11-2019

Improving Structure Evaluation Through Automatic Hierarchy Expansion

Brian McFee
New York University

Katherine M. Kinnaird
Smith College, kkinnaird@smith.edu

IMPROVING STRUCTURE EVALUATION THROUGH AUTOMATIC HIERARCHY EXPANSION

Brian McFee
Music and Audio Research Lab &
Center for Data Science
New York University
brian.mcfee@nyu.edu

Katherine M. Kinnaird
Department of Computer Science &
Statistical and Data Sciences Program
Smith College
kkinnaird@smith.edu

ABSTRACT

Structural segmentation is the task of partitioning a recording into non-overlapping time intervals, and labeling each segment with an identifying marker such as A, B, or *verse*. Hierarchical structure annotation expands this idea to allow an annotator to segment a song with multiple levels of granularity. While there has been recent progress in developing evaluation criteria for comparing two hierarchical annotations of the same recording, the existing methods have known deficiencies when dealing with inexact label matchings and sequential label repetition.

In this article, we investigate methods for automatically enhancing structural annotations by inferring (and expanding) hierarchical information from the segment labels. The proposed method complements existing techniques for comparing hierarchical structural annotations by coarsening or refining labels with variation markers to either collapse similarly labeled segments together, or separate identically labeled segments from each other. Using the multi-level structure annotations provided in the SALAMI dataset, we demonstrate that automatic hierarchy expansion allows structure comparison methods to more accurately assess similarity between annotations.

1. INTRODUCTION

In the music information retrieval (MIR) literature, the problem of *musical structure analysis* broadly concerns methods for automatically inferring relationships between moments in time within a piece [1, 10]. Substantial effort has been expended to develop computational techniques to infer various structures in recorded music, and the existence of reliable reference data and evaluation methodology is critical to accurately assess the efficacy of these methods. More broadly, reliable methods for comparing interpretations of musical structure can be informative for understanding human perception of music [13, 14].

Musical structure can be represented in a variety of ways, depending on the intended use case, ranging from (symbolic) staff notation, to chord annotations, lead sheets, *etc*. MIR research typically focuses on the *segmentation* problem, where the time extent of a recording is partitioned, and each partition is labeled with a descriptor that can be used to indicate repetitions, such as A, B, A, C or *verse, chorus, verse, bridge*. Much of the computational work in this area models musical structure as *flat*, meaning that there is exactly one partitioning of the piece, and the elements of the partition (*segments*) cannot be merged or subdivided to form larger or smaller structures.

In contrast, there is a long tradition in music theory of modeling music with *hierarchies* that simultaneously represent structure at multiple levels of granularity [2, 3]. Indeed, even when instructed to produce a flat segmentation of a piece, expert annotators will often encode latent hierarchical information by using variation markers in their segment labels, *e.g.*, $A, \ldots, A'$ or *verse, ..., verse_(instrumental)* [9, 12]. Although an annotator's choice of segment label may clearly be informative in these cases, this information is ignored by standard segmentation comparison methods. This owes, primarily, to an inability to directly support multi-level segmentations, which could be used to simultaneously represent both the original and simplified annotation as a coherent structure.

In recent years, there has been increasing interest and progress in developing datasets [8, 12], computational methods [16], representations [7], and evaluation criteria [6] for hierarchically structured music segmentations. However, little attention has been paid to exposing *latent* hierarchical structure encoded by segment labels for use in conjunction with these methods.

1.1 Our contributions

In this work, we develop a method for automatically exposing latent multi-level structure encoded by label similarity in music segmentations. The proposed *automatic hierarchy expansion* method operates by simultaneously contracting similar (but distinct) segment labels, and refining identically labeled (but distinct) segments. The contraction and refinement are combined with the original annotation into a hierarchical annotation, which can then be compared to other hierarchies using existing techniques.

Using the SALAMI dataset as a test case, we demonstrate that the proposed method is effective at identifying similarities across annotations that are not captured by previous methods. Finally, we leverage insights gained in developing the method to explore issues of internal consistency within multi-level structure annotations.

