

A META-ANALYSIS OF THE MIREX STRUCTURE SEGMENTATION TASK

Jordan B. L. Smith

Queen Mary, University of London
jblsmith@eecs.qmul.ac.uk

Elaine Chew

Queen Mary, University of London
elaine.chew@eecs.qmul.ac.uk

ABSTRACT

The Music Information Retrieval Evaluation eXchange (MIREX) serves an essential function in the MIR community, but researchers have noted that the anonymity of its datasets, while useful, has made it difficult to interpret the successes and failures of the algorithms. We use the results of the 2012 MIREX Structural Segmentation task, which was accompanied by anonymous ground truth, to conduct a meta-evaluation of the algorithms. We hope this demonstrates the benefits, to both the participants and evaluators of MIREX, of releasing more data in evaluation tasks.

Our aim is to learn more about the performance of the algorithms by studying how their success relates to properties of the annotations and recordings. We find that some evaluation metrics are redundant, and that several algorithms do not adequately model the true number of segments in typical annotations. We also use publicly available ground truth to identify many of the recordings in the MIREX test sets, allowing us to identify specific pieces on which algorithms generally performed poorly and to discover where the most improvement is needed.

1. INTRODUCTION

MIREX is a highly valued event in the MIR community. Modeled in large part on the evaluations conducted by the Text Retrieval Conference, its role is to establish benchmarks of performance and to allow the community to compare the efficacy of different approaches [5]. MIREX also stimulates competition and helps to drive innovation in areas that the community feels are valuable.

At previous ISMIR conferences, the problems facing MIREX have been a frequent topic of discussion. Some of these are challenges for any evaluation: e.g., the high cost of generating ground truth, and the legality of sharing the music in most test collections [5]. However, [14] points out some issues specific to MIREX, including the problem of hidden data: namely, the results published by MIREX do not identify the songs used in the evaluation or provide metadata other than a general description of the corpus. For example, in the Audio Key Detection task, participants can see how often their algorithm makes different kinds of mistakes—e.g., being off by a major fifth or by a relative key—but cannot see on which pieces their algorithm made the mistakes, or see other information related to the piece, such as the composer, instrumentation, or key.

In his call for more meta-evaluation, Urbano points out that the hidden nature of the MIREX testing data impedes the learning phase of algorithm development [20]. That is, although MIREX evaluations allow the community to compare the performance of state-of-the-art algorithms, it is difficult to learn from the mistakes of the algorithms without any information about the ground truth. For example, although a researcher could learn that their key-detection algorithm makes many parallel key errors, they would have no idea in what situations their algorithm is more likely to make these kinds of errors.

While improving MIREX by creating new ground truth or copyright-free datasets would be expensive, it would only require a change in publishing policies for MIREX to release more detailed evaluation data. In fact, such a change in policy was recently made for the Audio Structural Segmentation (SS) task: in 2012, MIREX posted not only the performance of each algorithm on each piece in the datasets, but the output of the algorithms and the matching annotation. This allows the community for the first time to look more deeply into the large-scale MIREX evaluation, and potentially identify patterns in the performance of the algorithms in order to improve them.

Urbano also recognized the need for more meta-evaluation in the MIR community [20]. Meta evaluation, or the analysis of evaluation systems, is a popular subject in the text retrieval community (e.g., [4]) but has received less attention in MIR. A meta-evaluation investigates whether an evaluation experiment accomplishes its purpose: do the metrics measure the desired quantities or qualities? Is the experiment fairly and effectively estimating the relative quality of different algorithms? With the release of ground truth data in the 2012 SS task, we can attempt to answer some of these questions.

The SS task was introduced to MIREX in 2009, and by then had already been the subject of a meta-evaluation of performance metrics: the merits and shortcomings of over a dozen previously used metrics was discussed in [10], which proposed an additional two. In [6], an extended evaluation of the algorithms submitted to the 2011 MIREX SS task, the authors accessed the algorithms submitted to MIREX in 2010 and tested them on the newly available SALAMI corpus [19]. They compared the algorithms' performance on the large- and small-scale segmentation data, on the different genres within SALAMI and on the different MIREX datasets. With the new MIREX task came a greater interest in generating test collections and designing appropriate methodologies: [13] made recommendations that were adapted by [19] in the creation of the SALAMI dataset, while [3] conceived an alternative methodology, leading to the INRIA collections of annotations.

