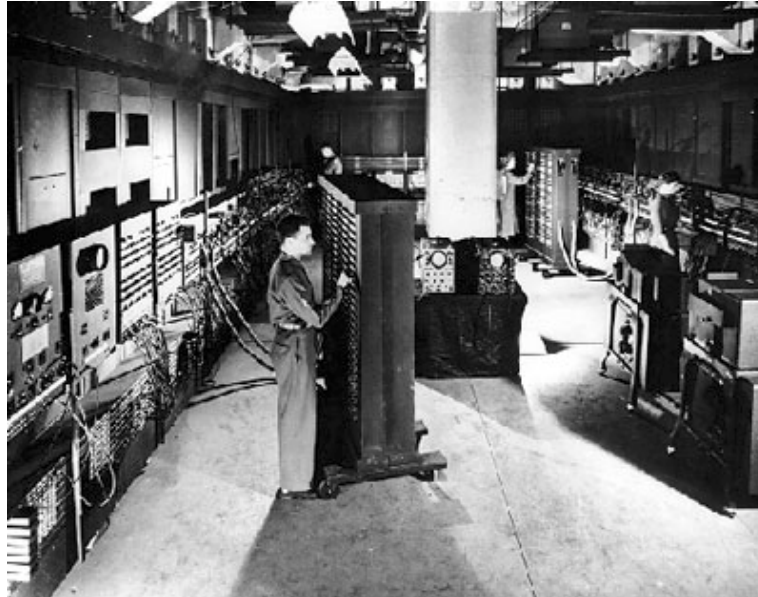


# ISMIR2004

## Panel 2: Discussion on the Audio Description Contest



# OUTLINE

- ❑ Panelists
- ❑ Contests (1 hour)
  - ❑ Overview: goal, data, and evaluation
  - ❑ Results
  - ❑ How to improve them: discussion
- ❑ Discussion (1 hour)
  - ❑ Role of evaluations
  - ❑ Lessons we have learned
  - ❑ Future contests



# ISMIR 2004 AUDIO DESCRIPTION CONTEST

1<sup>st</sup> World-wide competition on algorithms for audio description for Music IR

12 Laboratories

20 Participants

- Melody estimation
- Tempo induction
- Genre classification
- Artist identification
- Rhythm classification

## ISMIR2004

**Intl. Conf. Music Inform. Retrieval**

**Barcelona 10-14 October**

**<http://ismir2004.ismir.net/IsmirContest.html>**



## ISMIR2004

## PANELISTS

- Juan Bello
- Stephen Downie
- Dan Ellis
- Marc Leman
- Elias Pampalk
- George Tzanetakis



## PANELISTS

□ Juan Bello

□ Ste

□ Dar

□ Ma

□ Elia

□ Geo

### **Juan Bello**

□ Queen Mary University

□ Contest Participant

□ ISMIR 2005 Organizing team



## PANELISTS

### **J. Stephen Downie**

- School of Library and Information Science, U. Illinois at Urbana-Champaign

- IMERSEL project for MIR/MDL evaluation

- ISMIR Steering Committee



## PANELISTS

- Dan Ellis**
- LabRosa at Columbia University
- Ground truth and evaluation frameworks as well as cross-site evaluation experience (with HP and MIT)
- Contest participant

## PANELISTS

- Marc Leman**
- Ghent University, IPEM: Dept. of Musicology
- MAMI project
- ISMIR 2004 Program committee
-

## PANELISTS

- Elias Pampalk**
- PhD student at the Vienna University of Technology
- Machine Learning and Data Mining Group at the Austrian Research Institute for Artificial Intelligence
- Contest participant

## PANELISTS

□ Juan Bello

### **George Tzanetakis**

- Assistant professor at Computer Science Department at University of Victoria
- MARSYAS
- Contest participant

