A meta-analysis of the MIREX Structure Segmentation task

Introduction

The Music Information Retrieval Evaluation eXchange (MIREX) conducts benchmark evaluations for many MIR tasks. Keeping ground truth (GT) private increases its longevity and prevents over-learning. However, it also impedes learning from the results. [1]

In 2012 anonymized GT (i.e., annotations without titles) were released for the Structure Segmentation task. With this data, we can learn about how algorithms’ performance relates to properties of the annotations.

2012 Structure Segmentation Task

Four data sets
• MIREX09: Beatles, Carole King, Michael Jackson and Queen)
• MIREX10-1: INRIA segment-only annotations of RWC
• MIREX10-2: AIST labels of important sections of RWC
• MIREX12: SALAMI.

Fourteen evaluation metrics
Each penalizes over-segmentation, under-segmentation, or both:
• Pairwise retrieval (precision \(pw_p\), recall \(pw_r\) and f-measure \(pw_f\))
• Rand index (\(Rand\))
• Boundary retrieval with .5 and 3 second tolerance (%\(b_p, b_s, b_b, b_p, b_s, b_b\))
• Median true-to-claim (\(mt2c\) and claim-to-true (\(mc2t\))
• Discriminant and under-segmentation scores (\(So\) and \(Se\))

Five extra evaluation metrics
Released GT data allows new metrics to be computed:
• Average cluster purity (\(asp\)), speaker purity (\(asp\)) and summary K-measure (\(K\))
• Fragmentation score (\(1-f\)) and missed boundary score (\(1-m\))

Nine descriptive statistics
Additional parameters computed from GT may explain algorithm performance:
• Song length \(len\)
• Number of segments \(nss\) and labels \(nl\)
• Mean segment length \(mls\)
• Number of segments per label \(nslp\)
• The same statistics for the estimated description instead of the annotation (\(nss, nl, mls, nslp\))

Q1. Do the metrics behave as expected?
• Metrics excluded by MIREX (\(asp, asp, K, 1-m, 1-f\)) are indeed redundant.
• \(R\) behaves inconsistently: in ranking algorithms, it acts like an under-seg. metric, and in ranking recordings it acts like an over-seg. metric.
• \(So\) resembles summary metrics in ranking algorithms.
• Boundary retrieval with .5 and 3 second tolerances are not correlated.
• Poor precision seems to dominates \(by\). Precision is hard, recall is easy.

Q2. Does algorithm success depend on annotation properties?
Song length and segment length are correlated for annotations, but not for algorithms. Algorithms may be regularizing the length of sections too much instead of adapting to song length the way listeners do.

Q3. Are some datasets harder than others?
MIREX annotations are mostly public. We can use matching to de-anonymize songs. This allows us to observe the difficulty of subsets. We identified MIREX annotations by finding public matches where \(b_p > 0.9\).

On MIREX09, algorithms did better on Beatles songs than on the others. Beatles songs may be harder than others?

Conclusions
• Meta-evaluation is useful to understand behaviour of metrics.
• When information about the GT is published, we learn about algorithm performance and can devise adaptations.
• Releasing GT or GT metadata would only require a change in MIREX policy.

References & Datasets
• MIREX data: nema.lis.illinois.edu/nema_out/mirex2012-results/structure
• AIST: staff.aist.go.jp/m/mgoto/RWC-MDB/AIST-Annotations
• TUT: www.cs.tut.fi/sgn/arg/paulus/structure.html
• INRIA: musicdata.gforge.inria.fr/structureAnnotation.html
• Isophonics: www.isophonics.net/content/reference-annotations
• SALAME: dml.dial.mcgill.ca/research/salami/

Reading the graphs
Correlation given by Kendall’s \(\tau = p – (1-p)\), where \(p\) is the prob. that two lists rank a pair the same way.
We say: \(\tau > 0.8\) is a “strong”, \(\tau > 0.33\) is a “weak” correlation. Bonferroni correction applied.

Bold = strong correlation

related metrics = boxed together

background = significant

Sample data

<table>
<thead>
<tr>
<th>Correlation of evaluation metrics and properties of annotations and estimated output in ranking recordings</th>
<th>Correlation of boundary metrics in ranking algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>(pw_p)</td>
<td>(bp_3)</td>
</tr>
<tr>
<td>(pw_r)</td>
<td>(bp_3)</td>
</tr>
<tr>
<td>(S_o)</td>
<td>(1-f)</td>
</tr>
<tr>
<td>(asp)</td>
<td>(mc2t)</td>
</tr>
<tr>
<td>(K)</td>
<td>(br_3)</td>
</tr>
<tr>
<td>(1-m)</td>
<td>(mt2c)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correlation of labelling metrics in ranking algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>(b_p)</td>
</tr>
<tr>
<td>(b_s)</td>
</tr>
<tr>
<td>(b_b)</td>
</tr>
</tbody>
</table>

| Metrics excluded by MIREX \(asp, asp, K, 1-m, 1-f\) |

Average pairwise F-measure

<table>
<thead>
<tr>
<th>Carole King</th>
<th>Michael Jackson</th>
<th>Queen</th>
<th>Beatles</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
</tr>
</tbody>
</table>