USING PRIORS TO IMPROVE ESTIMATES OF MUSIC STRUCTURE

Jordan B. L. Smith **Masataka Goto**
National Institute of Advanced Industrial Science and Technology (AIST), Japan
jordan.smith@aist.go.jp, m.goto@aist.go.jp

ABSTRACT

Existing collections of annotations of musical structure possess many strong regularities: for example, the lengths of segments are approximately log-normally distributed, as is the number of segments per annotation; and the lengths of two adjacent segments are highly likely to have an integer ratio. Since many aspects of structural annotations are highly regular, but few of these regularities are taken into account by current algorithms, we propose several methods of improving predictions of musical structure by using their likelihood according to prior distributions. We test the use of priors to improve a committee of basic segmentation algorithms, and to improve a committee of cutting-edge approaches submitted to MIREX. In both cases, we are unable to improve on the best committee member, meaning that our proposed approach is outperformed by simple parameter tuning. The same negative result was found despite incorporating the priors in multiple ways. To explain the result, we show that although there is a correlation overall between output accuracy and prior likelihood, the weakness of the correlation in the high-likelihood region makes the proposed method infeasible. We suggest that to improve on the state of the art using prior likelihoods, these ought to be incorporated at a deeper level of the algorithm.

1. INTRODUCTION

One reason that the perception of structure in music is such a complex and compelling phenomenon is that it is a combination of 'bottom-up' and 'top-down' processes. It is bottom-up in the sense that a listener first performs grouping on short timescales before understanding the grouping at large timescales, but it is top-down in the sense that one has global expectations that can affect the way one perceives the music. For example, when hearing a new pop song for the first time, we expect there to be a chorus; even on our first hearing, we may identify the chorus partway through a song and already expect it to repeat later. After hearing a verse and a chorus, each 32 beats long, we may expect the bridge to be the same length when it starts.

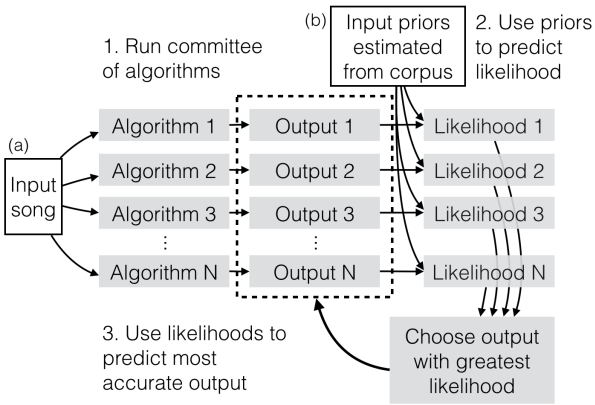


Figure 1. Proposed system overview.

This is important to recognize, since structure is ambiguous: for any piece, there are often multiple ways of interpreting it. As a result, we might never expect a purely bottom-up approach to be 100% correct; we need to also model the top-down influence, or what the listener 'brings' to the analysis.

For instance, consider the following analysis of a piece of music, with the A sections each 10 seconds long, and B 200 seconds long:

```
[-A-|-A-|-A-|------------B------------]
```

Even without knowing the piece of music, we can tell this is an unlikely analysis; it seems wrong to have the segments of the piece sized so asymmetrically. This example hints that the space of *plausible* analyses is limited (even if it is huge), and that listeners' intuitions about these limits inform the annotation process. Is there a way to embed such intuitions into music structure analysis algorithms? Can we employ a kind of 'top-down' critic to assess the likelihood of a given analysis?

We propose a system to accomplish this, illustrated in Figure 1. The inputs to the system are: (a) a song to analyze, and (b) a set of probability density functions (PDFs) estimated from a corpus of annotations. The input song is analyzed by a set of algorithms (step 1); the prior likelihood of each output is computed (step 2); and the estimated description with the highest prior likelihood is chosen (step 3). In contrast to the usual parameter tuning approach, in which a single parameter setting is fixed after evaluating performance over a corpus of songs, in our approach parameters can be tuned for each song, on the basis of prior likelihood.