In this article, we study the results of the 2012 MIREX SS task for these two purposes: first, to learn about the failure modes of the current state of the art algorithms, and second, to appraise the success of the task so far. In Section 2, we review the test sets, algorithms, and evaluation metrics involved in the SS task. In Section 3, we conduct a correlation analysis of the SS evaluation results to find which metrics and basic features could be measuring the same quantities. In Section 4, we show how the ground truth released in 2012 allows us to identify many of the anonymous pieces in the published results, and we survey some of the easiest and hardest songs to analyze.

2. THE STRUCTURE SEGMENTATION TASK

In this section we summarize the materials of the 2012 SS task: the algorithms submitted, the test data used, and the evaluation metrics calculated.

2.1 Algorithms

Five algorithms were evaluated in the 2012 MIREX SS task: KSP [8], MHRAF [11], OYZS, SBV [16] and SMGA [17]. KSP was submitted with four different parameter settings, and two versions of SMGA were submitted, resulting in nine algorithm runs. The algorithms are outlined in abstracts published with the MIREX results, although no abstract is posted for OYZS, and the difference between the two versions of SMGA is not specified in [17]. The four algorithms with abstracts are briefly compared below.

Among the most important differences between the algorithms is each one’s hypothesis about what musical structure is. Using the terminology of [12], the KSP algorithm uses the states hypothesis, which holds that sections are musically homogenous and distinct from one another. MHRAF uses the sequences hypothesis, which holds that sections are defined by distinct sequences. SMGA uses a combined approach, employing a novel feature representation that captures information about long-term repetitions and short-term homogeneity. Finally, SBV is based on the novel “system and contrast” theory of musical structure described in [2]. The algorithm expects that sections will consist of 4 groups of 4 measures, with the fourth group either conforming to or contrasting with the system of musical relationships established by the previous three groups. Other important differences between the algorithms are:

- MHRAF, SBV and SMGA all use harmonic features, while KSP uses a combination of harmonic and timbral features.
- All the algorithms except for MHRAF estimate boundaries first and then estimate segment labels; the MHRAF algorithm detects repetitions first and uses these to define the segmentation.
- SBV is the only algorithm to employ a beat-detection step to align the analysis frames.
- The parameters of the SBV algorithm were set from a test on the INRIA annotations of the RWC database, and a version of SMGA was previously tested on the Beatles and RWC datasets [18].

2.2 MIREX data

The 2012 SS task was evaluated using three datasets:

- MIREX09: A set of 297 pieces introduced in 2009, with annotations taken from the EP [22], Isophonics [24] and TUT collections [26]. According to the analysis in Section 4, of the 343 distinct annotations in these collections, the MIREX test set includes pieces by The Beatles, Carole King, Michael Jackson, and Queen.
- MIREX10: The RWC popular music database, which consists of 100 Japanese pop tunes, with annotations provided by AIST [7, 21] and by INRIA [3, 23]. The INRIA annotations only give boundaries and were first tested in 2010; the AIST annotations, introduced to MIREX in 2011, also provide section labels. INRIA annotations with segment labels were recently introduced [2].
- MIREX12: The SALAMI data [25], introduced to MIREX 2012, and consisting of approximately 859 pieces, over 500 of them with annotations by two listeners [19]. The pieces include popular, classical, world and jazz recordings, publicly available live recordings taken from the Internet Archive, and some pieces borrowed from the Isophonics and RWC collections.

Each of the collections were produced by a small number of listeners (fewer than 10 each) annotating the structure they perceived, but differed in their methodology. The Beatles annotations were adapted from musicologist Allan Pollack’s descriptions [15]. Many of the RWC annotations benefitted from a beat-tracked grid computed from a click track, and only the “obvious” sections were annotated, meaning there are some unannotated gaps in the descriptions [7]. Most SALAMI pieces were annotated by two listeners, and its annotations have separate layers for musical function, similarity at two timescales, and leading instrumentation. Finally, the INRIA annotations indicate boundaries according to a more concrete definition of segments [3].

