SYNTHESIZING FLEXIBLE, COMPOSITE HIERARCHICAL STRUCTURE FROM MUSIC DATASETS

Ilana Shapiro
UC San Diego
ilshapiro@ucsd.edu

ABSTRACT

Music is an innately hierarchical system, comprising multiple semantic levels informed by music theory. Such levels include formal structure segmentation, disjoint motif repetition, and harmonic and melodic contour. Historically, researchers in the music information retrieval community have focused on developing analyses for single levels in this hierarchy. However, existing research has addressed neither (1) how to combine arbitrarily many levels of structure analyses into a single unified model and (2) how to extract a representative such structure from a corpus of music, rather than just a single piece. In this work, we propose a novel data structure called the *semantic temporal graph* that captures the semantic (i.e. hierarchical music theoretic) relationships between levels of the hierarchy, as well as the temporal relationships between adjacent-level analyses. Furthermore, given a corpus of such graphs derived from individual pieces, we introduce a method rooted in stochastic optimization to derive a representative $centroid$ graph encoding the music dataset's overall structure. We provide a qualitative evaluation of the semantic temporal graph [where we.....], as well as a quantitative evaluation of the centroid graph [where we...].

1. INTRODUCTION

Music is both composed and comprehended with a hierarchical structure. Individual notes constitute the bottom of the hierarchy, followed by harmony, rhythmic patterns, motives, phrases, and finally large scale sections. Together, this composite hierarchy defines the overall structure of a piece [1].

Automatic identification of musical structure, also known as *music structure analysis* (MSA), continues to be a major interest to both musicologists and the MIR community. Research thus far has focused on the automatic contiguous segmentation (both flat and hierarchical) of musical form [1–13], which involves a boundary detection step followed by a segment labeling step, as well as motif detection [14–16], which looks for disjoint repeating musical patterns. More recently, researchers have also developed avenues for harmonic [17], functional harmonic [18], and melodic [19–21] contour extraction. The techniques used are diverse, ranging from matrix factorization to deep learning in both supervised and unsupervised settings. All of these tasks have been proposed in annual competitions of the Music Information Retrieval eXchange (MIREX) [22–25], which provides a standard format for their outputs.

To our knowledge, all existing MSA research addresses a single aspect of the compositional hierarchy, such as motif extraction, or melodic contour. There is currently no notion of how reconcile differing levels of the hierarchy into a single, unified model of structure, even though their amalgamation is central to a piece's compositional architecture and cohesive integrity. Indeed, Dai et al. have demonstrated empirically that the levels' contents are not formed in isolation, revealing significant interactions different structural levels [13]. In identifying the critical components necessary for integrating the hierarchical levels, we find that there are two central challenges: how to convey each level's semantic, music theoretic level in the hierarchy, and how to encapsulate the temporal relationships between the results of structural analyses at adjacent hierarchical levels.

Furthermore, prior MSA research has only addressed the problem of identifying structure in a single piece, and there is presently no methodology for describing the overall structure of a musical corpus, even across existing single-level analyses. The one exception is Oriol Nieto's proposed technique for merging multiple segment boundary annotations [1], but this is intended to be used with multiple boundary detection algorithms over a single piece to alleviate the problem of subjectivity, and does not address the problem of reconciling differing labels.

1.1 Contributions

To address the first gap, in Section 3, we develop the notion of a *semantic temporal graph* (STG), a k-partite directed acyclic graph (DAG) where the semantic, music theoretic levels of the compositional hierarchy are represented as levels in the k-partite structure, nodes represent structure labels that are the results of the relevant analysis at each level, and edges between nodes of adjacent levels convey the temporal relationships between those structure labels. Each node has an associated time interval determined by the relevant MSA algorithm. A node must have one or two parents at the level above it: one if its associated time interval is a total subset of its parents, and two if its time interval begins in one parent and ends in the other. In order to easily parse the results of MSA algorithms into this data structure, the standard MIREX format is adhered to.

Importantly, the STG is incredibly flexible, and supports the representation of arbitrarily many layers and layer types. Furthermore, the STG is totally decoupled from any specific MSA algorithm or input format, meaning that the chosen MSA algorithm for any level can be easily swapped out, as long as its output adheres to the standard MIREX format. This is crucial as single-level MSA algorithms are constantly improving, and the STG must be adaptable enough to accommodate this.