1.1 Related work

Most algorithms do at least some domain knowledge-based tuning, by putting a lower and/or upper bound on the length of segments, or by filtering features to reduce variations at certain timescales. These are important steps because, although musical structure is hierarchical, algorithms rarely attempt to predict this hierarchy and are evaluated only at a single level. (This status quo has been challenged by [5].)

However, a few algorithms have made greater use of domain knowledge, and their success has been noteworthy. Among the first optimization-based approaches to structure analyis was [7], who explicitly sought to define (in a top-down way) what constitutes a "good" analysis (i.e., one more likely to be in the space of plausible solutions). Later, [10] estimated the median segment length of a piece and used this as the preferred segment length in its search for an optimal segmentation; at the time, their algorithm outperformed the leader at MIREX. The AutoMashUpper system also uses a cost-based approach, rewarding solutions with "good" segment durations of 2, 4, or 8 downbeats, and penalizing ones deemed less likely, like 7 or 9 [2]. In the symbolic domain, [9] also used a cost-minimization approach, with costs increased for segments of unlikely duration or unlikely melodic contour; on one dataset, the approach outperformed a pack of leading melodic-segmentation algorithms.

The most direct way to use domain knowledge is to use supervised learning. Two examples include [14, 15], who each used machine learning to classify short excerpts as boundaries or not based on their resemblance to other short excerpts known to be boundaries. The performance of [15] exceeded the best MIREX result by nearly 10% f-measure for both 0.5- and 3-second thresholds, an enormous achievement.

Our intuition about what constrains the space of plausible analyses, as well as the success of previous algorithms in using domain knowledge and priors learned from corpora, suggest that taking full advantage of this prior knowledge is essential to designing effective algorithms.

In the next section, we survey some of these regularities, and explore the extent to which prior algorithms adhere to them. We detail our proposed algorithms and report our experimental results in Section 3. Alas, despite the solid foundations, no approach will be found to work. The significance of this negative result, and possible explanations for it, are discussed in Section 4.

2. REGULARITIES

In this section we briefly survey some regularities found in the SALAMI corpus of annotations [12], and describe the relationship between these regularities and algorithms that have participated in MIREX campaigns from 2012–14.

Although the time scale of the SALAMI annotations was not explicitly constrained in the Annotator's Guide [1], the length of annotated segments in the SALAMI corpus is roughly log-normally distributed, for both hierarchical levels. Figure 2a shows the PDF of the log segment length ($P(\log L_i)$) for the large and small hierarchical levels in SALAMI; this and all other PDFs in this paper were found using kernel density estimation (KDE). The number of segments within a piece (N) is also log-normally distributed (Figure 2b). If we take the log ratio of each segment's length to the median length of segments within that piece ($\log(L_i/L_{med})$), we obtain a PDF strongly concentrated at $\log(1) = 0$ (see Figure 2c), with additional spikes near ± 0.693, or $\log(2)$ and $\log(1/2)$, for segments of twice or half the median length. There is even more detail if we look at the log length ratio between adjacent segments ($\log(L_i/L_{i+1})$), a histogram of which is shown in Figure 3. Note that all the prominent peaks occur at ratios of small numbers. This makes sense if we consider that segments are usually a whole number of measures long. These properties are not specific to SALAMI annotations; similar distributions were reported by [1] for a completely different corpus of annotations.