## CONTESTS: INITIAL PROPOSAL

- Audio fingerprinting
- Music genre classification
- Music instrument classification
- Artist Id / Similarity
- Melody estimation
- Rhythm classification
- Tempo induction
- Key / Chord extraction
- Music structure analysis / Chorus detection



## CONTESTS: ACTUAL

- ❑ Melody estimation: 5 participants
- ❑ Artist Id: 2 participants
- ❑ Rhythm classification: 1 participants
- ❑ Music genre classification: 5 participants
- ❑ Tempo induction: 6 participants

## CONTEST OVERVIEW: WHY?

- ~50 researchers at MTG-UPF
- Annotated audio databases
- Computational resources
  - Massive storage
  - Cluster
- Enthusiasm



## CONTEST PRESENTATION: WHY?

- ~50 researchers at MTG-UPF

- 

### GOAL

- 

- Compare state-of-the-art audio description algorithms and systems relevant to MIR

Disclaimer:

- It does not represent the vast and multidisciplinary field of MIR/MDL

- It didn't mean to define as methodological basis for future initiatives

## CONTESTS: GENERAL ISSUES

- ❑ **Evaluation metrics** agreed among participants.
- ❑ **Databases** could have be expanded with contributions of the participants.
- ❑ **Training or testing data** available.
- ❑ Audio content provided as far as **copyright** allow, otherwise metadata is provided or low-level descriptors.
- ❑ **Evaluation** at environment located at UPF labs.
- ❑ **Anonymity** allowed.
- ❑ **Use of external frameworks** permitted.

## MELODY EXTRACTION






Compare algorithms for melody extraction within polyphonic audio: singing voice and solo instrument.

- Emilia Gómez
- Beesuan Ong
- Sebastian Streich
- Maarten Gratchen



## MELODY EXTRACTION: AUDIO DATA

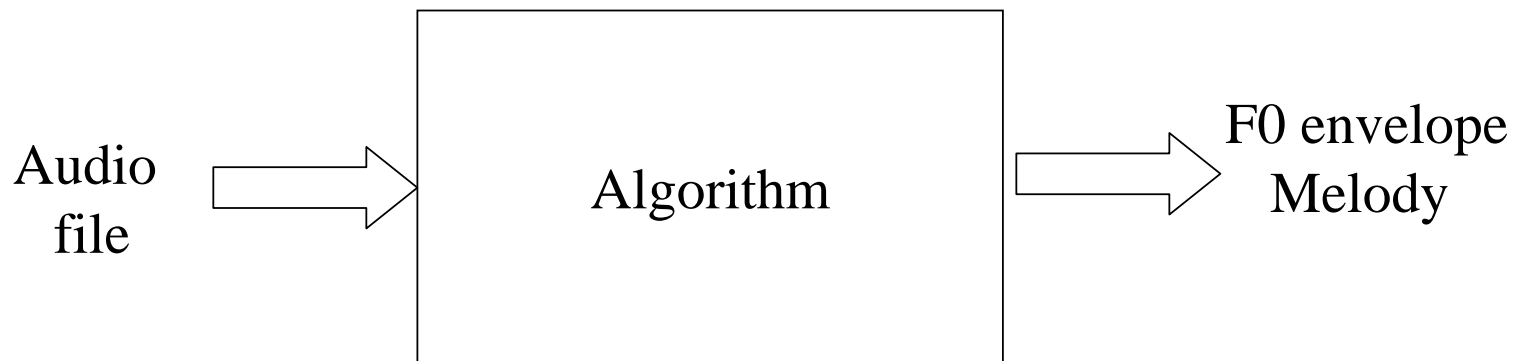
Tuning: 10 audio excerpts + melodic transcription of the predominant voice. Chosen to represent the problem:

-  2 MIDI synthesized
-  2 saxophone + background music
-  2 singing voice synthesizer + background music.
-  2 opera singing + background music.
-  2 items of pop music with singing voice.