2.3 Evaluation metrics

For the SS task, MIREX published 14 of the most common metrics reported in the literature: pairwise retrieval (precision pw_p , recall pw_r and f -measure pw_f), proposed by [9]; over- and under-segmentation scores (S_O and S_U , respectively), proposed in [10]; Rand index (R), a metric for comparing partitions of data first used for structural segmentation in the 2009 MIREX task; boundary retrieval with a specified tolerance of 3 seconds (precision b_{p3} , recall b_{r3} and f -measure b_{f3}) or 0.5 seconds ($b_{p.5}$, $b_{r.5}$, $b_{f.5}$); and the median distance from each true to the nearest claimed boundary ($mt2c$) and vice versa ($mc2t$). Since the output of each algorithm is also available [27], it is possible to evaluate the algorithms with metrics not published by MIREX. We did so for 5 other metrics: average cluster purity (acp) and speaker purity (asp), and their summary metric called the K-measure (K), mentioned by [10] as a potential metric in MIR; and the fragmentation and missed-boundary scores (f and m , respectively) used by [1].

Some of these metrics evaluate boundary estimation, and the others evaluate the grouping of sections. Boundary estimation measures either penalize over-segmentation, under-segmentation, or both. Similarly, grouping metrics either penalize the estimation of spurious similarity relationships, the omission of true similarity relationships, or both. Table 1 summarizes the general purpose of the different metrics. Although each metric is

Purpose of the metric	Boundary metrics	Label metrics
Summary metric	bf_3, bf_5	pw_f, R, K
Penalize over-segmentation (spurious boundaries and omitted similarity relationships)	$bp_3, bp_5, l-f, mc2t$	pw_r, S_o, asp
Penalize under-segmentation (omitted boundaries and spurious similarity relationships)	$br_3, br_5, l-m, mt2c$	pw_p, S_u, acp

Table 1. Summary of the metrics by evaluation purpose.

distinct, we expect the metrics in a single group will agree with each other.

3. CORRELATION ANALYSIS

With so many metrics we would like to know whether the metrics in fact measure different things. This problem can be posed in two ways: first, do the metrics differ in how they rank the algorithms? And second, do they differ in how they rank the difficulty of analyzing each recording?

Since our data (the evaluation metrics) are not normally distributed, we compute Kendall’s τ rather than the Pearson correlation. Consider all pairs of items in two ranked lists; if p is the probability that the lists agree on how to rank a pair, then $\tau = p - (1 - p)$ and ranges from $\tau = 1$ for identical rankings to $\tau = -1$ for reversed rankings. With independent lists of rankings, τ is a random variable with mean 0, and we can estimate the precision with which τ has been measured. In all the correlation plots that follow, we use the simple, conservative Bonferroni correction to determine which values for τ are significantly non-zero. Saying whether a given value of τ indicates a “strong” or “weak” correlation remains a subjective decision; we arbitrarily deem $|\tau| \geq 0.8$ as a strong correlation (for positive values, this means that two lists rank at least 9 in 10 pairs of items the same way), $|\tau| \geq 0.33$ as a weak correlation (2 in 3 pairs are ranked the same), and $|\tau| < 0.33$ as no correlation.

3.1 Correlation among metrics

The agreement between the metrics when the algorithms are ranked according to the median grade achieved is shown in Figure 1a. The trio of evaluation measures not used by MIREX (K , asp , acp) rank the algorithms very similarly to the pairwise retrieval metrics (pw_f , pw_r , and pw_p , respectively), supporting their exclusion from MIREX evaluations for being redundant. We expected each metric to be most similar to other metrics measuring the same type of error (under-segmentation, over-segmentation, and both together), but instead we find the over-segmentation metric S_o is more similar to the summary metrics pw_f and K ; and the intended summary metric R is more similar to the under-segmentation metrics S_u , pw_p and acp .

We can also see whether the ranking of the recordings according to difficulty depends on the metric. Here the results (see Figure 1b) conform more to our expectations: K , asp and acp are again found to be redundant;

S_o are grouped appropriately with pw_r and pw_p , but R now resembles an over-segmentation metric.