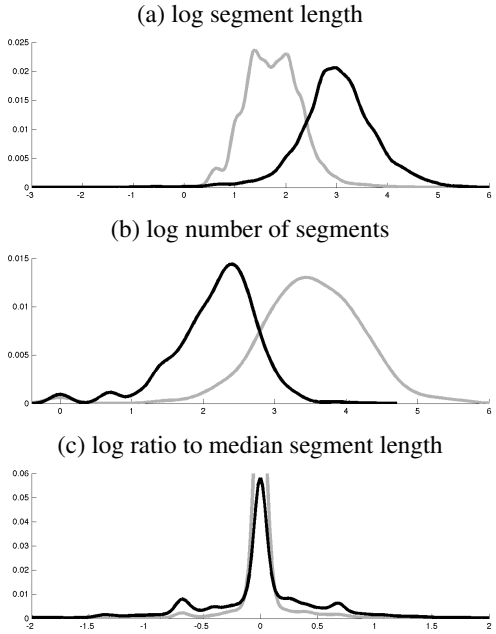


Figure 2. Estimated PDFs of three properties for annotations in SALAMI corpus; x-axis gives the value of the property, y-axis gives its relative probability. PDFs from large-scale segments shown in black, from small-scale segments in gray. In (c), the vertical axis is clipped to show detail; the gray line extends upwards to just over 0.1.

How closely do algorithms model these properties of the annotations? We looked at three years of participants in the MIREX Structural Segmentation task, 2012–14, and estimated PDFs for the same properties. Some examples are shown in Figure 4. Figure 4a shows PDFs for segment length estimated from each algorithm individually: some hew closely to the ground truth, but the majority underestimate the mean segment length. (Since precision is harder to achieve than recall, oversegmentation usually leads to

[1] Available at `http://ddmal.music.mcgill.ca/research/salami/annotations`.

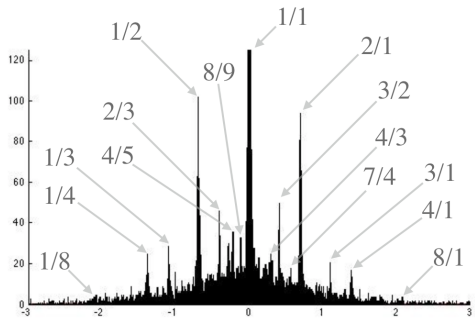


Figure 3. Histogram of $\log(L_i/L_{i+1})$ estimated from SALAMI annotations (y-axis truncated at 1/3 maximum). As shown, nearly every spike represents an integer ratio of segment lengths.

better evaluation scores. [13])

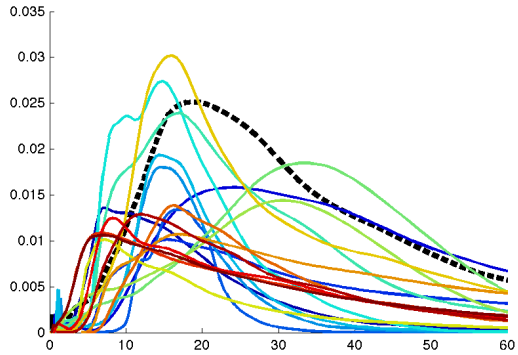
If segment length seems like a weak prior, consider instead Figure 4b, which compares PDFs for the log ratio of adjacent segment length. The characteristic side-lobes representing the high frequency of half- and double-length segments are prominent in only two of the algorithms, RBH1 and RBH3 (2013). This is likely because the algorithm [8] expects boundaries to occur on an 8-measure metrical grid, and snaps estimated boundaries to this grid. Performance (evaluated with f-measure and 3-second threshold) was mixed: RBH1 was close to the state of the art in 2013, while RBH3 was below-average. [2] On the other hand, the next-strongest side-lobes belong to SUG1, the second-best algorithm overall. SUG1 uses a convolutional neural network to classify short excerpts as containing boundaries or not; the method ends by picking peaks from a boundary-likelihood curve, without post-processing [15]. Although SUG1 obviously learns from annotated data, it learns from low-level features (a mel spectrogram) rather than high-level attributes like segment length ratios.

Does the fitness of the algorithms to the SALAMI-derived priors actually have an impact on their performance? We found this to be true by looking at the correlation between algorithm performance and prior likelihood. We took the output of the 18 unique segmentation algorithms that participated in MIREX from 2012–14 [3], and for each algorithm, computed the average log-likelihood of its estimated segments based on the KDE-derived PDFs from SALAMI. We also took the average performance of the algorithms on the three boundary retrieval metrics (f-measure, precision, and recall) with a threshold of 3 seconds. Figure 5 shows the correlation between the mean log-likelihoods (of various segment properties) and the evaluation metrics. There is a weak to moderate correlation between likelihood and f-measure for each of these properties, usually attributable to a strong correlation between likelihood and either precision or recall.

