How can we evaluate computer models of the highly subjective notion of music similarity?

Arthur Flexer

Austrian Research Institute for Artificial Intelligence (OFAI)

Vienna, Austria
The Austrian Research Institute for Artificial Intelligence - OFAI

- Intelligent Music Processing and Machine Learning Group
- Since ~ 1998
- 12 researchers
Arthur Flexer

- PhD in Psychology, minor degree in Computer Science
- 10 years of research in neuroscience, 10 years in MIR
- Vice-head of Intelligent Music Processing at OFAI

- music recommendation, playlist generation
- commercial applications of MIR
Automatic recommendation / Playlisting

millions of songs

query song

result list

Spotify™

DEEZER
Automatic recommendation / Playlisting

millions of songs

query song + similarity = result list

Spotify™
DEEZER

FM4 Soundpark Player - Mozilla Firefox

http://fire.urf.at/soundpark/syPlayer?mediaid=2141758&cmp=17787

Transferring data from www.soundpark.at...
Computation of similarity between songs

songs as audio

switching to frequencies

computing features

statistical models & similarity-metrics

\[ S(a_1, a_2) = ? \]
Computation of similarity between songs

query song

similar ?

similar ?
Computation of similarity between songs

$\max(S) = 97.9$

query song
Music recommendation - Wolperdinger

- In-house prototype
- 2.5 million 30sec songs
- Similarity based on timbre and rhythm information
- Audio based recommendation!
Are we there yet?
How can we evaluate our models of music similarity?

<table>
<thead>
<tr>
<th>23</th>
<th>45</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>77</td>
<td>100</td>
<td>45</td>
</tr>
<tr>
<td>100</td>
<td>77</td>
<td>23</td>
</tr>
</tbody>
</table>
How can we evaluate our models of music similarity?

Do these numbers correspond to a human assessment of music similarity?
MIREX -
Music Information Retrieval eXchange

- 2014: Grand Challenge on User Experience
- 2014: Audio Classification (Train/Test) Tasks, incorporating:
  - Audio US Pop Genre Classification
  - Audio Latin Genre Classification
  - Audio Music Mood Classification
  - Audio Classical Composer Identification
  - 2014: Audio K-POP Mood Classification
  - 2014: Audio K-POP Genre Classification
- 2014: Audio Cover Song Identification
- 2014: Audio Tag Classification
- 2014: Audio Music Similarity and Retrieval
- 2014: Symbolic Melodic Similarity
- 2014: Audio Onset Detection
- 2014: Audio Key Detection
- 2014: Real-time Audio to Score Alignment (a.k.a Score Following)
- 2014: Query by Singing/Humming
- 2014: Audio Melody Extraction
- 2014: Multiple Fundamental Frequency Estimation & Tracking
- 2014: Audio Chord Estimation
- 2014: Query by Tapping
- 2014: Audio Beat Tracking
- 2014: Structural Segmentation
- 2014: Audio Tempo Estimation
- 2014: Discovery of Repeated Themes & Sections
- 2014: Audio Downbeat Estimation
- 2014: Audio Fingerprinting
- 2014: Singing Voice Separation

International Society for MIR Conference (ISMIR)
Ten years of MIREX (Music Information Retrieval eXchange)

- Standardized testbeds allowing for fair comparison of MIR systems
- Range of different tasks
- Based on human evaluation
  - Directly: humans evaluating MIR system output
  - Indirectly: based on human annotations
Ten years of MIREX (Music Information Retrieval eXchange)

- Standardized testbeds allowing for fair comparison of MIR systems
- Range of different tasks
- Based on human evaluation
  - Directly: humans evaluating MIR system output
  - Indirectly: based on human annotations

What is the level of agreement between human raters/annotators?
What does this mean for the evaluation of MIR systems?
Audio music similarity

Audio music similarity

- Audio Music Similarity and Retrieval (AMS) task 2006-2013
  - 5000 song database
  - participating MIR systems compute 5000x5000 distance matrix
  - 60 randomly selected queries
  - return 5 closest candidate songs for each of the MIR systems
Audio music similarity

- Audio Music Similarity and Retrieval (AMS) task 2006-2013
  - 5000 song database
  - participating MIR systems compute 5000x5000 distance matrix
  - 60 randomly selected queries
  - return 5 closest candidate songs for each of the MIR systems
  - for each query/candidate pair, ask the human grader:
    - „Rate the similarity of the following Query-Candidate pairs. Assign a categorical similarity (Not similar, Somewhat Similar, or Very Similar) and a numeric similarity score. The numeric similarity score ranges from 0 (not similar) to 10 (very similar or identical).“
Audio music similarity