Finally, to address the second gap, in Section 4 we examine the problem of finding a *centroid*, or most representative, graph given a corpus of the k-partite semantic temporal DAGs derived from individual pieces. We use the label-aware graph edit distance as the similarity metric between two graphs. Given such a set of graphs G, we seek to construct the STG $g*$ that minimizes this distance from $g*$ to every graph in G. This is a constraint satisfaction problem, but one that is intractable to solve deterministically. Thus, we must rely on approximation techniques, and utilize Markov Chain Monte Carlo methods, demonstrating how to use the Metropolis Hastings algorithm to infer an optimal solution and thus arrive at the centroid graph most descriptive of the entire corpus by construction. [1]

2. RELATED WORK

The majority of existing MSA algorithms focus on the contiguous formal segmentation task: boundary detection, followed by section labeling. Segmentation algorithms can be flat or hierarchical (where each level is an increasingly granular contiguous segmentation of the piece). Existing approaches generally utilize matrix factorization, formal grammars, or more recently, neural networks. Oriol Nieto's Music Structure Analysis Framework (MSAF) toolkit [1] features most of the state-of-the art matrix factorization approaches for audio, including ordinal linear discriminant analysis [8], convex nonnegative matrix factorization [5], checkerboard [2], spectral clustering [9], the Structural Features algorithm [3], 2D-Fourier Magnitude Coefficients [6], and the Variable Markov Oracle [4]. These methods use variants of self-similarity matrices (SSMs) for boundary detection, which are symmetric matrices storing pair-wise comparisons between a given set of features. To assign labels, various methods including Gaussian mixture models and nearest neighbor search are employed [1].

More recently, both supervised [26] and unsupervised [12, 27] deep learning approaches to segmentation have studied, as have graph- and grammar-based approaches. Hernandez-Olivan et al. use the graph-based G-PELT algorithm [10], while Dai et al. employ a graph-based approach informed by interactions between sections, melody, harmony and rhythm [13]. In the grammar realm, Finkensiep et al. attempt a unified model of structure by using a minimal context-free grammar to combine repetition with formal prototypes in a tree-based approach, but they still only operate within the contiguous segmentation domain [11].

In the motif extraction domain, algorithms search for disjoint, repeating, and possibly overlapping patterns in a piece. They generally fall into three categories: string-based approaches (e.g. the Variable Markov Oracle [15]), where music data is represented as a one-dimensional pitch sequence and repeated patterns are detected with sub-string matching; geometry-based approaches (e.g. Hsiao et al's BPS-motif discovery algorithm [14]), where music data is represented as multidimensional point sets and *translatable subsets* identify repeating patterns; and feature-based (e.g. [16]) approaches, which learn features from music data, and retrieve patterns with clustering or classification of the features as repeated patterns.

Recent approaches in harmony identification are centered around neural networks, such as using multi-task learning with recurrent neural networks and long short term memory units to detect functional harmonic relationships in a piece [18], as well as developing transformer models to improve chord recognition through incorporating chord segmentation into the recognition process [17]. Until very recently, Justin Salamon's Melodia algorithm was the state of the art in melody extraction. It comprises sinusoid extraction, salience functions, contour creation, and melody selection, and automatically estimates the fundamental frequency corresponding to the pitch of the predominant melodic line of a piece of polyphonic music. Since then, approaches have shifted to neural networks, such as lightweight deep bidirectional LSTM models [20] and transformers [21].

To our knowledge, little work has been done in developing a unified model of structure for either a single piece or across a corpus. The closest we are aware of is Halley Young's *prototype graph*, a bipartite graph that represents musical form as a network of relationships between "prototype nodes" (specific musical elements) and the music they represent, as well as the music-theoretic relationships between musical spans and their respective nodes [28]. However, the prototype graph does not encode hierarchy, and it is limited to a single piece.

3. ABSTRACT REPRESENTATION

3.1 Semantic Temporal Graph

We seek a unified model of musical structure that captures semantic, music theoretic levels of the compositional hierarchy, as well as the temporal relationships between them. The *semantic temporal graph*, or STG, is a novel data structure serving as a meta-representation of musical structure that unifies these objectives. The STG is k-partite directed acyclic graph (DAG). Each of the k layers encodes a level in the semantic hierarchy dictated by music theory. From top to bottom, the music theoretic hierarchy is formed by large scale form, motifs, rhythmic patterns, harmony, and finally melodies constituting spe-

[1] Source code for this project can be found at https://github.com/ilanashapiro/constraints_project

cific note events with a certain timbre quality, duration, and pitch [1]. Individual levels themselves can form sub-hierarchies of increasing granularity, as is commonly seen with the segmentation algorithms corresponding to large scale sections.