Performing the same analysis with the boundary evaluation metrics, we again find that related metrics are somewhat redundant: bp_3 , $l-f$, and $mc2t$ are highly inter-correlated, as are br_5 , br_3 , $l-m$ and $mt2c$ (see Figure 2a). Interestingly, bp_5 does not correlate with bp_3 or the other over-segmentation metrics; locating boundaries to within 3 seconds and to within 0.5 seconds are perhaps two distinct skills. This discrepancy is not true for br_5 and br_3 , and hence is probably the cause of the surprising finding that the boundary f -measure summary metrics (bf_3 and bf_5) also do not intercorrelate significantly.

When ranking the recordings (Figure 2b), the groups of metrics (summary, over- and under-segmentation) are each consistent, but the summary metrics are also similar to the over-segmentation ones. Does this indicate that the algorithms are better at boundary precision than recall? In fact, the opposite is the case: mean bp_3 and bp_5 were simply consistently worse for all algorithms.

Lastly, while there is insufficient space to demonstrate it, the findings of this section were consistent across the datasets, albeit with some variation in significance levels.

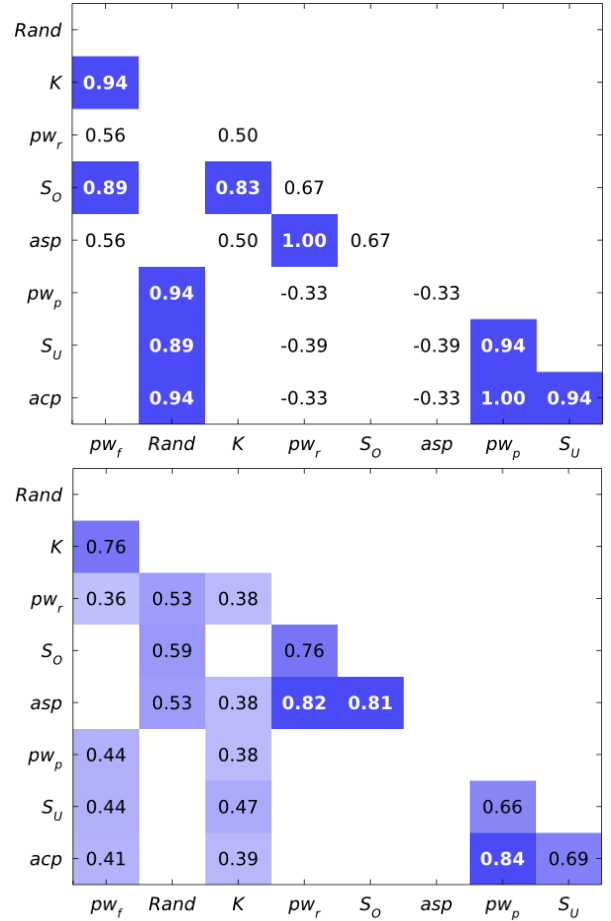


Figure 1a (top). Agreement (Kendall’s τ) between rankings of algorithms (according to median across all recordings) by different labelling metrics. All values of $\tau \geq 0.33$ are plotted. Shaded backgrounds indicate significance; bold values indicate strong agreements.

Figure 1b (above). Agreement in the ranking of recordings by different labelling metrics.

