QUANTIFYING REGULARITY IN MUSIC STRUCTURE ANALYSIS

Brian McFee
Music and Audio Research Lab, New York University
brian.mcfee@nyu.edu

ABSTRACT

This article describes objective measures of segment regularity for use in evaluating musical structure annotations. The core idea derives from identifying simple ratio relationships between segment durations (*e.g.*, 2:1 or 3:4), and can be implemented in both musical time (beats) or absolute time (seconds). Extensions are proposed to further quantify regularity within labeled segment groups, across hierarchical levels, and evaluate balance or uniformity of segment durations. The efficacy of the proposed methods is demonstrated through an empirical study of several standard datasets for music structure analysis.

The results indicate: 1) under reasonable assumptions of tempo stability, regularity can be reliably measured in absolute time, 2) most existing datasets exhibit regularity, 3) regularity interacts meaningfully with segment labelling, 4) regularity and balance are distinct concepts, and 5) multi-level segmentations exhibit cross-level regularity.

1. INTRODUCTION

Automatic music structure analysis can be thought of as being driven by four fundamental principles: homogeneity, novelty, repetition, and regularity. The first three principles have been fruitfully exploited in algorithm design, *e.g.*, the design of self-similarity matrices as an intermediate representation for boundary detection and section labeling. Similarly, these principles have led to the design of evaluation criteria which quantify the agreement between two annotations under one or more principles (*e.g.*, boundary detection metrics quantify agreement in novelty). The regularity principle, however, has proven to be somewhat trickier to integrate into algorithm design and evaluation. While a few methods have been proposed to promote or enforce regularity among automatically generated segmentations, there is at present no systematic method of *quantifying* the regularity of a temporal segmentation.

This paper describes a family of quantitative metrics to assess the regularity of temporal segmentations. The proposed metrics include formulations for unlabeled, labeled, and hierarchical segmentations, and do not depend on beat or downbeat estimations. Using these metrics, we analyze the reference annotations provided in several commonly used datasets for music structure analysis. The goal of this work is not to propose new algorithms for structure analysis, but rather to gain insights about how "regularity" manifests in existing structural annotations.

2. BACKGROUND AND RELATED WORK

Within the music information retrieval field, there is a well-established literature on music structure analysis, and multiple studies have proposed taxonomies of perceptual and musical properties used to inform the design of algorithms for the task [1–3]. In this work, we follow the most recent survey by Nieto *et al.* [3], which extends the earlier taxonomy of Paulus *et al.* [2] to include four governing principles of music structure analysis: **homogeneity** (segments tend toward self-similarity), **novelty** (segment boundaries coincide with perceptible changes), **repetition** (segments consist of and may be identified by repeating sequences), and most relevant to the current study: **regularity**, which broadly concerns the distribution of segment durations.

Regularity has been invoked in various forms by algorithm developers, though it is relatively under-explored compared to the other governing principles. Sargent *et al.* proposed an explicit objective to penalize deviation from expected segment durations (measured in beats) [4, 5]. While Sargent's penalty is a monotonic function of difference from the expected duration, Marmoret *et al.* proposed a penalty that promotes durations of specific multiples of bar lengths, *e.g.*, preferring segment durations to align with specific integer multiples of bars (8, 4, or 2) [6].

Other authors have proposed implicit models of segment regularity. McFee and Ellis proposed a clustering-based segmentation algorithm in which segments are penalized in proportion to their duration, and the influence of this penalty was optimized over training data [7]. Maezawa proposed a recurrent neural network model to learn the distribution of segment durations from training data [8]. In both cases, the notion of regularity is data-driven and implicit, rather than deriving from explicitly coded domain knowledge. The end result is qualitatively similar, in that the algorithms are incentivized to produce some segmentations over others according to the distribution of durations.

Outside of algorithm design, relatively little focus has been placed on identifying or quantifying regularity in music structure analysis. Most closely related is the work of Smith and Goto, who characterized the distribution of segment durations in the SALAMI dataset [9, 10]. Their study

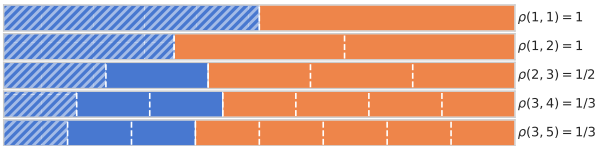

Figure 1. Examples of regular and irregular segmentations determined by d_1 (blue) and d_2 (orange). The patterned regions illustrate the largest unit which divides both d_1 and d_2, and multiples of this unit are marked by dashed lines.

yielded several findings: most notably from the perspective of regularity is the observation that the durations of adjacent segments tend to exhibit simple integer ratios. Smith and Goto exploited this and related observations derived from estimated segment duration to inform the selection of segmentation algorithms in an ensemble method.