1.2 Preliminaries

For a signal of duration T, we define a (flat) *segmentation* as a function $S : [0, T] \rightarrow V$ where V denotes a set of segment labels, e.g., $V = \{A, B, \ldots\}$. We define a *multi-level segmentation* (or *hierarchy*) as a sequence of segmentations $H = (S_0, S_1, \ldots)$, where S_0 maps to a single label, and subsequent segmentations S_i are ordered from *coarse* to *fine*. We assume that each segmentation S_i maps to a distinct vocabulary. Finally, we say that a hierarchy H is *monotonic* if for every level k, we have that

$$S_k(u) = S_k(v) \quad \Rightarrow \quad S_{k-1}(u) = S_{k-1}(v). \tag{1}$$

2. RELATED WORK

Methods for evaluating musical structure analysis algorithms, or more generally comparing two different (flat) structural annotations, broadly fall into two categories: boundary detection and label agreement. Boundary detection metrics capture the agreement between annotations in localizing moments of time when the piece transitions from one coherent segment to another [15]. Segment boundaries are entirely local phenomena, and the metrics do not attempt to encode any sense of long-term structure in the annotations.

Label agreement metrics, on the other hand, are globally informed, and derive from comparisons between the segment labels applied to short samples, typically 0.1s in duration. The *pairwise* precision metric [4] is defined by determining which time points i and j are both given the same label in the reference annotation, and checking to see if the estimate annotation also gives both time points the same label; the fraction of such pairs of time points determines a precision score. Exchanging the roles of the reference and estimate annotations yields definitions for recall and F_1-score. Similarly, the *normalized conditional entropy* measures [5] quantify agreement in terms of the mutual information between the two annotations.

Although label agreement metrics account for global structure, they have three notable shortcomings. First, they are sensitive to alignment errors: if two annotators are operating at different levels of granularity, this information can be obscured by the evaluation. Hierarchical structure evaluation measures address this by integrating multiple segmentations at different levels of granularity into a single hierarchy, and comparing hierarchies to one another [6].

Second, since they depend on frames in isolation from their surrounding context, label agreement metrics cannot distinguish a long segment A from two short segments aa that cover the same time extent. Typically, practitioners circumvent this issue by reporting boundary detection metrics as well as label agreement metrics, but the interactions between the two types of score are rarely easy to interpret.

Finally, label agreement metrics have no mechanism to exploit similarity encoded within segment *labels*: labels A and A' are considered equally distinct as A and B, even though the annotator is clearly implying some high-level similarity in the first case that is absent in the second. While one could modify such annotations directly and use a flat segment evaluation metric, doing so discards information that could still be useful for comparison purposes.

In this work, we address these three issues by deriving segment hierarchies which are informed by label similarity. Discarding variation markers ($A' \rightarrow A$) allows the evaluation to recover from superficially distinct segment labels, while adding counters ($aa \rightarrow a_0 a_1$) provides a way to distinguish a long segment from sequential repetitions of a short segment. By integrating these two modifications into a hierarchy, along with the original annotation, we preserve all of the information present in the annotation.

2.1 Hierarchical evaluation

The approach taken in this work is based on the L-measure method for multi-level segmentation comparison [6], summarized here for completeness. Given a hierarchy H, a *meet matrix* M is defined by the maximum level at which every pair of time instants (u, v) receive the same label:

$$M[u, v] := \max \{k \mid S_k(u) = S_k(v)\}. \tag{2}$$

The meet matrix induces a partial ordering over pairs of time instants, which is summarized by a set of triplets:

$$A(H) := \{(t, u, v) \mid M[t, u] > M[t, v]\}. \tag{3}$$

Finally, given two hierarchies H^R (the reference) and H^E (the estimate), a precision score is defined by comparing the two triplet sets:

$$\text{L-Precision}\left(H^R, H^E\right) := \frac{\left|A\left(H^R\right) \cap A\left(H^E\right)\right|}{\left|A\left(H^E\right)\right|}. \tag{4}$$

Recall is defined analogously by reversing the roles of reference and estimate, and an F_1-score (hereafter denoted as L-Measure) can be computed by the harmonic mean of precision and recall. The terms *reference* and *estimate* to distinguish between annotations derive from the method's use in comparing algorithm outputs to manual annotations. However, the method can generally compare between different annotations of equal status, *e.g.*, produced by two different human annotators. In this case, the terms *reference* and *estimate* are merely intended to identify the annotators, but not to confer privileged status to either.