Test: tuning set x 2 (20 excerpts)

## MELODY EXTRACTION: SUBMISSION FORMAT

- ❑ Input: file, mono and 44.1 kHz sampling rate.
- ❑ Output:
  - ❑ Option 1,2: predominant F0 envelope in Hz (hop size 256).
  - ❑ Option 3: *onset offset MIDIPitch*



## MELODY EXTRACTION: EVALUATION METRICS

Option 1: frame-based comparison of estimated  $f_0$  and reference  $f_0$  on logarithmic scale. 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.

Option 2: before computing the absolute difference, the values for  $f_0$  are mapped into one octave.

Option 3: Edit distance between melodies (Gratchen et al.)



## MELODY EXTRACTION: PARTICIPANTS

<b>ID</b>	<b>Name</b>	<b>Institution</b>
1	Anonymous	-
2	Sven Tappert	University of Berlin
3	Graham Poliner	Columbia University
4	Juan P. Bello	Centre for Digital Music, Queen Mary University of London

## MELODY EXTRACTION: PARTICIPANTS

<b>ID</b>	<b>Name</b>	<b>Institution</b>
1	Rui Pedro Paiva	University of Coimbra
2	Sven Tappert	University of Berlin
3	Graham Poliner	Columbia University
4	Juan P. Bello	Centre for Digital Music, Queen Mary University of London

# MELODY EXTRACTION: RESULTS

- ❑ *Algorithm with best performance: ID=1. ~65%*
- ❑ *Baseline: monophonic pitch tracker (SMSTools) ~38%*

ID		1				2			3			4				Baseline			
Option		1	2	3	Average12	1	2	Average12	1	2	Average12	1	2	3	Average12	1	2	Average12	
Training set	daisy2	75,23	75,23	4,94	75,23	38,65	69,06	53,86	78,22	78,74	78,48	78,13	78,66	6,92	78,40	68,52	71,38	69,95	
	daisy3	91,10	91,10	0,49	91,10	80,15	80,48	80,31	86,87	87,18	87,03	79,61	79,61	0,56	79,61	1,21	29,39	15,30	
	jazz2	67,82	68,56	6,80	68,19	21,05	55,74	38,40	74,99	74,99	74,99	59,70	67,86	9,66	63,78	46,16	57,90	52,03	
	jazz3	56,10	56,10	6,74	56,10	63,31	65,80	64,56	80,84	80,84	80,84	73,87	73,87	6,09	73,87	34,43	43,06	38,74	
	midi1	74,77	77,58	6,58	76,17	37,80	41,79	39,80	66,60	66,79	66,69	15,79	33,63	26,78	24,71	2,58	16,40	9,49	
	midi2	74,03	74,03	7,66	74,03	75,46	76,43	75,94	78,53	78,53	78,53	77,68	77,68	7,26	77,68	17,28	34,38	25,83	
	opera_fe																		
	m2	35,46	35,49	13,16	35,48	45,00	46,51	45,75	35,68	35,68	35,68	44,73	44,76	13,72	44,74	38,17	44,36	41,26	
	opera_mal																		
	e3	26,07	27,09	19,41	26,58	13,28	35,59	24,44	33,84	33,94	33,89	14,64	28,77	26,70	21,70	44,93	52,91	48,92	
pop1	60,92	61,10	11,69	61,01	17,16	39,26	28,21	55,43	55,43	55,43	25,95	34,74	26,19	30,35	14,40	18,29	16,35		
pop4	70,81	70,84	8,25	70,83	31,81	43,86	37,83	70,82	70,89	70,86	73,08	73,08	9,98	73,08	27,44	34,06	30,75		
Test set	daisy1	66,55	66,55	8,37	66,55	50,71	62,52	56,61	60,38	62,72	61,55	77,23	77,23	10,24	77,23	58,18	64,57	61,37	
	daisy4	89,58	89,58	6,01	89,58	69,22	79,94	74,58	65,04	67,67	66,36	61,94	66,15	8,42	64,04	42,63	53,31	47,97	
	jazz1	61,46	61,82	9,80	61,64	39,37	57,87	48,62	49,67	50,11	49,89	65,66	66,51	6,64	66,08	49,74	58,49	54,12	
	jazz4	78,26	78,26	1,96	78,26	32,83	56,77	44,80	46,41	47,61	47,01	61,11	67,06	4,56	64,08	25,12	34,24	29,68	
	midi3	64,20	64,22	5,30	64,21	61,47	64,37	62,92	50,93	51,42	51,17	42,22	58,30	19,43	50,26	32,59	38,63	35,61	
	midi4	71,97	74,54	5,12	73,25	47,21	52,91	50,06	35,83	41,58	38,71	20,78	37,87	24,82	29,33	2,85	13,91	8,38	
	opera_fe																		
	m4	46,96	46,96	9,22	46,96	55,84	56,36	56,10	20,04	23,51	21,77	44,40	44,40	8,64	44,40	23,44	38,94	31,19	
	opera_mal																		
	e5	46,51	47,19	22,79	46,85	18,42	49,74	34,08	29,43	30,43	29,93	8,58	34,32	31,39	21,45	70,25	74,18	72,21	
pop2	63,94	64,08	10,30	64,01	18,89	38,98	28,93	57,67	58,04	57,86	28,96	36,25	21,72	32,61	31,70	34,95	33,33		
pop3	73,02	73,73	8,10	73,37	26,11	43,56	34,83	45,64	46,69	46,17	62,85	73,17	12,63	68,01	23,31	31,24	27,27		
<b>Average</b>		<b>64,74</b>	<b>65,20</b>	<b>8,63</b>	<b>64,97</b>	<b>42,19</b>	<b>55,88</b>	<b>49,03</b>	<b>56,14</b>	<b>57,14</b>	<b>56,64</b>	<b>50,85</b>	<b>57,70</b>	<b>14,12</b>	<b>54,27</b>	<b>32,75</b>	<b>42,23</b>	<b>37,49</b>	