This work synthesizes and extends the above notions of regularity. The core idea is a direct extension of the "simple integer ratio" observation of Smith and Goto [9]. It generalizes and formalizes the definitions of regularity proposed by Sargent [5] and Marmoret [6], while also supporting analysis in absolute time rather than relying on potentially inaccurate beat and downbeat estimation.

3. METHODS

Although regularity may seem like a straightforward concept, most prior work stops short of providing a formal definition. For example, Sargent *et al.* define regularity as "segments of comparable size" or "conforming to a specific segment model" (*i.e.*, close to an expected value), and translate this high-level description into a penalty term that scales by divergence from an expected duration. This intuition captures situations where segments have uniform durations (fig. 1, top row), but does not include situations where one duration divides another (fig. 1, second row).

3.1 Temporal divisibility

The proposed notion of regularity derives from the question: *what do two segment durations have in common?* When two segments have equal duration—the most "regular" configuration possible—the answer is *everything*. The same is true, in a sense, when one duration divides the other: the longer duration consists of "regular" repetitions of the shorter duration. More interesting cases arise when the two durations are not integer multiples of each other, *e.g.*, the pair $(2, 3)$. In such cases, we can divide the segments into smaller pieces until a common unit is found that fits evenly into both. The less division is needed to achieve this, the more "regular" the segments appear.

This intuition can be formalized in terms of the greatest common divisor (gcd): in the case above, $\gcd\{2, 3\} = 1$. [1] Normalizing by the smaller of the two durations yields a simple expression that formalizes the question above:

$$\rho(d_1, d_2) := \frac{\gcd\{d_1, d_2\}}{\min\{d_1, d_2\}}. \tag{1}$$

Regularity of two segments is thus defined as the largest unit of relative time that divides both durations.

Because $\gcd$ and $\min$ are both associative operators, eq. (1) would directly generalize to support more than two segments under comparison. [2] However, both operators are sensitive to single input elements: a prime number may dominate the $\gcd$ calculation, while a small number would dominate the $\min$ calculation. This could in turn lead to an overly sensitive metric if applied naively to an entire segmentation. Instead, we may aggregate over pairwise comparisons between distinct segments:

$$R(S) := \frac{2}{|S| \cdot (|S| - 1)} \sum_{d_1 \neq d_2 \in S} \rho(d_1, d_2), \tag{2}$$

where S denotes the set of segment durations. This approach is more robust and lends itself naturally to useful extensions by restricting the pairs under comparison, as demonstrated in sections 3.4 and 3.5.

3.2 Musical time and absolute time

In eq. (1), durations d_1 and d_2 are assumed to be integer-valued so that $\gcd$ is well-defined. This is reasonable when time is measured in musical time (beats), but it is not directly applicable to durations measured in absolute time (seconds).

To resolve this, a two-stage pre-processing of durations is implemented. First, as in most segmentation evaluation metrics [11], durations are quantized with respect to a fixed frame rate f (*e.g.*, 10 Hz), so that $d \mapsto \lfloor d/f \rfloor$. This produces integer-valued durations measured in frames, though it is still possible that approximately commensurate durations would achieve a small ρ-value due to sampling and rounding in the floor operation. To combat this, in the second stage, ρ is computed for values $d_1 + \delta$ and $d_2 + \epsilon$ within a tolerance window $-w \leq \delta, \epsilon \leq w$. As in boundary detection metrics [11], the default w corresponds to 0.5 seconds, or equivalently, one beat at 120BPM. [3] The offsets δ, ϵ are chosen to maximize the score as follows:

$$\tilde{\rho}(d_1, d_2) := \max_{-w \leq \delta, \epsilon \leq w} \rho\left(\left\lfloor \frac{d_1 + \delta}{f} \right\rfloor, \left\lfloor \frac{d_2 + \epsilon}{f} \right\rfloor\right). \tag{3}$$

This modification allows eq. (3) to gracefully support tempo variations and small deviations of boundary positions, while still capturing the core principle of eq. (1). $\tilde{R}(S)$ is analogously defined as the average over all pairwise duration comparisons. The maximization in eq. (3) is performed by computing ρ over a grid of $(2w/f + 1) \times (2w/f + 1)$ sample points, which under the default values described below, is 11×11 and efficient in practice.