[2] Evaluation results in this paper differ from those reported at MIREX, since we re-evaluated the algorithm output with a 5-second 'warm-up' applied: boundaries within the first and final 5 seconds of pieces were ignored. This leads to lower results overall but better differentiation between algorithms.

[3] Of the 24 participants in these years, 5 used the same segmentation algorithm as another, and the data for one (FK2) were posted later than the others, and were excluded.

(a) PDFs of segment duration (in seconds)



(b) PDFs of log ratio of length of adjacent segments

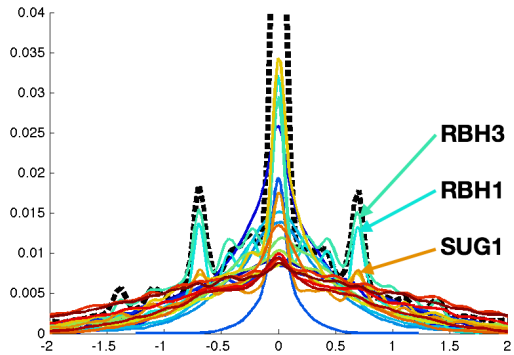


Figure 4. PDFs of properties estimated from SALAMI annotations (black dotted line) and from the output of MIREX algorithms (each algorithm in a different colour).

We have seen that most algorithms deviate substantially from the corpus; the algorithms' descriptions simply don't 'look' like the ground truth. Also, there is some evidence that an algorithm's accuracy is related to the prior likelihood of its output. Hence, it seems reasonable to ask: can we improve on these algorithms, or any set of algorithms, by maximizing their fitness to the priors?

3. USING PRIORS TO IMPROVE A COMMITTEE

Our proposed system is simple: for a single audio file, (1) run several existing structural analysis algorithms, (2) compute the log-likelihood of each prediction with respect to a corpus, and (3) choose the output that maximizes this. (See Figure 1.) For each of these steps, there are many ways to proceed.

3.1 Assembling a committee

We assembled two committees of algorithms: a committee of multiple parameterizations of two basic approaches (Foote [3] and Serra et al. [11]), and a committee of approaches drawn from MIREX. In the first case, we test

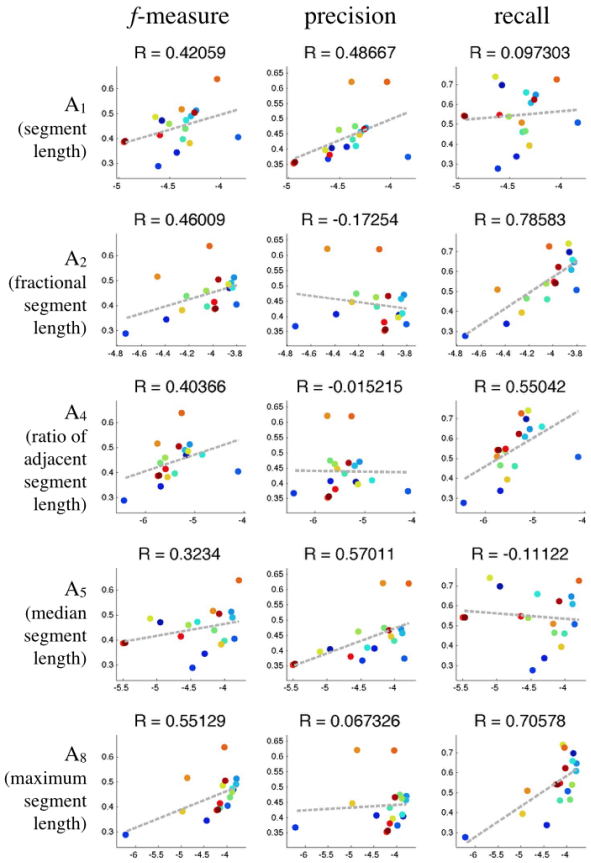


Figure 5. Scatter plots of mean log likelihood of algorithm output (x-axis) and f-measure, precision and recall (y-axis), for 18 MIREX segmentation algorithms, 2012–14. Correlations (Pearson's R) and lines of best fit are shown.

whether a set of more strictly bottom-up segmentation approaches can be improved with the top-down likelihood-maximizing process; in the second case, we test whether a set of already-optimized algorithms can also be improved.