## MELODY EXTRACTION: RESULTS SPEED

<b>ParticipantID</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>
Operating system	Windows	Linux (MATLAB)	Linux	Linux (MATLAB)
Average Time Per Audio Excerpt (in seconds)	3346,67	60,00	470,00	82,50
Average Time Per Audio Excerpt (in minutes)	55,78	1,00	7,83	1,38

Discussion on the evaluation methods and the corpora

- Does the contest reflect realistic needs?
- Is there enough data in the evaluation database?
- How to improve the evaluation framework?
- Possible related follow-on evaluations.

# RHYTHM CLASSIFICATION

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

□ Fabien Gouyon



## RHYTHM CLASSIFICATION : AUDIO DATA

8 rhythm classes: Samba, Slow Waltz, Viennese Waltz, Tango, Cha Cha, Rumba, Jive, Quickstep.

Participants were given a list of 488 training instances to download from:

*<http://www.ballroomdancers.com/Music/style.asp>*

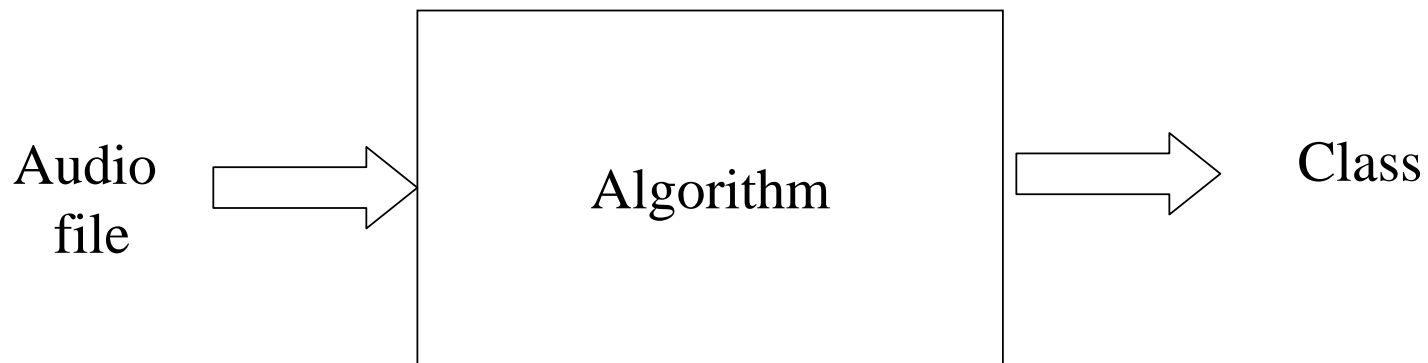
Test data (210 instances) come from the same source.