[1] We will assume for now that durations are integer-valued; the real-valued extension is described in section 3.2.

Table 1. Empirical mean $\tilde{\rho}$ for uniformly sampled segment durations $d_1 \leq d_2$ in the range $[4, 60]$ seconds, for different frame rates f (Hz) and tolerance window w (seconds).

	$w = 0$	$w = 0.25$	$w = 0.5$
$f = 40$	0.009	0.349	0.477
$f = 20$	0.016	0.343	0.477
$f = 10$	0.029	0.261	0.477
$f = 2$	0.111	0.111	0.427

3.3 Properties

Before going into extensions and applications, it is worth pausing to take note of a few properties of ρ, $\tilde{\rho}$, and R.

Boundedness Since $\gcd\{d_1, d_2\} \leq \min\{d_1, d_2\}$, eqs. (1) and (3) are bounded at 1, with equality when d_2 is an integer multiple of d_1 (or vice versa). The minimal value $1/\min\{d_1, d_2\}$ is achieved by relatively prime (d_1, d_2).

Scale-invariance For any positive rational c such that $c \cdot d_1$ and $c \cdot d_2$ are integers, $\rho(c \cdot d_1, c \cdot d_2) = \rho(d_1, d_2)$.

Attainable values $\rho(d_1, d_2) = 1/N$ for some positive integer N. This is because $c = 1/\gcd\{d_1, d_2\}$, which satisfies the scale invariance property above, implies

$$\rho(d_1, d_2) = \rho(c \cdot d_1, c \cdot d_2) = 1/\min\{c \cdot d_1, c \cdot d_2\}.$$

Expected value Table 1 reports the empirical mean $\tilde{\rho}$ for uniformly sampled duration pairs over the range $[4, 60]$ seconds. For the proposed default tolerance of $w = 0.5$, the mean $\tilde{\rho} \approx 0.477$ is stable for different frame rates f.

3.4 Extension 1: Section labels

Equation (2) averages over all unordered pairs of distinct segments. It often occurs that not all segments are relevant to include in this comparison: for example, introductory silences or crowd noise may exist outside of musical time and therefore not participate meaningfully in regularity. Similarly, sections with significant deviations in tempo from the remainder of the recording may result in low scores under eq. (3), and a case could be made that these should be treated separately.

More generally, one may consider a notion of restricted regularity that only compares segments with the same section label (*e.g.*, *verse* or *chorus*). Under suitable labeling conventions, this view encapsulates the examples listed above, and provides a simple mechanism to exclude segments with sporadically occurring labels. This idea can be implemented with a straightforward modification to eq. (2) where a collection of distinct segment pairs $P \subset S \times S$ is provided rather than the entire segmentation S:

$$R_{\mathsf{L}}(P) = \frac{1}{|P|} \sum_{(d_1, d_2) \in P} \rho(d_1, d_2). \tag{4}$$

[2] The associative property of gcd and min also implies that the edge case of a segmentation consisting of only one segment should produce a score of 1. This convention is adopted here.

[3] δ is constrained to $d + \delta \geq f$ so that eq. (3) is well-defined.

Label agreement is a simple way to generate the pair set P, though the definition above supports other schemes, *e.g.*automatic hierarchy expansion (for approximate agreement) [12]. Relatedly, the temporal proximity observation of Smith and Goto [9] can be implemented here by generating pairs of sequentially adjacent durations:

$$P = \{(d_i, d_{i+1}) \mid 0 \leq i < |S| - 1\}.$$

3.5 Extension 2: Hierarchical regularity

Equation (2) can be modified to evaluate the regularity of *hierarchical* segmentations. Note that eq. (2) operates on pairs of durations, but it does not require that the segments under comparison are disjoint in time or form a valid segmentation. If $H = (S_0, S_1, \dots)$ denotes a multi-level segmentation (with each S_i denoting now the collection of intervals at the ith segmentation level), a pair set P can be generated by matching each segment at level i to its maximally overlapping segment at each level $j < i$. The simplified case of a two-level hierarchy $H = (S_0, S_1)$ yields

$$P = \left\{(|s|, |t|) \;\middle|\; t \in S_1 \wedge s = \text{argmax}_{s \in S_0} |s \cap t|\right\},$$

where $|s|$ denotes the duration of interval s, and $|s \cap t|$ denotes the overlap duration between intervals s and t. Evaluating ρ on each such pair captures how evenly the hierarchy divides segments from one level to the next.