- Audio Music Similarity and Retrieval (AMS) task 2006-2013
  - 7000 song database
  - Participating MIR systems compute 7000x7000 distance matrix
  - 100 randomly selected queries
  - Return 5 closest candidate songs for each of the MIR systems
  - For each query/candidate pair, ask the human grader:
    - “Rate the similarity of the following Query-Candidate pairs. Assign a categorical similarity (Not similar, Somewhat Similar, or Very Similar) and a numeric similarity score. The numeric similarity score ranges from 0 (not similar) to 100 (very similar or identical).”
Audio music similarity

- Audio Music Similarity and Retrieval (AMS) task 2006-2013
  - 7000 song database
  - participating MIR systems compute $7000 \times 7000$ distance matrix
  - 50 randomly selected queries
  - return 10 closest candidate songs for each of the MIR systems
  - for each query/candidate pair, ask the human grader:
  - „Rate the similarity of the following Query-Candidate pairs. Assign a categorical similarity (Not similar, Somewhat Similar, or Very Similar) and a numeric similarity score. The numeric similarity score ranges from 0 (not similar) to 100 (very similar or identical).“
Audio music similarity

- Audio Music Similarity and Retrieval (AMS) task 2006-2013
  - 7000 song database
  - participating MIR systems compute 7000x7000 distance matrix
  - 50 randomly selected queries
  - return 10 closest candidate songs for each of the MIR systems
  - for each query/candidate pair, ask the human grader:
    - „Rate the similarity“ of the following Query-Candidate pairs. Assign a categorical similarity (Not similar, Somewhat Similar, or Very Similar) and a numeric similarity score. The numeric similarity score ranges from 0 (not similar) to 100 (very similar or identical).“
Rate the similarity!
Rate the similarity!

0 … 100 ?

query song ↔ candidate song
Rate the similarity!

Instrumentation?

Timbre?

Melody?

Harmony?

Mood?

Lyrics?

Cultural Background?
Rate the similarity!
Rate the similarity!
Rate the similarity!

- Listening history?
- Musical Preferences?
- Mood of the day?
- Musical training?
- Musical context?
Rate the similarity!

- Factors that influence human music perception

Fig. 1 Factors that influence human music perception.
Rate the similarity!

- Instrumentation?
- Timbre?
- Rhythm?
- Melody?
- Mood of the day?
- Mood?
- Tempo?
- Lyrics?
- Musical context?
- Cultural Background?
- Listening history?
- Harmony?
- Musical Preferences?
- Musical training?

FINE SCORE: 86
Inter-rater agreement in AMS
Inter-rater agreement in AMS

- AMS 2006 is the only year with multiple graders
- each query/candidate pair evaluated by three different human graders
- each grader gives a FINE score between 0 … 10 (not … very similar)
Inter-rater agreement in AMS

- AMS 2006 is the only year with multiple graders
- each query/candidate pair evaluated by three different human graders
- each grader gives a FINE score between 0 ... 10 (not ... very similar)
- correlation between pairs of graders

\[
\begin{array}{|c|c|c|}
\hline
 & \text{grader1} & \text{grader2} & \text{grader3} \\
\hline
\text{grader1} & 1.00 & 0.43 & 0.37 \\
\text{grader2} & 0.43 & 1.00 & 0.40 \\
\text{grader3} & 0.37 & 0.40 & 1.00 \\
\hline
\end{array}
\]

**Table 1.** Correlation of FINE scores between pairs of human graders.
Inter-rater agreement in AMS

- inter-rater agreement for different intervals of FINE scores
Inter-rater agreement in AMS

- Inter-rater agreement for different intervals of FINE scores
Inter-rater agreement in AMS

- inter-rater agreement for different intervals of FINE scores
Inter-rater agreement in AMS

- look at very similar ratings in the [9,10] interval

<table>
<thead>
<tr>
<th></th>
<th>grader1</th>
<th>grader2</th>
<th>grader3</th>
</tr>
</thead>
<tbody>
<tr>
<td>grader1</td>
<td>9.57</td>
<td>6.66</td>
<td>5.99</td>
</tr>
<tr>
<td>grader2</td>
<td>6.60</td>
<td>9.55</td>
<td>6.67</td>
</tr>
<tr>
<td>grader3</td>
<td>6.62</td>
<td>6.87</td>
<td>9.69</td>
</tr>
</tbody>
</table>