Foote uses a checkerboard kernel to identify discontinuities between homogeneous regions in a self-similarity matrix (SSM), and his remains a classic approach to segmentation. Although surpassed in evaluations such as MIREX, the simplicity and effectiveness of the algorithm means it is still commonly used as a model to improve upon (e.g., see [4]). In contrast to Foote, Serra et al. [11] aim to use both repetition and novelty to locate boundaries. In practice, both algorithms require several design choices: which audio features to use, what amount of smoothing to apply, etc. We ran each algorithm with a small factorial range of settings, including three features (HPCP, MFCCs, and tempograms), for a total of 40 unique settings—hence, 40 committee members. We ran the algorithms on 773 songs within the public SALAMI corpus (version 1.9). Feature extraction and algorithm computation were both easily handled using MSAF [6].

The output of the algorithms that participated in MIREX is publically available, so we simply assembled it, along with the reported algorithm performance, for a MIREX committee of 23 members. We restricted ourselves to the SALAMI portion of the MIREX evaluation, which overlaps significantly with the public half of SALAMI but is not identical.

3.2 Computing likelihoods

We looked at the distribution of several attributes of the SALAMI corpus, listed below. Of these, A_{1-4} are estimated on a per-segment basis and A_{5-9} are global attributes of a description.

- A_1 Segment length (L_i)
- A_2 Fractional segment length (L_i / song length)
- A_3 Ratio of L_i to median segment length
- A_4 Ratio of adjacent segment lengths (L_i/L_{i+1})
- A_5 Median segment length (median of L_i)
- A_6 Number of segments
- A_7 Minimum segment length
- A_8 Maximum segment length
- A_9 Standard deviation of segment length

For attributes A_{1-4}, we took the average across segments. Although log likelihoods are designed to be summed, taking the sum of $\log P(L_i)$ would punish descriptions with more boundaries, regardless of how probable the segments are. (In fact, we did test taking the sum instead of the mean across segments, and the results were generally much poorer.)

3.3 Electing a winner

Once we have computed all of the log likelihoods, how should they be combined, and how should we use these values to elect an answer? Without any *a priori* reason to prefer one over another, we tested multiple approaches:

- choose the description that maximizes the likelihood of attribute A_i;
- choose the description that maximizes a summary statistic over all attributes;
- use a linear model to predict f-measure based on the likelihoods;
- use a linear model with interactions;
- use a quadratic model.

As two summary statistics, we used the sum and the minimum of the log likelihoods of A_i. Using the sum optimizes the general fitness; using the minimum penalizes descriptions with any unlikely attributes.

3.4 Experiment and results

With 5-fold cross-validation, we tested all versions of the algorithm, using both the Foote/Serra and MIREX committees. For each fold, the prior PDFs were estimated only using annotations from the training set.

As a baseline, we used simple parameter tuning: i.e., simply pick the committee member with the greatest average success on the training set. For reference, we also computed the mean f-measure of all committee members, and

Attribute	f (3 sec)	f (0.5 sec)
A_1 (mean)	0.4230	0.1051
A_2 (mean)	0.4156	0.0958
A_3 (mean)	0.4176	0.1140
A_4 (mean)	0.4194	0.1072
A_5	0.3597	0.0863
A_6	0.3781	0.0991
A_7	0.0603	0.0124
A_8	0.3907	0.0961
A_9	0.3956	0.0950
$\sum A_i$	0.4260	0.1093
Min A_i	0.4206	0.1046
Linear model	0.4399	0.0845
Interactions model	0.4451	0.0688
Quadratic model	**0.4494**	0.0739
Baseline	0.4439	**0.1151**
Committee mean	0.2826	0.0691
Theoretical max	0.6015	0.2572

Table 1. Average f-measure (at two thresholds) achieved by different decision criteria for Foote-and-Serra committee.