3.6 Extension 3: Balance

Equation (1) captures a form of regularity where durations are related by simple ratios. This differs from previous notions of regularity, which were designed to favor segments of equal duration [5]. This notion can be recovered by replacing the $\min$ normalization in eq. (1) by $\max$:

$$\beta(d_1, d_2) := \frac{\gcd\{d_1, d_2\}}{\max\{d_1, d_2\}}. \tag{5}$$

Equation (5) thus captures the *balance* of d_1 and d_2: a score of 1 is only achieved when $d_1 = d_2$, a score of $1/2$ is achieved when they are related by a factor of 2, and so on. In general, $\beta(d_1, d_2) \leq \rho(d_1, d_2)$, and it otherwise inherits the boundedness, scale-invariance, and integer reciprocal properties noted above. Repeating the calculations behind table 1 for $\tilde{\beta}$ results in an expected value of 0.216 for uniformly random durations and $w = 0.5$.

As above, this also gives rise to an aggregate pairwise score $B(S)$, sampled versions $\tilde{\beta}$ and $\tilde{B}$, and labeled and hierarchical variations.

4. EXPERIMENTS

The proposed metrics are evaluated on the reference annotations provided by a variety of commonly used structure analysis datasets spanning multiple genres:

Beatles (TUT) 174 Beatles songs using the TUT segmentations [13] and Isophonics beat annotations [14].

HarmonixSet 912 popular songs with segment and beat annotations [15].

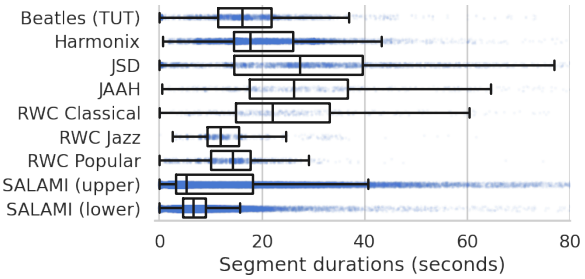

Figure 2. Segment durations for each dataset.

Jazz Structure Dataset (JSD) 340 tracks [16]. For the labeled metrics, the *chorus* and *theme* counter fields are discarded from segment label strings, and segments labeled *silence* are treated as mutually distinct.

Jazz Audio-aligned Harmony (JAAH) 113 tracks with labels derived from the *parts* annotations [17].

Real-world Computing (RWC) 211 tracks (100 popular, 61 classical, 50 jazz) [18]. For labeled metrics, segments labeled as *"nothing"* are treated as mutually distinct, and labels are simplified by discarding parenthetical variations (*e.g.*, *"chorus A (+1)"* $\mapsto$ *"chorus A"*).

SALAMI 1359 tracks from the publicly available dataset [10], consisting of 4486 annotations (2243 upper, 2243 lower). Sections labeled as *"Z"* or *"silence"* are treated as mutually distinct for labeled metrics, and variation markers are discarded (*e.g.*, $A' \mapsto A$).

Figure 2 illustrates the distribution of segment durations for each dataset.

The evaluation seeks to explore the following questions:

1. How do the absolute time metrics ($\tilde{\rho}$, $\tilde{\beta}$) differ from the musical time metrics (ρ, β)?
2. Do structure annotations exhibit regularity and/or balance? Does this vary with genre?
3. Are multi-level segmentations regular across levels?

In service of the first question, we compared scores derived from absolute time (using the approach described in section 3.2) to the simpler forms derived from integer-valued durations measured in beats. This analysis is restricted to the datasets with reference beat annotations: Beatles, HarmonixSet, JAAH, and RWC. Each segment boundary is mapped to its nearest beat, and segment durations d are measured in beats between the start and end boundaries. A preliminary study revealed sensitivities to rounding error in beat position identification, which were resolved by including a maximization over $\{d-1, d, d+1\}$. The (Pearson) correlation was then computed between the musical-time and absolute-time metrics for each dataset.

For the second question, unlabeled and labeled forms of the absolute time metrics were computed. As a point of comparison, metrics were also computed under restriction to adjacent segments [9], denoted here as R_{S}, B_{S}, *etc*.

Table 2. Mean regularity and balance scores using musical time, both unlabeled (R, B) and labeled (R_{L}, B_{L}).