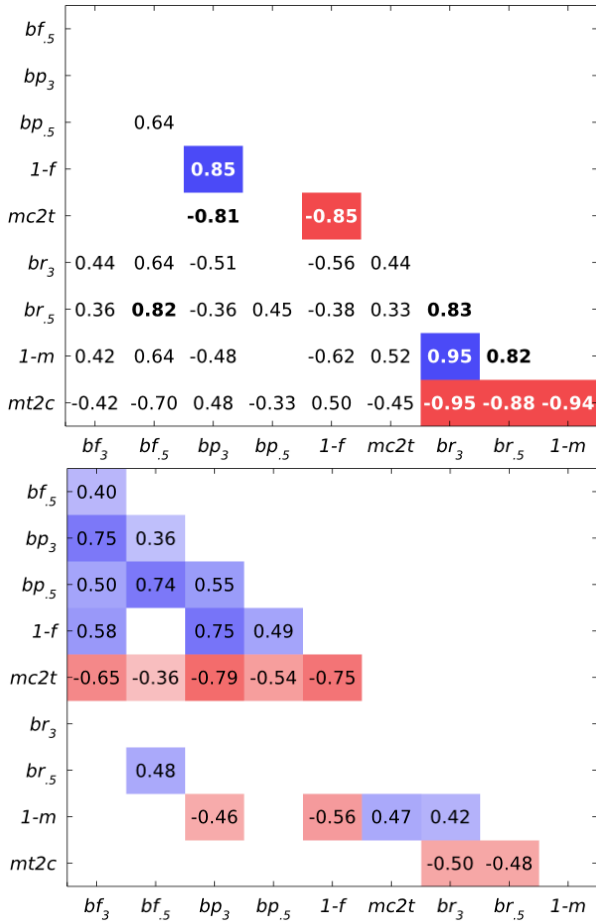


Figure 2a (top). Agreement in the ranking of algorithms by different boundary metrics.

Figure 2b (above). Agreement in the ranking of recordings by different boundary metrics.

3.2 Correlation with ground truth properties

The preceding analysis uses only the evaluation data, not the additional ground truth information made available for the 2012 SS task. With the ground truth we can push the correlation one step further and check whether there are simple properties of the annotated and estimated descriptions that strongly influence the evaluation metrics. For example, it could be that the algorithms simply find longer songs to be more difficult to analyze.

We tested the correlation between all the preceding metrics and the length of the recording (len), as well as ten other properties. Four are properties of the annotation: number of segments (ns_a), number of unique labels (nl_a), mean segment length (msl_a) and the number of segments per label ($nspl_a$). The next four are the same properties for the estimated description (ns_e , nl_e , msl_e , $nspl_e$). We also take the number of extra segments estimated ($ns_e - ns_a$) as a “direct” over-segmentation measure for boundaries (ob), and likewise the number of extra labels ($nl_e - nl_a$) as the over-segmentation measure for labels (ol).

The correlation between these properties reveals an interesting mismatch between the algorithms and the annotations with regard to the mean segment length. In the annotations, song length correlates significantly with mean segment length ($\tau = 0.37$), but hardly at all with the number of segments (0.22). The pattern is reversed

among the algorithm outputs, where song length barely correlates with mean segment length (0.25) but strongly with the number of segments (0.72). It appears that the algorithms do not model how listeners tend to identify a number of sections that is stable across most pieces: while the middle half of the values of ns_e ranges from 7 and 13 segments, the middle values for msl_e for most algorithms range from 11 to 20 segments. The two exceptions are MHRAF and OYZS, for which both msl_e and ns_e match the distributions seen in the annotations.

This shortcoming of the algorithms is reflected in the strong dependence of boundary estimation metrics on the mean segment length in the annotation (msl_a): worse values of b_p , $1-f$ and $mc2t$, all of which punish spurious boundary estimation, are correlated with longer annotated segments.

Figure 3 also shows that the label evaluation metrics seem more sensitive to the number of unique labels in the annotation (nl_a). When nl_a is high, so are asp and S_o (these reflect over-segmentation errors, which are more difficult to make with more fine-grained annotations), while pw_p and acp are reduced, indicating a susceptibility to under-segmentation errors. This dependence suggests that algorithms have difficulty estimating the number of unique segment types heard by a listener.

4. IDENTIFYING MIREX RECORDINGS

In the 2012 SS task, ground truth and the estimated analysis of each algorithm were provided for each piece.