Format: 30-s instances, real audio decompressed to .wav format, 44100 Hz, 16 bits, mono



## RHYTHM CLASSIFICATION : SUBMISSION FORMAT

- ❑ Input: file, mono and 44.1 kHz sampling rate.
- ❑ Output:
  - ❑ A class among 8 possibilities



Percentage of correctly classified instances



## RHYTHM CLASSIFICATION : PARTICIPANT

1 algorithm submitted

Name	Institutions
Thomas Lidy, Andreas Pesenhofer and Andreas Rauber	Vienna University of Technology and ec3 - eCommerce Competence Center Vienna

## RHYTHM CLASSIFICATION : RESULTS

- ❑ Algorithm performance: ~82% correct classification (210 test instances)
- ❑ Comparison with published results on comparable data (10-fold cross-validations on 698 instances):
  - ❑ Gouyon, Dixon, Pampalk, Widmer, Proc. AES25, 2004, ~79%
  - ❑ Dixon, Gouyon, Widmer, Proc. ISMIR 2004, ~96%



Discussion on the evaluation methods and the corpora

- Does the contest reflect realistic needs?
- Is there enough data in the evaluation database?
- How to improve the evaluation framework?
- Possible related follow-on evaluations.

# MUSIC GENRE CLASSIFICATION

□ The task is to classify songs into genres like Magnatune has organized its site.

classical 

electronic 

jazz\_blues 

metal\_punk 

rock\_pop 

world 

Pedro Cano

Markus Koppenberger

Nicolas Wack

Jose Pedro Mahedero

# MUSIC GENRE CLASSIFICATION : AUDIO DATA

## **Training and development set**

A training and a development available participants. The training and development set consist each of:

classical: 320 pieces.

electronic: 115 pieces

jazz\_blues: 26 pieces

metal\_punk: 45 pieces

rock\_pop: 101 pieces

world: 122 pieces

## **Evaluation**

740 pieces



## MUSIC GENRE CLASSIFICATION : SUBMISSION FORMAT

- ❑ A **framework** composed of python scripts was distributed
- ❑ Participants submitted
  - ❑ Descriptor extractor: wav → features
  - ❑ Train model: features + classes →
  - ❑ Evaluate model: features → classes



# MUSIC GENRE CLASSIFICATION : EVALUATION METRICS

Percentage of correctly classified instances normalized by the genre frequency

$$\sum_{c \in \text{genres}} p_c \cdot \text{guessed}_c$$

## MUSIC GENRE CLASSIFICATION : PARTICIPANTS

Name	Institutions
Thomas Lidy and Andreas Rauber	Vienna University of Technology
Dan Ellis Brian Whitman	Columbia University MIT
Kris West	Univ. of East Anglia
Elias Pampalk	ÖFAI
George Tzanetakis	Univ. of Victoria

## MUSIC GENRE CLASSIFICATION : RESULTS

Name	Results (% Acc / Acc normalized)
Thomas Lidy and Andreas Rauber	70.4 / 55.7
Dan Ellis and Brian Whitman	64 / 51
Kris West	78.3 / 67.2
Elias Pampalk	82.3 / 78.8
George Tzanetakis	71.3 / 58.6

## MUSIC GENRE CLASSIFICATION : RESULTS

□ Robustness to cropping 25 sec middle

□ Lidy 70.4 / 55.7 → 63.4 / 52.1

□ Tzanetakis 71.3 / 58.6 → 57.5 / 24

Discussion on the evaluation methods and the corpora

- Does the contest reflect realistic needs?
- Is there enough data in the evaluation database?
- How to improve it?
- Possible related follow-on evaluations.