	R	R_{L}	B	B_{L}
Beatles (TUT)	0.681	0.847	0.459	0.834
Harmonix	0.728	0.799	0.524	0.731
JAAH	0.741	0.878	0.488	0.869
RWC Classical	0.599	0.789	0.391	0.765
RWC Jazz	0.914	0.949	0.789	0.945
RWC Popular	0.820	0.958	0.587	0.945

Table 3. Mean regularity and balance scores using absolute time, both unlabeled ($\tilde{R}$, $\tilde{B}$) and labeled ($\tilde{R}_{\mathsf{L}}$, $\tilde{B}_{\mathsf{L}}$).

	$\tilde{R}$	$\tilde{R}_{\mathsf{L}}$	$\tilde{B}$	$\tilde{B}_{\mathsf{L}}$
Beatles (TUT)	0.704	0.820	0.394	0.805
Harmonix	0.730	0.789	0.498	0.719
JSD	0.732	0.606	0.344	0.591
JAAH	0.646	0.793	0.411	0.784
RWC Classical	0.506	0.709	0.298	0.673
RWC Jazz	0.818	0.856	0.720	0.838
RWC Popular	0.791	0.941	0.560	0.925
SALAMI (upper)	0.776	0.719	0.373	0.619
SALAMI (lower)	0.875	0.889	0.684	0.840

For the third question, we restrict attention to the SALAMI dataset, and evaluate hierarchical regularity and balance using the paired upper- and lower-level annotations for each track.

5. RESULTS

5.1 Musical time vs. absolute time

Table 2 reports the average value for the regularity and balance metrics on each of the datasets listed above for which segment durations can be reliably measured in beats. As should be expected, the labeled forms are generally substantially higher than the unlabeled forms. Each dataset exhibits high labeled regularity (significantly above 0.5), as well as high labeled balance, indicating that similarly labeled segments do consistently span equivalent durations.

Table 3 summarizes the absolute-time metrics across all datasets, and Figure 3 illustrates the correlation between these and the musical time data reported in table 2. The correlations are generally high (above 0.6), with a few notable exceptions in the jazz and classical datasets. These exceptions may be explained by the tempo distributions of each dataset, illustrated in fig. 4. Recall that the absolute time metric uses a tolerance window of 0.5 seconds, equivalent to one beat at 120BPM. If a track is much slower—*e.g.*, RWC Classical with median tempo of 87.1, or RWC Jazz with median tempo of 89.4—the maximization in eq. (3) will not cover a full beat, so a larger window may be warranted. However, note that if the tempo is *stable*, this becomes less of an issue because absolute- and musical-time are approximately proportional, which is exploited by the scale-invariance property of ρ.

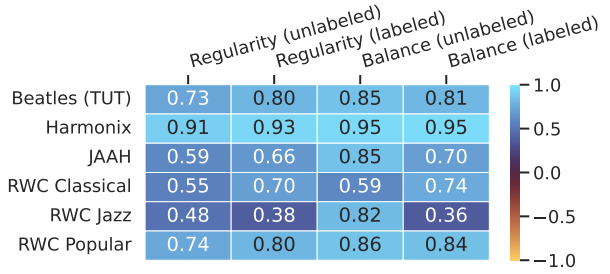

Figure 3. Pearson correlation between musical- and absolute-time metrics for each dataset.

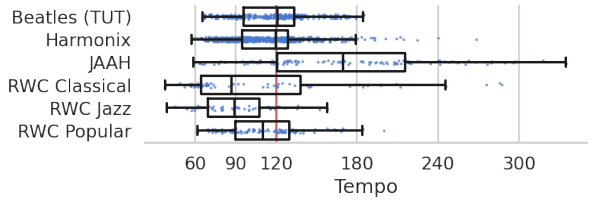

Figure 4. Tempo derived from reference beat annotations. Each point corresponds to the mean tempo for one recording. 120BPM is marked in red as a reference point.

Figure 5 illustrates the distributions of tempo stability, measured as the standard deviation of inter-beat-interval. Datasets with high tempo stability tend to exhibit high correlation in fig. 3 even when they contain many low-tempo tracks (*e.g.*, Beatles, Harmonix, and RWC Pop).