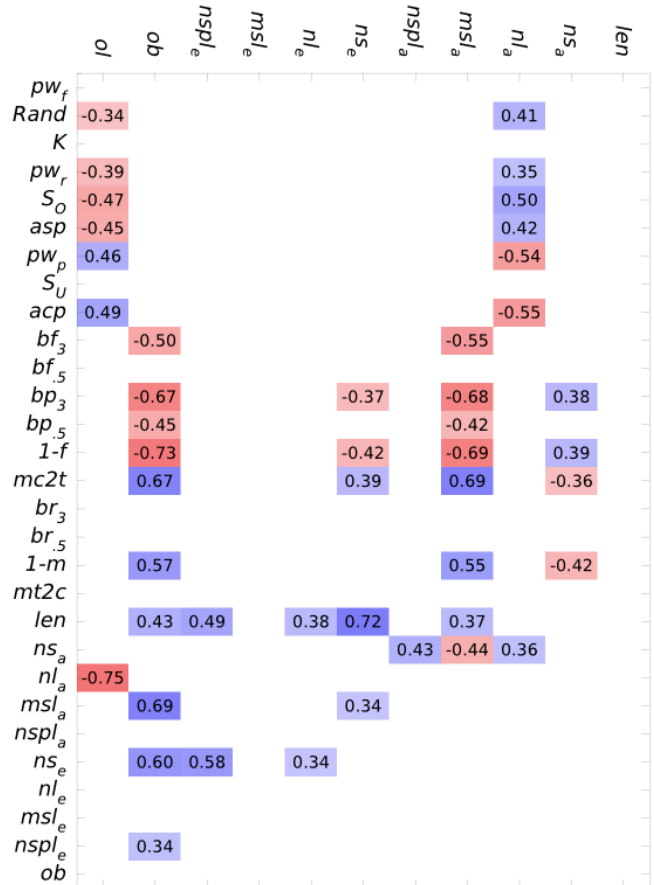


Figure 3. Agreement in the ranking of recordings between metrics and properties of the annotated and estimated descriptions.

Since the ground truth collections used by MIREX are publically available, we can try to identify the recordings in the evaluation by matching the anonymized ground truth published by MIREX with public ground truth.

Before we identify the recordings, we acknowledge that there are advantages to keeping test data private: it is difficult for the designers of algorithms to overfit to hidden data, which means the same data can be reused in successive years; this is useful since ground truth is expensive to create. However, to learn from an evaluation with private data would require the task moderators to conduct meta-evaluation, which is also costly. Moreover, for the SS task, most of the annotation data is already public, and the Beatles and RWC datasets are already widely distributed—indeed, RWC was designed in order to be distributable at cost, without regard for copyright issues. Hence the main advantages of private data do not apply to this task.

The collections used in the SS task are summarized in Table 2. We downloaded all the publicly available annotations for these sources [21–26] to compile a grand public corpus, as well as all MIREX output and annotation for the SS task [27]. For every anonymous MIREX annotation, we searched for public annotations where the lengths of the pieces differed by less than 15 seconds, and computed the boundary f -measure between them. If the boundary f -measure exceeded 0.99, we assumed a match was found. Checking many of the matches informally, it was clear the match was correct.

The number of annotations in the four test collections is provided in Table 1, as well as the number of pieces that were positively matched with an existing annotation. The greatest number of pieces missed were in the SALAMI collection. This is to be expected since half of the SALAMI data remains private.

Associating the MIREX results with actual recordings allows us to search for possible commonalities between the recordings that were “easiest” and “hardest” to annotate. The piece with the highest median pw_f is The Beatles’ “Her Majesty,” a 30-second song with just one section. When a song has just one section, any algorithm is guaranteed to get pairwise precision of 1, and the only boundaries in the song are within 3 seconds of its beginning and end, ensuring boundary recall of 1 as well. The next-best Beatles song, “I Will”, is an instance where both the states and sequences hypotheses apply well: the repeating sections are relatively homogeneous, but contain distinct harmonic sequences. Also, like “Her Maj-

MIREX dataset	Dataset contents	Number of pieces	Number of pieces identified
mrx09	Beatles, Queen, Michael Jackson	297	274
mrx10_1	RWC Popular	100	100
mrx10_2	RWC Popular	100	100
sal	SALAMI	1000	674

Table 2. Summary of the annotations identified in each corpus

esty,” the song is short and contains few sections, reducing the chance of under-segmentation.

On the opposite end, the worst overall performance in the MIREX09 dataset is on songs by Queen and Michael Jackson. At the bottom is Jackson’s “They Don’t Care About Us”. The nine algorithms’ output for this song, along with the ground truth and pw_f scores, are shown in Figure 4. This song is highly repetitive, especially harmonically, although the sung portions are very distinct from the instrumental sections: intro (before 0:42), outro (after 3:50), and interlude (2:40 to 3:10). Most of the estimated descriptions discover this overall structure, but fail to differentiate between the similar verses and choruses. Perhaps algorithms will need to employ more than just harmonic features to improve performance on a case like this.