Although defined for multi-level segmentations, the L-measure can also be applied to compare flat segmentations by including a vacuous segment S_0 which produces a single segment spanning the entire duration. This results in an evaluation which is similar to the pairwise frame similarity metric [4], differing only in that it compares triples rather than pairs. For consistency across experiments, we will employ this method (with the S_0 segment) when comparing flat segmentations in the remainder of this article.

3. METHODS

This work proposes a method for expanding flat annotations to include both more nuanced and coarser structural information. Such expansions seek to address the three shortcomings of label agreement metrics detailed in Section 2. We also explore the concept of *monotonicity* within the hierarchy resulting from applying our methods to several levels of flat annotations.

3.1 Automatic Hierarchy Expansion

Here, we propose an automatic hierarchical expansion for any 'flat' annotations. Our method expands a flat annotation into a hierarchy with three levels. The first level is a *contraction* of the variation markers. The second level is the original annotation. The third level is a *refinement* of the labels by making each instance of a label unique by adding counters to the label.

For a concrete example, consider Figure 1. The left part of the image shows the flat annotation which is repeated on the right side of the image as the middle level of the hierarchy. The contraction level, shown in green, removed the variation markers of the A repetition. The result is that the contraction part of the hierarchy has two kinds of repetitions instead of three.

The refined level of the hierarchy, shown in blue, has at most one block per line. For clarity, the refinement level is created directly from the contraction level of the hierarchy. For each instance of a label in the contraction level, we append a counter (starting with 0) to form a new label. If instead we had conducted this refinement starting at the middle level, we would have ended up with the annotation labels $\{A0, A'0, B0, B1, B2\}$ instead of $\{A0, A1, B0, B1, B2\}$. Both methods produce equivalent results, but the latter is easier to interpret.

Although the expansion described above is most easily understood when applied to flat inputs, it can also be applied to hierarchical inputs by expanding each level independently and combining the results. An example of this multi-level hierarchy expansion is given in Figure 2. We also note that whenever a contraction (or refinement) leaves the segmentation unchanged, the redundant level is omitted from the expanded hierarchy as it produces no additional content. Note that the expansions of the upper and lower annotations in Figure 2 only have two levels each; this is due to the lack of variation markers within the original annotations, meaning that there is nothing to contract.

3.2 Monotonicity

The L-measure described in section 2.1 hinges upon the definition of the meet matrix M (see eq. (2)), which can be interpreted as measuring the similarity between two time points by the depth in the hierarchy at which they receive the same label. When expanding a flat segmentation S, the result is guaranteed to be a monotonic hierarchy. If $S(u) \neq S(v)$, then the contraction level *may* assign u and v the same label, but the refinement level will not. Conversely,

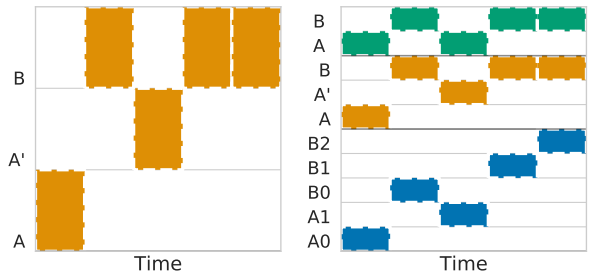

Figure 1. An example of automatic hierarchy expansion. A flat segmentation (left) with segments (A, B, A', B, B) is expanded into a three-level hierarchy (right). The contraction level (green, top) removes variation markers, while the refinement level (blue, bottom) adds counters to each instance of a segment label. The center level (orange) preserves the original annotation.