5.2 Unlabeled and labeled regularity

Figure 6 illustrates the relationship between labeled and unlabeled regularity metrics. Consistent with the summary in table 3, the unlabeled regularity scores are generally quite dispersed, while the labeled scores skew higher, confirming that segments belonging to differently labeled sections may not conform to regular duration relationships. Two exceptions to this observation are JSD and SALAMI (upper). In both cases, labeled regularity decreases from the unlabeled scores. These cases may be explained by the use of short *silence* segments, which divide evenly into most other segments, contributing many large values to the

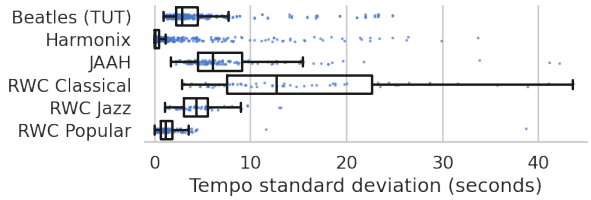

Figure 5. Tempo stability for each dataset, as measured by the standard deviation of local tempo derived from inter-beat intervals in the reference annotations. Each point represents the standard deviation of tempo for one recording.

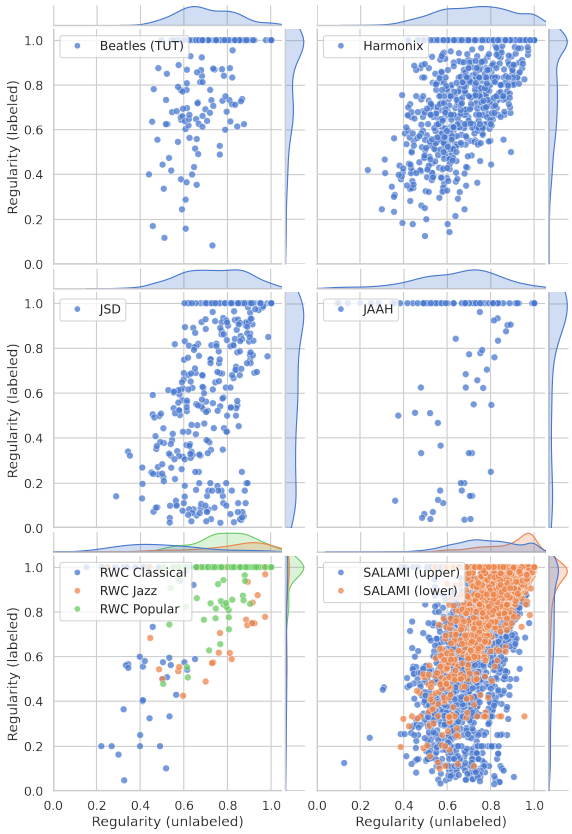

Figure 6. Labeled vs. unlabeled regularity metrics for each annotation in each dataset.

average in eq. (2). In the labeled regularity calculation, each *silence* segment is treated as distinct, eliminating this source of inflation. Segments of this nature are less prevalent in the other datasets (*e.g.*, RWC or JAAH).

Table 4 summarizes the results of regularity and balance when computed on adjacent segment pairs. While there are clear regularity trends, confirming the prior work of Smith and Goto, the effect is not generally as prevalent as the label-agreement results reported in table 3.

5.3 Balance vs. Regularity

Figure 7 illustrates the distribution of the difference between labeled regularity and labeled balance in each

Table 4. Sequential regularity and balance metrics in both musical time (R_S, B_S) and absolute time ($\tilde{R}_S$, $\tilde{B}_S$).

	R_S	$\tilde{R}_S$	B_S	$\tilde{B}_S$
Beatles (TUT)	0.666	0.651	0.399	0.398
Harmonix	0.720	0.696	0.500	0.483
JSD	—	0.729	—	0.479
JAAH	0.786	0.699	0.563	0.495
RWC Classical	0.604	0.521	0.380	0.311
RWC Jazz	0.935	0.872	0.818	0.780
RWC Popular	0.839	0.806	0.592	0.562
SALAMI (upper)	—	0.746	—	0.420
SALAMI (lower)	—	0.882	—	0.753

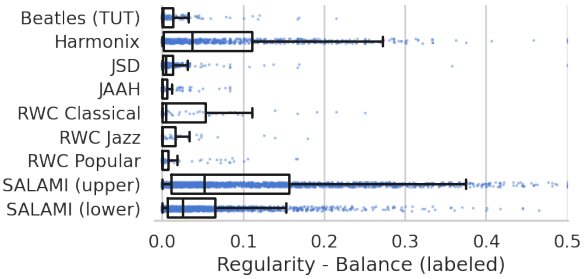

Figure 7. The distributions of difference between labeled regularity and balance: $\Delta = \tilde{R}_{\mathsf{L}} - \tilde{B}_{\mathsf{L}}$ for each dataset.