On the other hand, the annotation also labels the end of the song differently from the chorus, even though they sound similar. That is, the annotation conflates musical similarity with musical function, the situation discussed by [13]. To improve performance, MIREX may wish to update its ground truth to reflect primarily musical similarity; or, algorithms should aim to characterize the semantic labels usually ignored in an analysis.

Conspicuously, 17 of the easiest 20 songs (again, with respect to pw_f) are Beatles tunes, while only 2 of the most difficult 20 songs are—the rest being Michael Jackson, Queen and Carole King songs. Taking the median pw_f across the algorithms and comparing this value for the 274 annotations identified as one of these four artists, a Kruskal-Wallis test confirms that the groups differ. A multiple comparison test reveals that pw_f is significantly greater for the Beatles group than the three others. The simplest explanation is that the songs by the other artists are simply more challenging to analyze than the bulk of the Beatles catalogue. However, this may be evidence that the community is overlearning on the Beatles dataset, which has been widely distributed and used as a test collection for at least 6 years.

5. CONCLUSION

We revisited the 2012 MIREX Structure Segmentation task to better understand the performance of the algo-

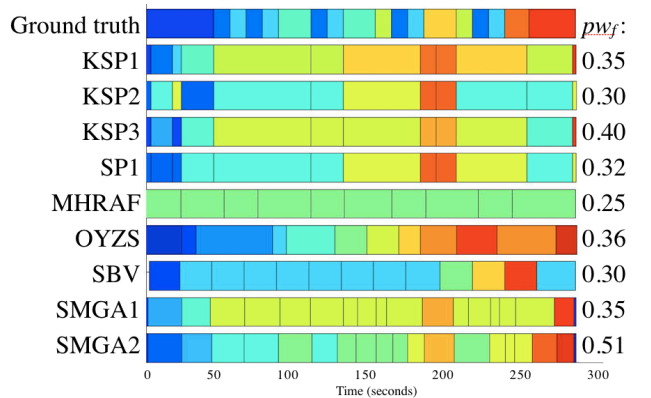


Figure 4. Annotated ground truth and algorithm output for “They Don’t Care About Us” by Michael Jackson. The median pw_f achieved by these algorithms was among the lowest in all of MIREX 2012.

rhythms and the behaviour of the evaluation metrics. Using a correlation analysis, we identified the same metrics excluded from MIREX as redundant (K , asp , acp) and one as unstable and biased (R). Thanks to the release of the ground truth and algorithm output with the 2012 MIREX SS task, we were able to investigate the relationship between evaluation metrics and simple properties of the annotated and estimated descriptions, identifying the lack of regularity in the number of segments per song as a hindrance to many submissions.

We hope that this investigation serves as a positive example of the kind of learning that can be accomplished through meta-analysis. Other MIREX tasks could benefit from the release of algorithm output data and information about the ground truth. While the MIR community must weigh the value of open evaluations with the cost of new datasets, note that it is not necessary to release all the ground truth to benefit a meta-analysis: indeed, this analysis focused mainly on non-identifying parameters of the annotations (Section 3.2) and only a few high- and low-performing songs.

6. ACKNOWLEDGEMENTS

This research was supported in part by the Social Sciences and Humanities Research Council of Canada and a QMUL EPSRC Doctoral Training Account studentship.