Attribute	f (3 sec)	f (0.5 sec)
A_1 (mean)	0.6273	0.2733
A_2 (mean)	0.3487	0.0996
A_3 (mean)	0.3487	0.0996
A_4 (mean)	0.3487	0.0996
A_5	0.3916	0.1385
A_6	0.3768	0.1594
A_7	0.3487	0.0996
A_8	0.4662	0.1356
A_9	0.4233	0.1514
$\sum A_i$	0.6273	0.2733
Min A_i	0.6273	0.2733
Linear model	0.5591	0.4005
Interactions model	0.6273	0.4005
Quadratic model	0.6273	0.4005
Baseline	**0.6273**	**0.4005**
Committee mean	0.2826	0.0691
Theoretical max	0.7345	0.5157

Table 2. Average f-measure (at two thresholds) achieved by different decision criteria for MIREX committee.

the theoretical maximum—i.e., the average of the highest-scoring estimates for each song.

The results are shown in Tables 1 and 2. Among all the variations of the proposed method, there were only two instances that surpassed the baseline: the quadratic and interactions models for the Foote-Serra committee, with a 3-second tolerance level. They surpassed the baseline f-measure by 0.0055 and 0.0012, respectively. Given the number of trials conducted, this small amount of success could easily have come by chance.

4. DISCUSSION

Negative results are not normally conclusive: in this case, the reader may suspect that with a small twist, our proposed method may yet succeed. For example, what if we examined subsets by genre, or considered conditional probabilities? In fact, this process of tweaking is how our experiments came about. Our first effort to solve the problem used a small committee of solely Foote-based algorithms, and a set of four log likelihoods. When tests with this initial system gave us a negative result, we tried varying each of the parts of the system—adding more members to the committee, including more PDFs, using increasingly sophisticated regression approaches—until we had assembled the large-scale experiment reported here. And we conducted several more informal tests—looking at subsets of the data, varying the method of characterizing the PDFs (instead of with KDE, they can be modelled with plain histograms, or normal curves can be fitted to some distributions), looking at subcommittees (e.g., removing top-performing and low-performing outlier members) and computing two-dimensional priors (to model, for example, the fact that segment length is not independent of when a segment begins)—all to no avail.

The consistency of the negative result—only two trials out of 56 exceeded the baseline, and only by the slimmest of margins—suggests a dead end. But in order to draw conclusions from this negative result, we must try to understand *why* the approach failed.

Earlier, in Figure 5, we saw that algorithm performance could, over many trials, correlate with the prior likelihood of their output. But what happens when we dig deeper and look at the relationship between each individual output's correctness and its likelihood, as in Figure 6? On the one hand, there is a clear positive trend overall, since there are no examples in the upper-left corner—that is, there are no predictions that have low likelihood but that are close to correct. And the examples with the highest f-measure are also *among* those with the highest likelihoods. Thanks to this relationship, the committees can, despite the noise, generally choose an output that is at least, or slightly better than, the committee's average; that's why, in Tables 1 and 2, nearly all of the algorithms exceeded the average result of the committee.

On the other hand, trying to *find* the high-f-measure predictions based on their prior likelihood is clearly futile when we consider only the rightmost region of the plot, a zoomed-in portion of which is shown in the lower part of Figure 6. Even these predictions, with the greatest fit to the priors, range widely in accuracy: there are plenty above the baseline (0.44), but also plenty below it, including a large number of predictions that contain zero correct boundaries. Figure 6 shows that having a high log-likelihood is a necessary but not sufficient condition to be correct, and it is a condition that most algorithms already achieve.