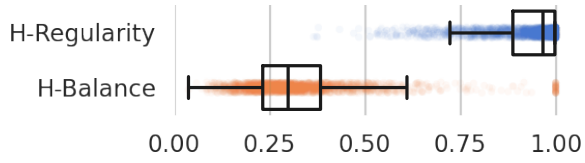

Figure 8. Hierarchical scores on SALAMI.

dataset. The balance scores cannot exceed the regularity scores—so the difference is non-negative—though each dataset does exhibit very high correlation between regularity and balance: all correlation coefficients exceed 0.95. While some datasets generally tend to match balance and regularity (Beatles, JSD, JAAH, RWC Jazz and Pop), others diverge substantially (Harmonix, RWC Classical, SALAMI). This demonstrates that regularity and balance are indeed distinct qualities of segmentation.

5.4 Hierarchical regularity

Figure 8 illustrates the distribution of hierarchical regularity and balance scores on the SALAMI dataset. As expected, the balance scores tend to be low due to the shorter duration of segments in the lower level annotations.

Interestingly, the regularity scores are generally quite high, with a median value of 0.969. This can be interpreted broadly as confirming that upper-level segments are comprised of whole repetitions of lower-level segment durations. While this may be intuitively expected given the annotation rules, it is not an obvious conclusion from the single-level analyses in the previous section. Figure 6 illustrates that lower-level segmentations tend to be highly regular ($\tilde{R}_{\mathsf{L}} \approx 0.889$) and highly balanced ($\tilde{B}_{\mathsf{L}} \approx 0.840$), while upper-level segmentations are slightly less regular ($\tilde{R}_{\mathsf{L}} \approx 0.719$) and often less balanced ($\tilde{B}_{\mathsf{L}} \approx 0.619$).

6. DISCUSSION

From the findings above, we can draw some conclusions about the role of regularity in music structure analysis.

First, because these analyses are conducted on reference annotations (not model outputs), the results reflect the pbehavior of human annotators, and not algorithms. The distribution plots in fig. 6 indicate that although the mean regularity scores are generally high across datasets, there is considerable variability across individual tracks. While these results derive from the absolute time metrics, the high correlation with the musical time metrics suggests that this is generally *not* explained by tempo variation, and rather reflects widespread and meaningful structural irregularity in many datasets. This suggests that regularity, if taken as a design principle in segmentation algorithms, should be treated with some care to allow for irregular segmentations when warranted by the track in question.

Second, the discrepancy between labeled and unlabeled metrics can be quite large (Beatles, Harmonix, RWC Classical and Pop). This corresponds to non-trivial interactions between the regularity and repetition principles (as related to segment label agreement), which had not been identified in previous studies. Modeling and fruitfully exploiting these interactions would be an interesting direction for future work in structure analysis algorithms.

Third, some datasets exhibit significant discrepancies between regularity and balance (Harmonix, SALAMI). This demonstrates that segment durations in fact exhibit more complex patterns than simple equivalence.

7. LIMITATIONS

The proposed methods are applicable to quantitative evaluation of segmentations, but they do exhibit some limitations. First, the absolute time definition does appear to exhibit sensitivity to tempo variation, in particular as it relates to the choice of tolerance window. In situations where high tempo variation may be expected, it may be preferable to either apply the musical time formulation using estimated beat positions (if they are reliable), or adapt the tolerance window to fit the (estimated) tempo of the track.

Second, short segments may artificially inflate scores by being easily divisible into long segments. This is partially addressed by the labeled extension, as short segments tend to be sporadic and unrelated to the majority of a track, *e.g.*, a short *silence* segment at the beginning or end.

Finally, the proposed metrics do not easily lend themselves to differentiable formulations which may be integrated as learning objectives or penalties in current gradient-based learning frameworks. While it may be possible to do so, *e.g.*, by pre-computing a look-up table of pairwise duration comparisons, other difficulties may arise in adapting the ideas into practical segmentation algorithms. Still, the proposed metrics may be more easily integrated as post-processing steps, *e.g.*, to identify meaningful levels to include in a multi-level segmentation, or to select among a collection of proposed segmentations generated by an ensemble of methods.

8. ACKNOWLEDGMENTS

The author thanks Qingyang (Tom) Xi and Meinard Müller for helpful discussions and early feedback.

