A meta-analysis of the MIREX Structure Segmentation task

Jordan B. L. Smith
<j.smith@qmul.ac.uk>
Elaine Chew
<elaine.chew@qmul.ac.uk>
Queen Mary University of London

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_{p3}, b_{r3}, b_{f3}; $b_{p.5}$, $b_{r.5}$, $b_{f.5}$)
- Median true-to-claim (***mt2c***) and claim-to-true (***mc2t***) distance
- Over- and under-segmentation scores (S_O and S_U)

Five extra evaluation metrics

Released GT data allows new metrics to be computed:

- Average cluster purity (***acp***), 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 (ns_a) and labels (nl_a)
- mean segment length (msl_a)
- number of segments per label ($nslp_a$)
- the same statistics for the estimated description instead of the annotation (ns_e, nl_e, msl_e, $nslp_e$)

Q1. Do the metrics behave as expected?

- Metrics excluded by MIREX (*asp, acp, 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.
- S_O resembles summary metrics in ranking algorithms.
- Boundary retrieval with .5 and 3 second tolerances are not correlated.
- Poor precision seems to dominates b_f. Precision is hard, recall is easy.

Reading the graphs

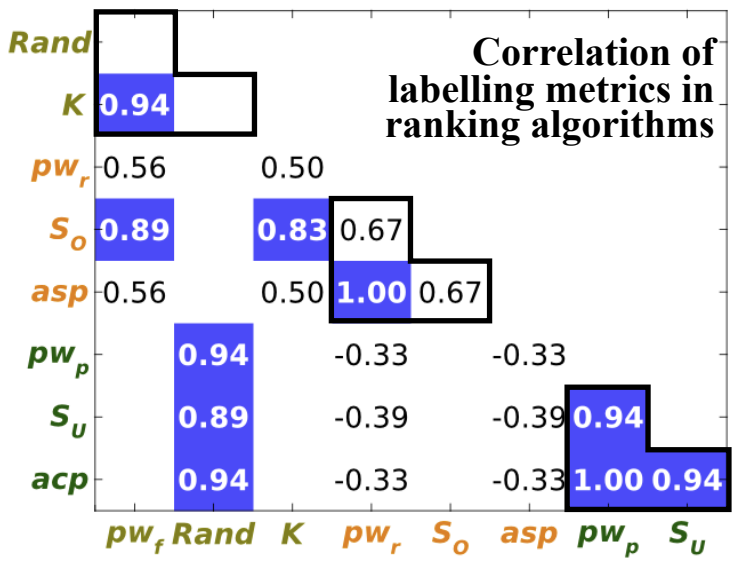
Correlation given by Kendall's $\tau = p - (1 - p)$, where p is the prob. that two lists rank a pair the same way.

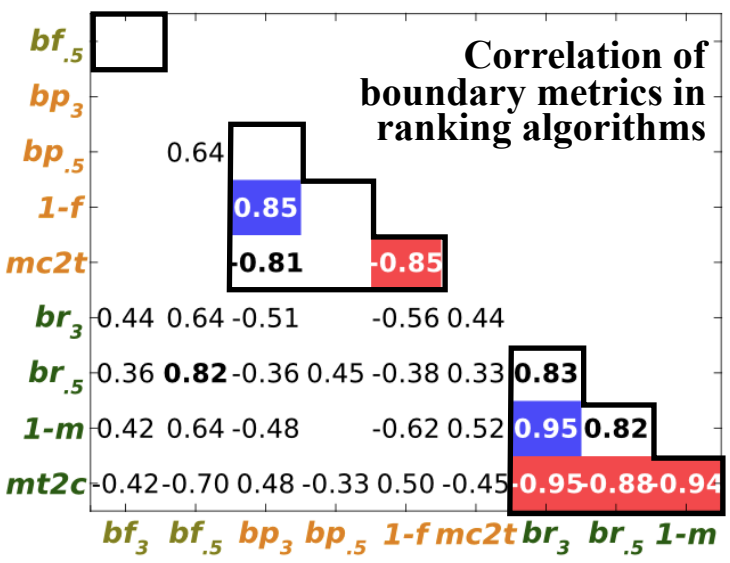
We say: $\tau \geq .8$ is a "strong", $\tau \geq .33$ is a "weak" correlation. Bonferroni correction applied.