The uppermost points in Figure 6 represent a few lucky, perfect estimates of the true structure. Their distribution reveals another important point: that although the prior PDFs derive from the ground truth, the prior likelihood of

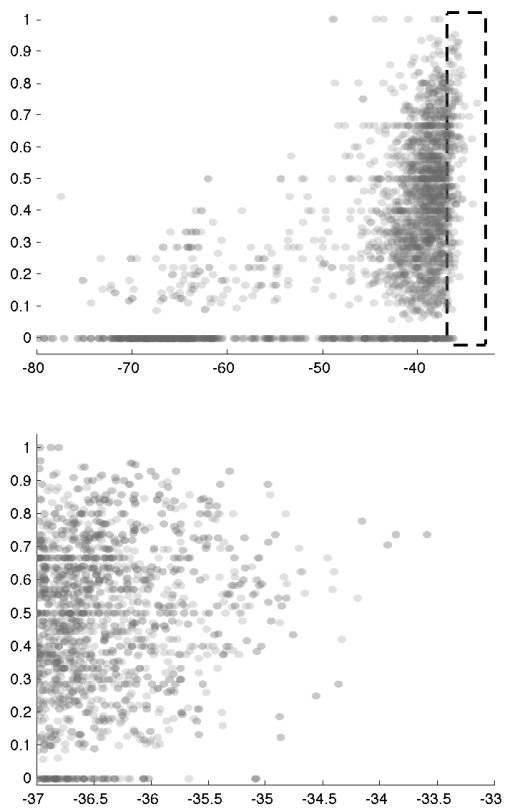


Figure 6. Above: scatter plot of $\sum \log P(A_i)$ (x-axis) vs. f-measure (y-axis, 3-second threshold) of algorithm outputs for Foote/Serra committee on all data. (The heavy horizontal lines are caused by the fact that f-measure is often the product of common fractions.) Below: inset of plot indicated by rectangle.

many annotations is moderate. The fat tails of the PDFs in Figure 2 represent a large set of descriptions that are unlikely to ever be predicted by a prior likelihood-based approach. For example, consider the analyses shown in Figure 7.[4] One algorithm achieved a perfect f-measure (with 3-second threshold), and the likelihood of the description (measured with respect to attribute A_4) was close to that of the annotated description. But a second estimate had a slightly higher prior likelihood thanks to its more consistent segment lengths, and a very poor f-measure.

To sum up the factors that appear to limit the effectiveness of our approach:

1. Although a high f-measure tends to come with a higher prior likelihood, the reverse is not true: plenty of highly probable descriptions are very poor.
2. The moderate correlation between algorithm success and prior likelihood is irrelevant, since we are interested only in the high-likelihood region of estimated descriptions.

[4] Although the MIREX data are anonymized, many songs can be identified by comparing the ground truth to known datasets. [13]

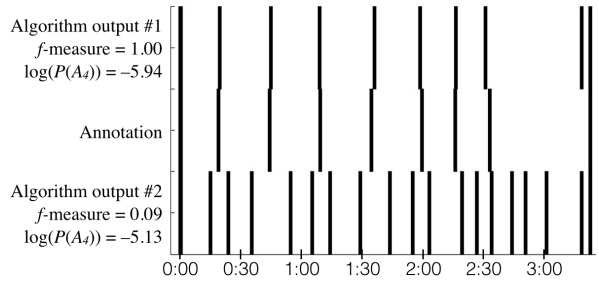


Figure 7. Two algorithmic estimates, compared to the ground truth (middle). The estimates differ somewhat in the likelihood of A_4 (adjacent segment length), but drastically in f-measure. The song is "Rock With You" by Michael Jackson, SALAMI ID 1616.

3. Among high-likelihood descriptions, the correlation between success and likelihood is much weaker: many likely descriptions are poor, and many annotations have low likelihood.

5. CONCLUSION AND FUTURE WORK

We proposed and tested a novel committee-based approach to structural analysis. We motivated the approach by discussing the strong regularities displayed by annotations of music structure. But after a long stretch of negative results, we have concluded that the approach seems unviable: the relationship between a description's prior likelihood and its evaluated score seems to be too weak, especially in the high-likelihood region we are interested in.