## ARTIST IDENTIFICATION

Compare algorithms for artist identification given 3 songs after training with 7 songs

Pedro Cano

Markus Koppenberger

Nicolas Wack

Jose Pedro Mahedero



# ARTIST IDENTIFICATION : AUDIO DATA

## **Training and Development set**

Low-level features (HTK MFCC) corresponding to songs of 105 artists from uspop2002.

The training set includes 7 songs from each artist and the development 3 songs.

## **Evaluation**

205 Artists

7 songs/artist development and 3 for evaluation



# ARTIST IDENTIFICATION : AUDIO DATA

## **Training and Development set**

Low-level features (HTK MFCC) corresponding to songs of 105 artists from uspop2002.

The training set includes 7 songs from each artist and the development 3 songs.

## **Evaluation**

**30 Artists**

7 songs/artist development and 3 for evaluation



## ARTIST IDENTIFICATION : SUBMISSION FORMAT

- ❑ A **framework** composed of python scripts was distributed
- ❑ Participants submitted
  - ❑ Descriptor extractor: wav → features
  - ❑ Train model: features + classes →
  - ❑ Evaluate model: features → classes



Percentage of correctly identified artists



## ARTIST IDENTIFICATION : PARTICIPANTS

Name	Institutions
Thomas Lidy and Andreas Rauber	Vienna University of Technology
Dan Ellis Brian Whitman	Columbia University MIT

## ARTIST IDENTIFICATION : RESULTS

Name	Results (% Acc out of 90 queries)
Thomas Lidy and Andreas Rauber	28 %
Dan Ellis and Brian Whitman	34 %

### Comparison with genre results

- Lidy 55.7 % (6 class) → 28 % (30 class)
- Ellis/Whitman 51 % (6 class) → 34 % (30 class)

Discussion on the evaluation methods and the corpora

- Does the contest reflect realistic needs?
- Is there enough data in the evaluation database?
- How to improve it?
- Possible related follow-on evaluations.

# TEMPO INDUCTION

Compare algorithms that given an input audio file output its basic tempo (i.e. a scalar in BPM)

□ Fabien Gouyon



## TEMPO INDUCTION : AUDIO DATA

### Preparatory data:

7 instances have been given to the participants together with their tempo values in order to compare whether algorithms yield the same outputs when ran in participants' labs and on our machines, and to check proper formatting of algorithm outputs.

### Test set:

3199 tempo-annotated instances

2 to 30 seconds, 50 to 250 BPM,

approximately constant tempi





mono, .wav,  $F_s = 44100\text{Hz}$ , 16 bit resolution.

total duration: ~45140 seconds (i.e. ~12 h 36 mn)