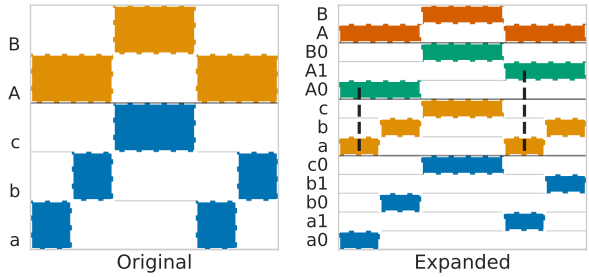

Figure 2. Automatic hierarchy expansion is not guaranteed to preserve monotonicity when applied independently to each level of a hierarchy (left). The dashed lines indicate two instants which receive the same label a in one level, but different labels ($A0$ and $A1$) in a preceding level.

if $S(u) = S(v)$, then the contraction level *must* preserve this equivalence, while the refinement level may not.

However, when applying automatic expansion independently to each level of a hierarchical annotation, the result may not be monotonic. Figure 2 illustrates this effect, where the refinement of the upper level (orange, left plot) results in violations of monotonicity, indicated by the dashed lines in the right plot.

While one could preserve monotonicity by only contracting at the highest level and refining at the lowest level, this is undesirable for three reasons. First, there may be informative structure encoded by variation markers in the intermediate levels which would be missed. Second, annotators may not be internally consistent (*i.e.*, monotonic) from upper to lower-level, so monotonicity would be violated from the start. And finally, because the L-measure depends only on the maximum level of agreement, it does not strictly *require* monotonicity to operate, though the results may be somewhat counter-intuitive. Still, the L-measure definition is most intuitive when the underlying annotations are monotonic, so it is worth investigating the effects of hierarchy expansion on monotonicity.

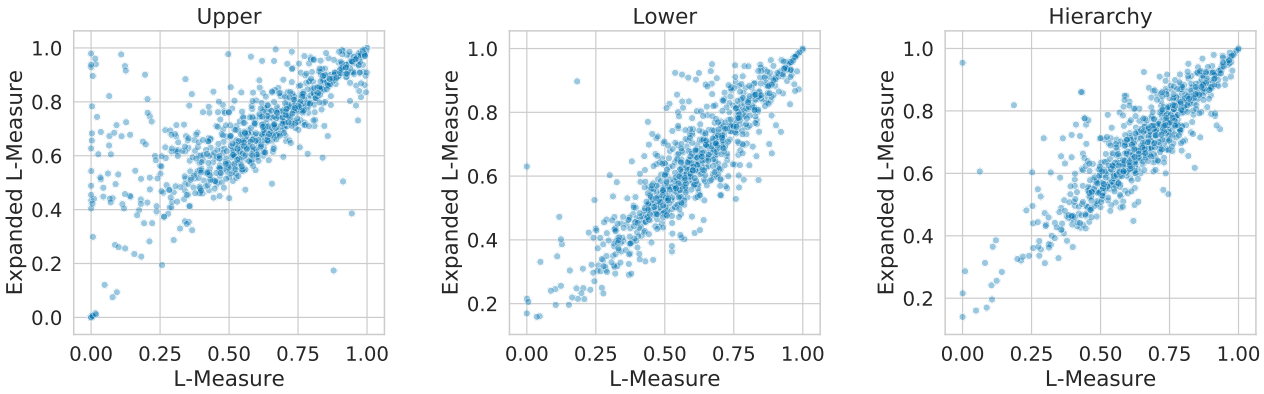

Figure 3. L-measure applied to pairs of annotations in the SALAMI dataset before and after automatic expansion. Left: upper-level annotations; middle: lower-level annotations; right: hierarchical annotations.

4. EXPERIMENTS

We evaluated the effect of automatic hierarchy expansion on flat annotations in the SALAMI dataset [12]. This dataset was selected for two reasons. First, it contains multiple reference annotations (by different annotators). Second, each annotation includes segmentations at different levels of granularity (*upper* and *lower*), which can be treated separately or combined into one hierarchy. All experiments were conducted using the L-measure implementation included in `mir_eval` version 0.5 [11].