7. REFERENCES

- [1] Abdallah, S., Noland, K., Sandler, M., Casey, M., & Rhodes, C. 2005. Theory and evaluation of a Bayesian music structure extractor. In *Proc. ISMIR*. London, UK. 420–5.
- [2] Bimbot, F., Deruty, E., Sargent, G., & Vincent, E. 2012. Semiotic structure labeling of music pieces: Concepts, methods and annotation conventions. In *Proc. ISMIR*. Porto, Portugal. 235–40.
- [3] Bimbot, F., Le Blouch, O., Sargent, G., & Vincent, E. 2010. Decomposition into autonomous and comparable blocks: A structural description of music pieces. In *Proc. ISMIR*. 189–94).
- [4] Buckley, C. and E. M. Voorhees. 2005. Retrieval system evaluation. In *TREC: Experiment and Evaluation in Information Retrieval*. E. M. Voorhees and D. K. Harman, eds. MIT Press, Cambridge, MA.
- [5] Downie, J. S. 2008. The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research. *Acoustical Science and Technology* 29, 247–55.
- [6] Ehmann, A. F., M. Bay, J. S. Downie, I. Fujinaga, and D. De Roure. 2011. Music structure segmentation algorithm evaluation: Expanding on MIREX 2010 analyses and datasets. In *Proc. ISMIR*. Miami, FL. 561–6.
- [7] Goto, M. 2006. AIST annotation for the RWC music database. In *Proc. ISMIR*. Victoria, Canada. 359–60.
- [8] Kaiser, F., Sikora, T., & Peeters, G. 2012. MIREX 2012 - Music Structural Segmentation Task: IrcamStructure submission. In *Late-breaking and demo session, ISMIR*.
- [9] Levy, M. & Sandler, M. 2008. Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 16 (2). 318–26.
- [10] Lukashevich, H. 2008. Towards quantitative measures of evaluating song segmentation. In *Proc. ISMIR*. Philadelphia, PA. 375–80.
- [11] Martin, B., Hanna, P., Robine, M., & Ferraro, P. 2012. Structural analysis of harmonic features using string matching techniques. In *Late-breaking and demo session, ISMIR*.
- [12] Peeters, G. 2004. Deriving musical structures from signal analysis for music audio summary generation: “sequence” and “state” approach. In G. Goos, J. Hartmanis, and J. van Leeuwen (Eds.), *Computer Music Modeling and Retrieval*, 2771: 169–85. Springer Berlin / Heidelberg.
- [13] Peeters, G. & Deruty, E. 2009. Is music structure annotation multi-dimensional? A proposal for robust local music annotation. In *Proceedings of the International Workshop on Learning the Semantics of Audio Signals*. Graz, Austria. 75–90.
- [14] Peeters, G., Urbano, J., & Jones, G. J. F. 2012. Notes from the ISMIR 2012 late-breaking session on evaluation in music information retrieval. In *Late-breaking and demo session, ISMIR*.
- [15] Pollack, A. 2001. “Notes on ... Series.” http://www.icce.rug.nl/~soundscapes/DATABASES/AWP/awp-notes_on.shtml
- [16] Sargent, G., Bimbot, F., & Vincent, E. 2012. A music structure inference algorithm based on morphological analysis. In *Late-breaking and demo session, ISMIR*.
- [17] Serrà, J., Müller, M., Grosche, P., & Arcos, J. L. 2012a. The importance of detecting boundaries in music structure annotation. In *Late-breaking and demo session, ISMIR*.
- [18] Serrà, J., Müller, M., Grosche, P., & Arcos, J. L. 2012b. Unsupervised detection of music boundaries by time series structure features. In *Proceedings of the AAAI International Conference on Artificial Intelligence*, Toronto, Ontario, Canada. 1613–9.
- [19] Smith, J. B. L., Burgoyne, J. A., Fujinaga, I., De Roure, D., & Downie, J. S. 2011. Design and creation of a large-scale database of structural annotations. In *Proc. ISMIR*. Miami, FL. 555–60.
- [20] Urbano, J. 2011. Information retrieval meta-evaluation: Challenges and opportunities in the music domain. In *Proc. ISMIR*. Miami, FL. 609–14.
- [21] AIST: <http://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/>
- [22] EP: http://www.ifs.tuwien.ac.at/mir/audiosegmentation.html#anchor_corpus
- [23] INRIA: <http://musicdata.gforge.inria.fr/structureAnnotation.html>
- [24] Isophonics: <http://www.isophonics.net/content/reference-annotations>
- [25] SALAMI: <http://ddmal.music.mcgill.ca/research/salami/annotations>
- [26] TUT: <http://www.cs.tut.fi/sgn/arg/paulus/structure.html>
- [27] MIREX data: http://nema.lis.illinois.edu/nema_out/mirex2012/results/struct/