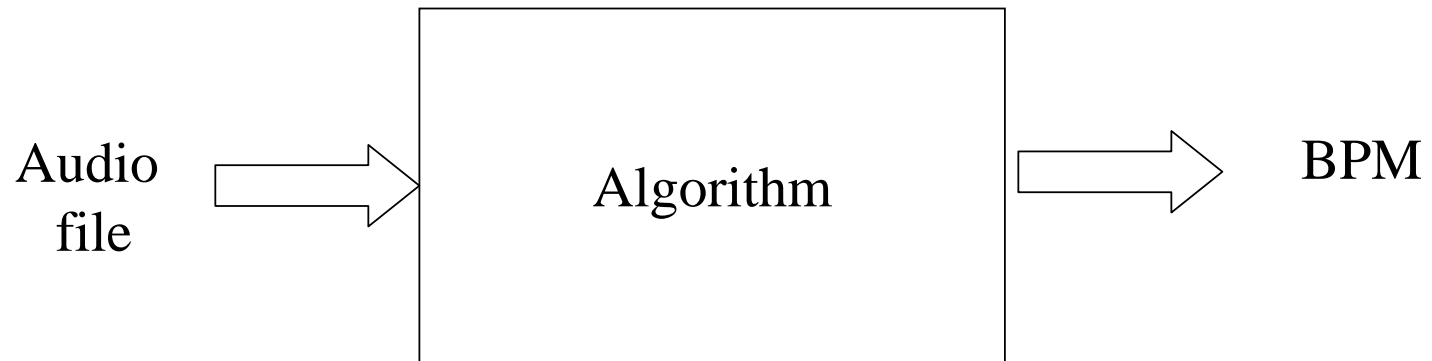
## TEMPO INDUCTION : AUDIO DATA (2)

### Test data details:

1. Loops: from a commercial provider, N=2036, a few bars of drum, bass or electronic loops (Rock, House, Ambient, Electronic, Techno), tempi given 
2. Ballroom dance music excerpts: from a commercial provider (free version in real audio format), N=698, 8 dance styles, tempi given 
3. Personal collection, N=465, musician placed beats manually (ground-truth tempo computed as the mean of the inter-beat interval)   


## TEMPO INDUCTION : SUBMISSION FORMAT

- ❑ Input: file, mono and 44.1 kHz sampling rate.
- ❑ Output: a scalar in BPM



## TEMPO INDUCTION : EVALUATION METRICS

1. *Accuracy 1*: percentage of correct tempi, in BPM, within 4% accuracy
2. *Accuracy 2*: percentage of correct tempo, in BPM, within 4% accuracy, considering half, double, three times,  $1/3$  and  $2/3$  of ground-truth tempo as correct
3. In addition, the 2 previous accuracy measures have been computed on part of the test data (personal collection) *distorted by several processes*, i.e. downsampling/resampling, gsm encoding/decoding, filtering, addition of reverberation and white noise (with a signal-to-noise ratio of 20 dB).



## TEMPO INDUCTION : PARTICIPANTS

12 algorithms:

Name	Institution	# entries
Miguel Alonso	ENST Paris	2
Simon Dixon	OEFAI, Vienna	3
Anssi Klapuri	Univ. Tampere	1
Martin McKinney	Philips	1
George Tzanetakis	Univ. Victoria	3
Christian Uhle	Fraunhofer	1
+ Eric Scheirer *	-	1

(\*) open-source, downloaded from the web



## TEMPO INDUCTION : RESULTS

- ❑ Best algorithm: Klapuri's
  - ❑ Accuracy 1 ~66.9%
  - ❑ Accuracy 2 ~84.3%
  - ❑ Almost no percentage point loss with noisy data (similarly to McKinney and Dixon ISMIR03)
  
- ❑ Fastest algorithm: Dixon BeatRoot induction stage
  
- ❑ Scores range from 66.9% to 22.4% for accuracy 1 and 84.3% to 49.8% for accuracy 2
  
- ❑ Percentage point loss with noisy data range from around 0 to 28.
  
- ❑ Computation time (processing time / excerpt length) range from 0.02 to 15



Discussion on the evaluation methods and the corpora

- Does the contest reflect realistic needs?
- Is there enough data in the evaluation database?
- How to improve it?
- Possible related follow-on evaluations.

## GENERAL ISSUES

- ❑ Need of evaluations
- ❑ Encouraging participation
  - ❑ J. Bello statement
- ❑ Audio and metadata quality issues.
- ❑ Copyright issues
- ❑ Future contests
- ❑ Summary



## ENCOURAGING PARTICIPATION

A great deal of work is independent of the number of participants.

Audio Description Contest allowed

- Anonymous submission
- Different platforms and frameworks

Sources of not participation

- Short time for preparation and submission
- Disagreement on the methodology
- Participants do not always pursue the contest goals.
- O.S. or other technical problems?
- Licensing issues?