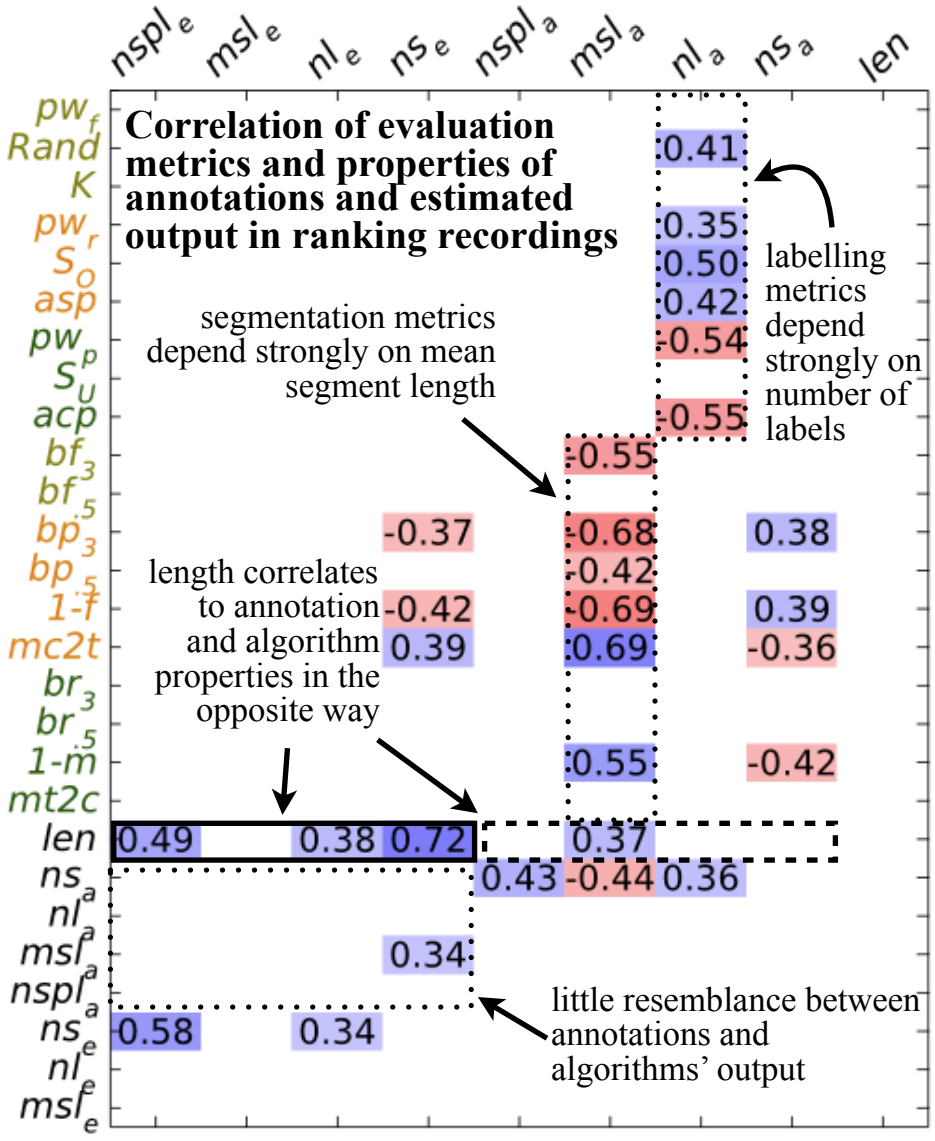
background = significant

related metrics = boxed together

bold = strong correlation







Q2. Does algorithm success depend on annotation properties?

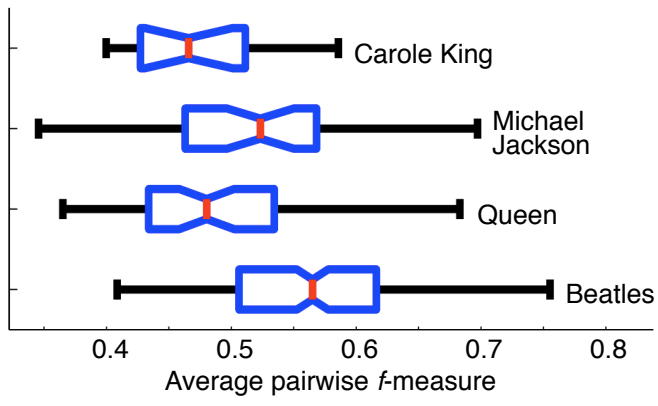
Song length and segment length are correlated for annotations, but not for algorithms. Algorithms may be regularizing the length of sections too much instead of adapting to song length the way listeners do.

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 $b_{f.5} \geq .99$.

On MIREX09, algorithms did better on Beatles songs than on the others. Beatles songs may be easier to analyze, or the algorithms may be tuned too strongly to the Beatles.



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

- [1] Urbano, J. 2011. Information retrieval meta- evaluation: Challenges and opportunities in the music domain. In Proc. ISMIR. Miami, FL. 609–14.

- MIREX data: nema.lis.illinois.edu/nema_out/mirex2012/results/struct/
- AIST: staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/
- EP: www.ifs.tuwien.ac.at/mir/audiosegmentation.html#anchor_corpus
- INRIA: musicdata.gforge.inria.fr/structureAnnotation.html
- Isophonics: www.isophonics.net/content/reference-annotations
- SALAMI: ddmal.music.mcgill.ca/research/salami/annotations
- TUT: www.cs.tut.fi/sgn/arg/paulus/structure.html

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada
Queen Mary University of London
centre for digital music