**Table 2.** Pairwise inter-rater agreement for FINE scores from interval $v = [9, 10]$. 
Inter-rater agreement in AMS

- Look at very similar ratings in the [9,10] interval

<table>
<thead>
<tr>
<th></th>
<th>grader1</th>
<th>grader2</th>
<th>grader3</th>
</tr>
</thead>
<tbody>
<tr>
<td>grader1</td>
<td>9.57</td>
<td>6.66</td>
<td>5.99</td>
</tr>
<tr>
<td>grader2</td>
<td>6.60</td>
<td>9.55</td>
<td>6.67</td>
</tr>
<tr>
<td>grader3</td>
<td>6.62</td>
<td>6.87</td>
<td>9.69</td>
</tr>
</tbody>
</table>

**Table 2.** Pairwise inter-rater agreement for FINE scores from interval \( v = [9, 10] \).
Inter-rater agreement in AMS

- look at very similar ratings in the [9, 10] interval

<table>
<thead>
<tr>
<th></th>
<th>grader1</th>
<th>grader2</th>
<th>grader3</th>
</tr>
</thead>
<tbody>
<tr>
<td>grader1</td>
<td>9.57</td>
<td>6.66</td>
<td>5.99</td>
</tr>
<tr>
<td>grader2</td>
<td>6.60</td>
<td>9.55</td>
<td>6.67</td>
</tr>
<tr>
<td>grader3</td>
<td>6.62</td>
<td>6.87</td>
<td>9.69</td>
</tr>
</tbody>
</table>

average = 6.54

Table 2. Pairwise inter-rater agreement for FINE scores from interval $v = [9, 10]$. 
Inter-rater agreement in AMS

- look at very similar ratings in the [9,10] interval
- what sounds very similar to one grader, will on average receive a score of only 6.54 from other graders
- this constitutes an upper bound for average FINE scores in AMS

<table>
<thead>
<tr>
<th></th>
<th>grader1</th>
<th>grader2</th>
<th>grader3</th>
</tr>
</thead>
<tbody>
<tr>
<td>grader1</td>
<td>9.57</td>
<td>6.66</td>
<td>5.99</td>
</tr>
<tr>
<td>grader2</td>
<td>6.60</td>
<td>9.55</td>
<td>6.67</td>
</tr>
<tr>
<td>grader3</td>
<td>6.62</td>
<td>6.87</td>
<td>9.69</td>
</tr>
</tbody>
</table>

Table 2. Pairwise inter-rater agreement for FINE scores from interval $v = [9, 10]$. 
Comparison to the upper bound

• compare top performing systems 2007, 2009 – 2013 to upper bound
Comparison to the upper bound

• compare top performing systems 2007, 2009 – 2013 to upper bound
Comparison to the upper bound

- upper bound has already been reached in 2009
Comparison to the upper bound

- PS2 performed at the same level with top performers in all following years
Comparison to the upper bound

- PS2 performed at same level with top performers in all following years
Summary

• low inter-rater agreement in human evaluation of AMS

• upper bound for MIR systems as measured via subjective gradings

• upper bound has already been reached and cannot be surpassed in the future
Why is that?
Why the low inter-rater agreement?
Experimental design

Type of algorithm

Independent variable
(presumed cause, manipulated by researcher)

FINE similarity rating

Dependent variable
(presumed effect, measured by researcher)
Experimental design

Independent variable
(presumed cause,
manipulated by researcher)

Control variable

gender, age,
musical training/experience/preference
type of music, ...

Type of algorithm

FINE similarity rating

Dependent variable
(presumed effect,
measured by researcher)
Experimental design

**Independent variable**
(presumed cause, manipulated by researcher)

**Control variable**

**Dependent variable**
(presumed effect, measured by researcher)

- gender, age,
- musical training/experience/preference
- type of music, ...