## ENCOURAGING PARTICIPATION:

A contestant/future host view:

- Lack of participation
- Well documented and widely referenced systems
- Somehow results should serve as a guideline for people interested on using these systems as front-end
- Intimidating to put your name on the line
- What do we need to do to have YOU guys participating next year
- Are we reflecting the interests of the community?  
(e.g. audio2score alignment, beat tracking)



## AUDIO QUALITY ISSUES

- ❑ Audio quality issues (J. Reiss, ISMIR2004)
- ❑ Is it always important?



## ANNOTATION METADATA

- ❑ Annotation issues (M. Lesaffre, M. Leman et al. 2004)
- ❑ MTGDB as a repository?



## COPYRIGHT ISSUES

All rights reserved

- ❑ Easier to get metadata

  - ❑ AMG

  - ❑ Musicologist sites

- ❑ Real data

but cannot be distributed.

Possible solutions:

- ❑ Distribute features only

- ❑ Citation ( < 30 sec)



- ❑ Easier to distribute audio

- ❑ Could greatly benefit from MIR tools

but

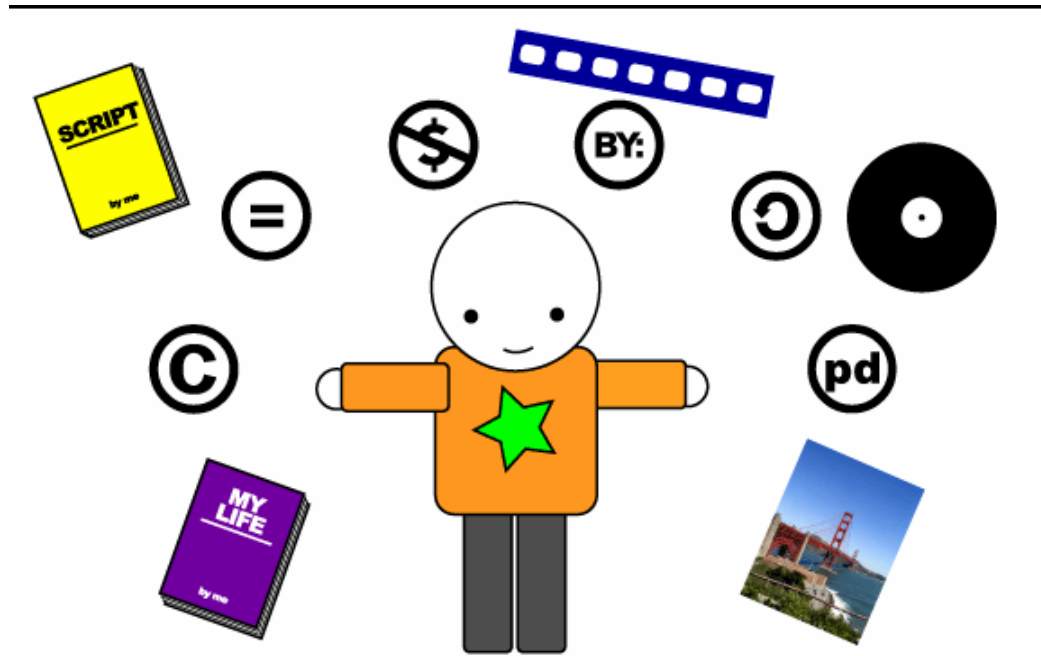
- ❑ Unknown artists

- ❑ Less metadata

# COPYRIGHT ISSUES: CREATIVECOMMONS.ORG



<http://creativecommons.org>



# COPYRIGHT ISSUES: CREATIVECOMMONS.ORG



## FUTURE INITIATIVES

- ❑ ISMIR 2005, QueenMary?
- ❑ UPF?
- ❑ Different labs involved in the preparation of different contests.
- ❑ IMERSEL



# SUMMARY