4.1 Expansion on flat annotations

For each SALAMI track with two annotators, we first computed the L-measure between the two *upper* annotations. Because neither annotator has a privileged status as reference, we computed the "L-measure" as the harmonic mean of L-precision and L-recall. We then applied the hierarchy expansion procedure to each annotation, and the recomputed the L-measure on the expansions. Comparing the L-measure before and after expansion of a single level allows us to quantify the amount of structural similarity implicitly coded in the segment labels. This process was then repeated for the *lower*-level annotations.

Figure 3 (left and middle plots) summarizes the results of this experiment. As a general trend, expansion has substantial impact on the *upper* level, and less impact on the *lower* level. More specifically, expansion of the upper level produces a change in L-measure of 0.107 ± 0.168 (mean $\pm$ standard deviation), while expansion of the lower-level produces a change of 0.038 ± 0.09. The trend is generally positive at the upper level (with a few exceptions), while the lower-level changes are more symmetric.

Figure 4 illustrates two extreme cases where automatic hierarchy expansion dramatically changes the L-measure between *upper* annotations. [1] In the first case, track 242 improves from 0 to 0.979, because the refinement of the second annotation ($AA \rightarrow A_0A_1$) agrees with the first annotation (AA'), and the contraction of the first annotation

[1] Qualitatively similar examples can be observed for the *lower* annotations, which are omitted here for brevity.

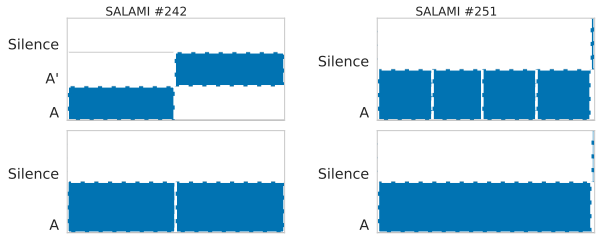

Figure 4. Two extreme examples where hierarchy expansion changes the L-measure from flat (upper) annotations. Left: track 242 increases by $+0.979$; right: track 251 decreases by -0.705.

($AA' \rightarrow AA$) matches with the second annotation. These annotations effectively encode the same information, differing only in the use of variation markers. The second case, track 251, decreased from 0.879 to 0.174 after expansion. This is explained by the first annotation explicitly marking repeated A sections, which are refined into unique sections $(A0, A1, A2, A3)$ by expansion. This structure is absent from the second annotation, which covers the entire duration by a single A segment. Prior to expansion, label-agreement metrics over-estimate the similarity between these annotations. Hierarchy expansion exposes this oversight, resulting in a more accurate comparison.

4.2 Expansion on hierarchical annotations

Extending the analysis of the previous section, we combined each annotator's upper and lower segmentations into a hierarchical annotation H. We then applied hierarchy expansion to each level of the hierarchy, resulting in a new hierarchy H^*. Finally, we computed the L-measure between pairs of hierarchies before and after expansion, which provides a more holistic view of how expansion affects measured agreement between annotators.

The results of expansion comparison for hierarchies are summarized in Figure 3 (right). Overall, the differences are qualitatively similar to the *lower*-level comparison, producing differences in L-measure of 0.048 ± 0.090. While

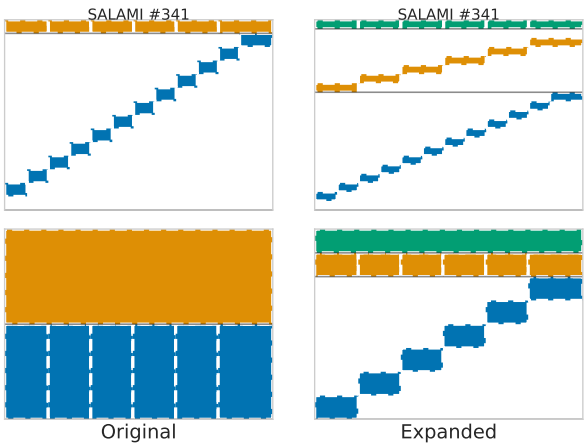

Figure 5. Automatic expansion significantly improves L-measure agreement between two hierarchical annotations of track 341 (increase of $+0.954$). Left: the original hierarchies; right: the expanded hierarchies. Segment labels are suppressed to enhance legibility.