We began the article by pointing out some mismatches between the properties of algorithmic estimates of structure and the ground truth, and we suggested that this may be because algorithms do not model top-down factors in perception. For a listener, top-down factors interact with bottom-up factors; in contrast, our algorithm applies bottom-up considerations first (by collecting the committee of estimates), and then applies the top-down considerations *post hoc*. This may be the central weakness of our algorithm. Perhaps, if the top-down influence were modelled earlier on, an estimate like the top one in Figure 7 could be fine-tuned, the boundaries shifted slightly to give a more probable output, rather than rejected early on because of its low likelihood. One algorithm that is ready to test this as future work is the optimization approach of [10]. Although the authors model only a few basic priors, it could be improved by including more.

6. ACKNOWLEDGEMENTS

A version of this work was shown to participants at the Dagstuhl Workshop on Computational Music Structure Analysis. We are very grateful to the attendees for their helpful suggestions, and to the workshop organizers and the Leibniz Center for Informatics for hosting us. This work was supported in part by OngaCREST, CREST, JST.

7. REFERENCES

[1] Frédéric Bimbot, Gabriel Sargent, Emmanuel Deruty, Corentin Guichaoua, and Emmanuel Vincent. Semiotic description of music structure: An introduction to the Quaero/Metiss structural annotations. In *Proceedings of the AES 53rd Conference on Semantic Audio*, 2014.

[2] Matthew E. P. Davies, Philippe Hamel, Kazuyoshi Yoshii, and Masataka Goto. AutoMashUpper: Automatic creation of multi-song music mashups. *IEEE Transactions on Audio, Speech, and Language Processing*, 22(12):1726–1737, 2014.

[3] Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia & Expo*, pages 452–455, 2000.

[4] Florian Kaiser and Geoffroy Peeters. A simple fusion method of state and sequence segmentation for music structure discovery. In *Proceedings of ISMIR*, pages 257–262, Curitiba, Brazil, 2013.

[5] Brian McFee, Oriol Nieto, and Juan Pablo Bello. Hierarchical evaluation of segment boundary detection. In *Proceedings of ISMIR*, Málaga, Spain, 2015.

[6] Oriol Nieto and Juan Pablo Bello. Systematic exploration of computational music structure research. In *Proceedings of ISMIR*, New York, NY, USA, 2016.

[7] Jouni Paulus and Anssi Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech & Language Processing*, 17(6):1159–1170, 2009.

[8] Bruno Rocha, Niels Bogaards, and Aline Honingh. Segmentation and timbre similarity in electronic dance music. In *Proceedings of the Sound and Music Computing Conference*, pages 754–761, Stockholm, Sweden, 2013.

[9] Marcelo Rodríguez-López, Anja Volk, and Dimitrios Bountouridis. Multi-strategy segmentation of melodies. In *Proceedings of ISMIR*, pages 207–212, Taipei, Taiwan, November 2014.

[10] Gabriel Sargent, Frédéric Bimbot, and Emmanuel Vincent. A regularity-constrained Viterbi algorithm and its application to the structural segmentation of songs. In *Proceedings of ISMIR*, pages 483–488, Miami, FL, USA, 2011.

[11] Joan Serrà, Meinard Müller, Peter Grosche, and Josep Ll. Arcos. Unsupervised detection of music boundaries by time series structure features. In *Proceedings of the AAAI International Conference on Artificial Intelligence*, pages 1613–1619, Toronto, Canada, 2012.

[12] Jordan B. L. Smith, J. Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. Design and creation of a large-scale database of structural annotations. In *Proceedings of ISMIR*, pages 555–560, Miami, FL, USA, 2011.

[13] Jordan B. L. Smith and Elaine Chew. A meta-analysis of the MIREX Structure Segmentation task. In *Proceedings of ISMIR*, pages 251–256, Curitiba, Brazil, 2013.

[14] Douglas Turnbull, Gert Lanckriet, Elias Pampalk, and Masataka Goto. A supervised approach for detecting boundaries in music using difference features and boosting. In *Proceedings of ISMIR*, pages 51–54, Vienna, Austria, 2007.

[15] Karen Ullrich, Jan Schlüter, and Thomas Grill. Boundary detection in music structure analysis using convolutional neural networks. In *Proceedings of ISMIR*, pages 417–422, Taipei, Taiwan, November 2014.