THE 2006 MUSIC INFORMATION RETRIEVAL EVALUATION EXCHANGE (MIREX 2006) RESULTS OVERVIEW

The IMIRSEL Group led by J. Stephen Downie
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

MIREX Results

Audio Onset Detection

Participant	Avg. F-Measure
A. Röbel (3)	0.788
A. Röbel (2)	0.780
A. Röbel (1)	0.777
Du, Li & Liu	0.762
P. Brossier (hfc)	0.734
S. Dixon (sf)	0.726
P. Brossier (dual)	0.724
P. Brossier (complex)	0.721
S. Dixon (rcd)	0.716
S. Dixon (cd)	0.710
P. Brossier (specdiff)	0.707
S. Dixon (wpd)	0.685
S. Dixon (nwpd)	0.620

Audio Tempo Extraction

Participant	P-Score
A. Klapuri	0.806
Davies & Plumbley	0.776
Alonso, David & Richard (2)	0.724
Alonso, David & Richard (1)	0.693
D. P. W. Ellis	0.673
Antonopoulos, Pikrakis & Theodoridis	0.669
P. Brossier	0.628

Audio Cover Song Identification

Participant	Avg. Perf.
D. P. W. Ellis	2.306
K. Lee (1)	1.106
K. Lee (2)	0.951
Sailer & Dressler	0.639
Lidy & Rauber	0.451
K. West (likely)	0.354
T. Pohle	0.351
K. West (trans)	0.309

Audio Beat Tracking

Participant	Avg. P-Score
S. Dixon	0.407
D. P. W. Ellis	0.401
A. Klapuri	0.395
Davies & Plumbley	0.394
P. Brossier	0.391

Audio Melody Extraction (ADC 2004 Dataset)

Participant	Overall Accuracy
K. Dressler	82.50%
Ryynänen & Klapuri	77.30%
Poliner & Ellis	71.90%
Sutton, Vincent, Plumbley & Bello	58.20%
P. Brossier	49.60%

Audio Melody Extraction (MIREX 2005 Dataset)

Participant	Overall Accuracy
K. Dressler	73.20%
Ryynänen & Klapuri	67.90%
Poliner & Ellis	63.00%
Sutton, Vincent, Plumbley & Bello	53.70%
P. Brossier	31.90%

Audio Music Similarity and Retrieval

Participant	Avg. Fine Score
E. Pampalk	0.430
T. Pohle	0.423
V. Soares	0.404
Lidy & Rauber	0.393
K. West (trans)	0.372
K. West (likely)	0.339

Symbolic Melodic Similarity (RISM Collection)

Participant	ADR	Binary*
Typke, Wiering & Veltkamp	0.715	0.733
Ferraro & Hanna	0.707	0.744
Frieler & Müllensiefen (2)	0.670	0.711
A. L. Uitdenbogerd	0.577	0.717
Frieler & Müllensiefen (3)	0.555	0.650
Lemström, Mikkilä, Mäkinen & Ukkonen (2)	0.541	0.422
Lemström, Mikkilä, Mäkinen & Ukkonen (1)	0.268	0.283
Frieler & Müllensiefen (1)	0.000	0.333

Symbolic Melodic Similarity (Karaoke Collection)

Participant	ADR	Binary*
Typke, Wiering & Veltkamp	0.819	0.440
A. L. Uitdenbogerd	0.378	0.407
Ferraro & Hanna	0.150	0.347
Lemström, Mikkilä, Mäkinen & Ukkonen (2)	0.000	0.293
Lemström, Mikkilä, Mäkinen & Ukkonen (1)	0.000	0.233

Symbolic Melodic Similarity (Mixed Polyphonic Collection)

Participant	ADR	Binary*
Typke, Wiering & Veltkamp	0.784	0.833
A. L. Uitdenbogerd	0.587	0.628
Ferraro & Hanna	0.218	0.483
Lemström, Mikkilä, Mäkinen & Ukkonen (1)	0.070	0.267
Lemström, Mikkilä, Mäkinen & Ukkonen (2)	0.000	0.272

❖ *Binary*: Binary Relevance Judgment (Not Similar=0, Somewhat Similar=1, Very Similar=1)*

Query-by-Singing/Humming (Task 1)

Participant	MRR
Wu & Li (1)	0.926
Wu & Li (2)	0.900
Jang & Lee	0.883
López & Rocamora	0.800
Lemström, Mikkilä, Mäkinen & Ukkonen	0.688
C. Sailer (ear)	0.568
Typke, Wiering & Veltkamp (2)	0.390
C. Sailer (warp)	0.348
A. L. Uitdenbogerd (2)	0.288
C. Sailer (midi)	0.283
Ferraro & Hanna	0.218
A. L. Uitdenbogerd (1)	0.205
Typke, Wiering & Veltkamp (1)	0.196

Query-by-Singing/Humming (Task 2)

Participant	Mean Precision
Jang & Lee	0.926
Lemström, Mikkilä, Mäkinen & Ukkonen	0.722
C. Sailer (midi)	0.649
C. Sailer (ear)	0.587
Typke, Wiering & Veltkamp (1)	0.468
C. Sailer (warp)	0.415
Typke, Wiering & Veltkamp (2)	0.401
Ferraro & Hanna	0.309
A. L. Uitdenbogerd (2)	0.238
A. L. Uitdenbogerd (1)	0.163

Score Following

Participant	Total Precision
Cont & Schwarz	82.90%
M. Puckette	29.75%

MIREX 2006 Challenges

◆ Data quality issues with regard to the selection of test sets and the identification of problematic formats and non-representative content

◆ Basic content-management issues with regard to multiple formats and ever growing result set sizes

◆ Difficulties performing sanity-check verifications of result sets as some individual returns are now approaching 0.5 Gigabytes

◆ Scaling issues between the "at home" testing of submissions and the formal "in lab" running of the task

◆Balancing the need for depth and breadth in evaluation data with the burden placed on the voluntary human evaluators

Key Issues for Future MIREX

◆ Continue to explore more statistical significance testing procedures beyond Friedman's ANOVA test

◆ Continue refinement of Evalutron 6000 technology and procedures

◆ Continue to establish a more formal organizational structure for future MIREX contests

◆ Continue to develop the evaluator software and establish an open-source evaluation API

◆ Make useful evaluation data publicly available year round

◆ Establish a webservices-based IMIRSEL/M2K online system prototype

Special Thanks to: *The Andrew W. Mellon Foundation, the National Science Foundation (Grant No. NSF IIS-0327371), the content providers and the MIR community, the Automated Learning Group (ALG) at the National Center for Supercomputing Applications at UIUC, Paul Lamere of Sun Labs, all of the IMIRSEL group – Mert Bay, Andreas Ehmann, Joe Futrelle, Anatoliy Gruzd, Xiao Hu, M. Cameron Jones, Jin Ha (Gina) Lee, Martin McCrory, David Tcheng, Qin Wei, Kris West, Byounggil Yoo, and all the volunteers who evaluated the MIREX results, the GSLIS technology services team, and the ISMIR 2006 organizing committee*