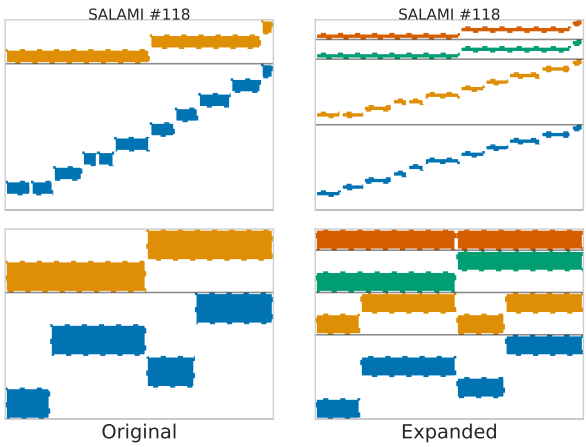

Figure 6. An example where automatic expansion significantly reduces L-measure agreement between two hierarchical annotations (decrease of -0.223). Left: the original hierarchies; right: the expanded hierarchies.

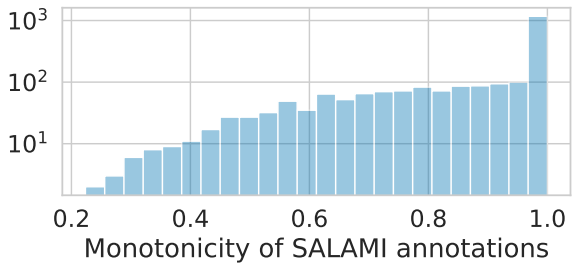

Figure 7. The distribution of monotonicity scores across all segment hierarchies in SALAMI (median: 0.98).

less dramatic than the *upper*-level comparison, the trend is still generally positive, with over 77% of comparisons increasing in value after applying hierarchy expansion. This indicates that even when evaluating with hierarchical annotations, there is still some latent structure encoded in the segment labels which current methods do not account for.

Figure 5 illustrates an example where expansion increases L-measure between two hierarchies, from 0 to 0.954. The second hierarchy (bottom left) assigns the same label to each segment, though the lower level is divided in to repetitions. Expansion separates these repeated segments in both annotations, exposing the common structure shared by both hierarchies (right two subplots).

Figure 6 illustrates the opposite case, where expansion exposes disagreement, decreasing score from 0.841 to 0.618. In this case, looking only at the upper level of the two annotations (left plots, orange level) would indicate considerable agreement between the two annotations, though the lower levels (blue) diverge significantly.

These selected examples are the extreme cases where L-measures deviated the most after hierarchy expansion. In both cases, we find that the divergence can be easily explained by visual inspection, which validates that hierarchy expansion behaves as expected. Note that these cases are relatively unusual, and most changes in scores are much smaller in magnitude. We therefore conclude that hierarchy expansion is effective at recovering from exceptional cases while not detrimentally affecting the common cases.

4.3 Quantifying monotonicity

The experiment described in Section 4.1 started with flat segmentations, and is therefore guaranteed to produce monotonic hierarchies. As noted in Section 3.2, this is not generally true when expanding hierarchies. This raises the question of the importance of monotonicity on hierarchical segmentation evaluation, and whether a given annotator is internally consistent between upper and lower levels.

The definition of monotonicity given in eq. (1) is binary, but it can be relaxed by instead measuring the *proportion* of time instants u and v where agreement at level k implies agreement at level $k-1$. This is calculated exactly by the pairwise recall measure [4], when S_k is treated as the reference and S_{k-1} is the estimate. We thus computed pairwise recall between lower and upper segmentations for each annotation: measures close to 1 are highly monotonic, and lower values indicate violations of monotonicity. The results are summarized by the distribution plot in Figure 7. Overall, the median monotonicity score across all annotations was 0.98, though there appears to be a heavy tail of non-monotonic annotations.