9. REFERENCES

[1] G. Peeters, "Deriving musical structures from signal analysis for music audio summary generation: "sequence" and "state" approach," in *Computer Music Modeling and Retrieval, International Symposium, CMMR 2003,Montpellier, France, May 26-27, 2003, Revised Papers*, ser. Lecture Notes in Computer Science, U. K. Wiil, Ed., vol. 2771. Springer, 2003, pp. 143–166. [Online]. Available: https://doi.org/10.1007/978-3-540-39900-1_14

[2] J. Paulus, M. Müller, and A. Klapuri, "State of the art report: Audio-based music structure analysis." in *Proceedings of the 11th International Society for Music Information Retrieval Conference*. ISMIR, Aug. 2010, pp. 625–636. [Online]. Available: https://doi.org/10.5281/zenodo.1417289

[3] O. Nieto, G. J. Mysore, C.-i. Wang, J. B. L. Smith, J. Schlüter, T. Grill, and B. McFee, "Audio-based music structure analysis: Current trends, open challenges, and applications," *Transactions of the International Society for Music Information Retrieval*, Dec 2020.

[4] G. Sargent, F. Bimbot, and E. Vincent, "A regularity-constrained viterbi algorithm and its application to the structural segmentation of songs." in *Proceedings of the 12th International Society for Music Information Retrieval Conference*. ISMIR, Oct. 2011, pp. 483–488. [Online]. Available: https://doi.org/10.5281/zenodo.1415950

[5] ——, "Estimating the structural segmentation of popular music pieces under regularity constraints," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 344–358, 2017.

[6] A. Marmoret, J. E. Cohen, and F. Bimbot, "Barwise music structure analysis with the correlation block-matching segmentation algorithm," *Transactions of the International Society for Music Information Retrieval*, Nov 2023.

[7] B. McFee and D. P. W. Ellis, "Learning to segment songs with ordinal linear discriminant analysis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5197–5201.

[8] A. Maezawa, "Music boundary detection based on a hybrid deep model of novelty, homogeneity, repetition and duration," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 206–210.

[9] J. B. L. Smith and M. Goto, "Using priors to improve estimates of music structure." in *Proceedings of the 17th International Society for Music Information Retrieval Conference*. ISMIR, Aug. 2016, pp. 554–560. [Online]. Available: https://doi.org/10.5281/zenodo.1416916

[10] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. D. Roure, and J. S. Downie, "Design and creation of a large-scale database of structural annotations." in *Proceedings of the 12th International Society for Music Information Retrieval Conference*. ISMIR, Oct. 2011, pp. 555–560. [Online]. Available: https://doi.org/10.5281/zenodo.1416884

[11] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "mir_eval: A transparent implementation of common mir metrics." in *Proceedings of the 15th International Society for Music Information Retrieval Conference*. ISMIR, Oct. 2014, pp. 367–372. [Online]. Available: https://doi.org/10.5281/zenodo.1416528

[12] B. McFee and K. Kinnaird, "Improving structure evaluation through automatic hierarchy expansion," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*. ISMIR, Nov. 2019, pp. 152–158. [Online]. Available: https://doi.org/10.5281/zenodo.3527764

[13] J. Paulus, "Improving markov model based music piece structure labelling with acoustic information." in *Proceedings of the 11th International Society for Music Information Retrieval Conference*. ISMIR, Aug. 2010, pp. 303–308. [Online]. Available: https://doi.org/10.5281/zenodo.1416732

[14] C. Harte, "Towards automatic extraction of harmony information from music signals," Ph.D. dissertation, Department of Electronic Engineering, Queen Mary, University of London, 2010.

[15] O. Nieto, M. McCallum, M. Davies, A. Robertson, A. Stark, and E. Egozy, "The Harmonix Set: Beats, downbeats, and functional segment annotations of western popular music," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*. ISMIR, Nov. 2019, pp. 565–572. [Online]. Available: https://doi.org/10.5281/zenodo.3527870

[16] S. Balke, J. Reck, C. WeiSS, J. AbeSSer, and M. Müller, "JSD: A dataset for structure analysis in jazz music," *Transactions of the International Society for Music Information Retrieval*, Nov 2022.

[17] V. Eremenko, E. Demirel, B. Bozkurt, and X. Serra, "Audio-aligned jazz harmony dataset for automatic chord transcription and corpus-based research," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*. ISMIR, Sep. 2018, pp. 483–490. [Online]. Available: https://doi.org/10.5281/zenodo.1492457

[18] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical and jazz music databases." in *Proceedings of the 3rd*

International Conference on Music Information Retrieval. ISMIR, Oct. 2002. [Online]. Available: https://doi.org/10.5281/zenodo.1416474