. ISMIR, pages 531-536, 2004