Looking more carefully into the data, we observed that a significant portion of monotonicity violations could be explained by the use of variation markers in the upper-level segmentation. These annotations are specifically problematic because the lower segment a may correspond to distinct upper labels A and A'. More than 60% of hierarchies that do not use variation markers in the upper level are perfectly monotonic, while only 27% of hierarchies with upper variations are monotonic.

Figure 8 illustrates the distribution of monotonicity

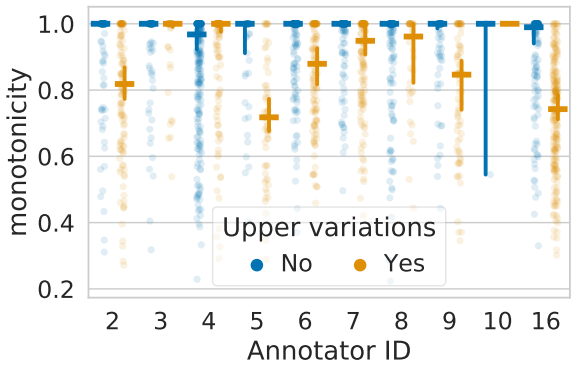

Figure 8. Monotonicity measurements for each SALAMI annotation, grouped by annotator. Median values and 95% bootstrap confidence intervals are indicated by bars.

scores for each individual annotator, sub-divided according to the presence or absence of variation markers in the upper level. Figure 8 shows that use of upper variation markers consistently coincide with lower monotonicity score.

In these experiments, we identified that variation markers can introduce unnecessary differences between sections. What is more, our investigations suggest that the use of variation markers coincides with reduction in monotonicity. The contraction level in the automatic hierarchy expansion seeks to address this issue. However, expanding multiple levels can introduce monotonicity violations. Combining the investigations in this section with the results from Section 4.2, we conclude any new violations created by the contraction of the lower level and the refinement of the upper levels are not substantially detrimental compared to the overall improvements conferred by automatic hierarchy expansion.

4.4 Permutation stability

It is natural to ask whether the previous results are due to introducing multiple hierarchical levels (independent of labels), or if the specific manner in which the contraction and refinement levels are constructed matters. If the effects of hierarchy expansion on L-measure are primarily due to additional levels, but not their specific label structure, we expect that expanding the segmentation with randomly permuted labels should produce comparable results.

To test this idea, we took inspiration from statistical permutation testing, and conducted the following experiment on each level of flat segmentations.

- For each annotation S, construct its hierarchy expansion H and compute the L-recall from S to H.
- (Repeat): randomly permute the labels of S to produce new flat segmentation P, and expand P to new hierarchy H^P. Compute L-recall from S to H^P.

We then compared the distribution of recall scores arising from the (S, H) comparisons to distribution arising from (S, H^P) comparisons. Since the expansion H contains S, the recall score will be identically 1.[2] Note that the expansion H^P will have equivalent refinement level to that of H because each segment is uniquely labeled, so the differences induced by permutation are confined to the middle and upper (contraction) levels.

Level	Mean (original)	Mean (permutation)	KS
Upper	0.992	0.603	0.940
Lower	0.996	0.468	0.977

Table 1. Results of the permutation-expansion test on upper- and lower-level segmentations. *KS* reports the 2-sample Kolmogorov-Smirnov test statistic between expansion and permuted expansion comparisons.

For each annotation, 20 independent permutations were generated. For each level, we report the mean L-recall over original expansions and permuted expansions. We then calculated the 2-sample Kolmogorov-Smirnov test statistic (KS) to determine if the two samples could plausibly be generated from the same underlying distribution. Table 1 summarizes the results of the experiment. In both cases, this null hypothesis was rejected with p-value numerically indistinguishable from 0, indicating that the effects of the hierarchy expansion on the L-measure depend on both the additional hierarchical information and its specific label structure.