Type of algorithm → FINE similarity rating
Experimental design

Control variable, keep it constant

Type of algorithm → FINE similarity rating

male
professional musician
piano concertos
Experimental design

Control variable, keep it constant

Type of algorithm → FINE similarity rating

male
professional musician
piano concertos

very specialized, limited generality
## Experimental design

### Control variable, monitor it

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>gender</th>
<th>musician</th>
<th>genre</th>
<th>FINE score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>male</td>
<td>musician</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>male</td>
<td>musician</td>
<td>rock</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>male</td>
<td>non</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>male</td>
<td>non</td>
<td>rock</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>female</td>
<td>musician</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>female</td>
<td>musician</td>
<td>rock</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>female</td>
<td>non</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>female</td>
<td>non</td>
<td>rock</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>male</td>
<td>musician</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>male</td>
<td>musician</td>
<td>rock</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>male</td>
<td>non</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>male</td>
<td>non</td>
<td>rock</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>female</td>
<td>musician</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>female</td>
<td>musician</td>
<td>rock</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>female</td>
<td>non</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>female</td>
<td>non</td>
<td>rock</td>
<td></td>
</tr>
</tbody>
</table>
### Experimental design

**Control variable, monitor it**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>gender</th>
<th>musician</th>
<th>genre</th>
<th>FINE score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>male</td>
<td>musician</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>male</td>
<td>musician</td>
<td>rock</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>male</td>
<td>non</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>male</td>
<td>non</td>
<td>rock</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>female</td>
<td>musician</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>female</td>
<td>musician</td>
<td>rock</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>female</td>
<td>non</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>female</td>
<td>non</td>
<td>rock</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>male</td>
<td>musician</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>male</td>
<td>musician</td>
<td>rock</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>male</td>
<td>non</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>male</td>
<td>non</td>
<td>rock</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>female</td>
<td>musician</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>female</td>
<td>musician</td>
<td>rock</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>female</td>
<td>non</td>
<td>classic</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>female</td>
<td>non</td>
<td>rock</td>
<td></td>
</tr>
</tbody>
</table>
Experimental design

Type of algorithm

Independent variable
(presumed cause, manipulated by researcher)

FINE similarity rating

Dependent variable
(presumed effect, measured by researcher)
Rate the similarity!

„Rate the similarity of the following Query-Candidate pairs. Assign a categorical similarity (Not similar, Somewhat Similar, or Very Similar) and a numeric similarity score. The numeric similarity score ranges from 0 (not similar) to 10 (very similar or identical).“
Rate the similarity!

• „Rate the similarity of the following Query-Candidate pairs. Assign a categorical similarity (Not similar, Somewhat Similar, or Very Similar) and a numeric similarity score. The numeric similarity score ranges from 0 (not similar) to 10 (very similar or identical).“

• ask a more specific question?
  • more fine-grained notion of similarity
Rate the similarity!

- „Rate the similarity of the following Query-Candidate pairs. Assign a categorical similarity (Not similar, Somewhat Similar, or Very Similar) and a numeric similarity score. The numeric similarity score ranges from 0 (not similar) to 10 (very similar or identical).“

- ask a more specific question?
  - more fine-grained notion of similarity

- does something like abstract music similarity even exist?
  - evaluation of complete MIR systems
  - centered around specific task
  - much clearer goal of evaluation
Grand challenge user experience

• „Holistic, user-centered evaluation of the user experience in interacting with complete, user-facing music information retrieval (MIR) systems”
Grand challenge user experience

- „Holistic, user-centered evaluation of the user experience in interacting with complete, user-facing music information retrieval (MIR) systems”
Grand challenge user experience

• „Holistic, user-centered evaluation of the user experience in interacting with complete, user-facing music information retrieval (MIR) systems”

• Very hard to find significant differences between systems

• Interface seems to be more important than music similarity behind it
Open questions

• How can we evaluate computer models of music similarity?
  • because this prevents progress in MIR!
Open questions

• How can we evaluate computer models of music similarity?
  • because this prevents progress in MIR!

• Should we ask more specific questions?
  • but what exactly are these questions?
Open questions

• How can we evaluate computer models of music similarity?
  • because this prevents progress in MIR!

• Should we ask more specific questions?
  • but what exactly are these questions?

• Will more elaborate experiment designs help?
  • but are they tractable?
Open questions

• How can we evaluate computer models of music similarity?
  • because this prevents progress in MIR!
• Should we ask more specific questions?
  • but what exactly are these questions?
• Will more elaborate experiment designs help?
  • but are they tractable?
• Should we move to a more holistic approach?
  • will interface design cloud questions of music similarity?
Open questions

• How can we evaluate computer models of music similarity?
  • because this prevents progress in MIR!

• Should we ask more specific questions?
  • but what exactly are these questions?

• Will more elaborate experiment designs help?
  • but are they tractable?

• Should we move to a more holistic approach?
  • will interface design cloud questions of music similarity?

• Do radically different approaches to evaluation exist?