5. CONCLUSION

The automatic hierarchy expansion method proposed in this article provides a flexible framework for retaining subtle differences in annotations, while simultaneously exposing coarse similarity. By leveraging ideas from hierarchical structure evaluation, the proposed method is able to illuminate detailed structure latent in the annotations, and recover from problematic edge cases not handled by previous methods. Moreover, the segment refinement technique provides a way to simultaneously evaluate segment boundaries and repetitions, which has been problematic for previous, frame-based evaluations.

Although our focus of the experiments in this work relies on structural segmentation labels in the SALAMI dataset, the general ideas may be applied more broadly, *e.g.*, to the functional section labels used in the Isophonics and TUT annotations [9]. Alternatively, this automatic hierarchy expansion could be applied to other musical concepts where hierarchies naturally occur, such as chord labels (*root, quality, extensions*) or instrumentation (*family, instrument, register*). This could provide a robust alternative evaluation technique for classification problems where adhering to a flat vocabulary is problematic, but where modeling full taxonomies might also be intractable.

[2] When S consists of a single segment spanning the entire duration, it will produce an L-recall of 0. However, we note that the permutation procedure on such annotations will have no effect.

6. REFERENCES

[1] Roger B Dannenberg and Masataka Goto. Music structure analysis from acoustic signals. In *Handbook of Signal Processing in Acoustics*, pages 305–331. Springer, 2008.

[2] Katherine M. Kinnaird. Aligned hierarchies: A multi-scale structure-based representation for music-based data streams. In *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016*, pages 337–343, 2016.

[3] Fred Lerdahl and Ray S. Jackendoff. *A generative theory of tonal music*. MIT press, 1985.

[4] Mark Levy and Mark Sandler. Structural segmentation of musical audio by constrained clustering. *IEEE transactions on audio, speech, and language processing*, 16(2):318–326, 2008.

[5] Hanna M. Lukashevich. Towards quantitative measures of evaluating song segmentation. In *ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008*, pages 375–380, 2008.

[6] Brian McFee, Oriol Nieto, Morwaread M. Farbood, and Juan Pablo Bello. Evaluating hierarchical structure in music annotations. *Frontiers in Psychology*, 8:1337, 2017.

[7] Melissa R. McGuirl, Katherine M. Kinnaird, Claire Savard, and Erin H. Bugbee. SE and SNL diagrams: Flexible data structures for MIR. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, pages 341–347, 2018.

[8] Oriol Nieto and Juan Pablo Bello. Systematic exploration of computational music structure research. In *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016*, pages 547–553, 2016.

[9] Jouni Paulus and Anssi Klapuri. Music structure analysis by finding repeated parts. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, pages 59–68. ACM, 2006.

[10] Jouni Paulus, Meinard Müller, and Anssi Klapuri. State of the art report: Audio-based music structure analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010, Utrecht, Netherlands, August 9-13, 2010*, pages 625–636, 2010.

[11] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel 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 2014, Taipei, Taiwan, October 27-31, 2014*, pages 367–372, 2014.

[12] Jordan B. L. Smith, John 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 the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*, pages 555–560, 2011.

[13] Jordan B. L. Smith and Elaine Chew. Automatic interpretation of music structure analyses: A validated technique for post-hoc estimation of the rationale for an annotation. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, pages 435–441, 2017.

[14] Jordan B. L. Smith, Ching-Hua Chuan, and Elaine Chew. Audio properties of perceived boundaries in music. *IEEE Transactions on Multimedia*, 16(5):1219–1228, Aug 2014.

[15] Douglas Turnbull, Gert R. G. Lanckriet, Elias Pampalk, and Masataka Goto. A supervised approach for detecting boundaries in music using difference features and boosting. In *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, Austria, September 23-27, 2007*, pages 51–54, 2007.

[16] Karen Ullrich, Jan Schlüter, and Thomas Grill. Boundary detection in music structure analysis using convolutional neural networks. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, pages 417–422, 2014.