In the STG, nodes encode labels from the relevant MSA algorithm. Each node also has an associated time interval given by the algorithm. At each level, these intervals are converted to indices by ordering their start times. Edges encode temporal relationships between nodes of adjacent levels. Specifically, for node n at level i, n must have either one or two parents in level $i-1$: one if its associated time interval is a total subset of its parent's, and two if its time interval begins in one parent's, and ends in the other's.

3.2 Building the Graph

In order to visualize the STG more clearly, we must first establish how it is built. In this paper, we build STGs that unify hierarchical formal segmentation and disjoint motif repetition, with plans in the near future to add layers for rhythm, harmony, and melody. In order to parse the results of any MSA algorithm into a format suitable for the STG, we must first consider the standard MIREX formats they adhere to.

The structural segmentation task was most recently proposed in the 2017 MIREX competition [22]. It defines the following standard output format:

```
<onset_time(sec)>\t<offset_time(sec)>\t<label>\n
<onset_time(sec)>\t<offset_time(sec)>\t<label>\n
...
```

where \t denotes a tab and \n denotes the end of line. The $<$ and $>$ characters are not included. Thus, an example output file could look like:

```
0.000  5.223  A
5.223  15.101 B
15.101 20.334 A
```

Hierarchical segmentations will have multiple of these lists separated by a newline.

MIREX 2017 also dictates the standard output format for motif detection algorithms [23]. Ontimes (in seconds) of notes in the pattern in the left-hand column, with their MIDI note numbers on the right. Each occurrence of a discovered pattern is given before moving on to the next pattern, and occurrences do not have to be of the same length. An example output would resemble:

```
pattern1
occurrence1
7.00000, 45.00000
...
11.00000, 60.00000
occurrence2
31.00000, 57.00000
...
patternM
occurrence1
9.00000, 58.00000
...
occurrencem
100.00000, 62.00000
...
```

Given these standard MIREX formats, we must parse this data into a format suitable for the STG. Each "chunk" of the data, separated by a newline, forms a level in the STG, and each line in a chunk corresponds to a node. Consider Figure 1, which demonstrates the parsing process for a segmentation hierarchy with two levels. We start with the MIREX segmentation output in step 1. In step 2, we number each segmentation level (i.e. text chunk) and extract the segment label and time interval from each line. This gives us the preliminary node label **S**{segment label}**L**{segment level number}**I**{time interval}. Finally, in step 3, we transform each time interval into an index representing that line's position in the level, sorted by interval start time. In the case of segmentation, this matches the line order in the MIREX output. We now have the final segmentation node label **S**{segment label}**L**{segment level number}**N**{label index within level}.

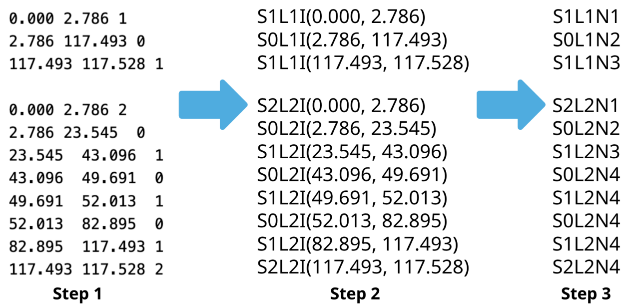

Figure 1. Parsing Formal Segmentation

We repeat this process for motifs as in Figure 2. Motif analyses are single level, i.e. they do not output sub-hierarchies like segmentation can do. Consider the MIREX motif output in step 1. Each pattern (i.e. motif) occurrence's time interval is defined as the ontime of its first note to its last, and we extract pattern and occurrence numbers in step 2 to arrive at the preliminary node label **P**{pattern number}**O**{occurrence number}**I**{time interval}. We then order the preliminary labels by start time, as unlike in segmentation, pattern occurrences are not ordered chronologically in MIREX format. We then assign indices to this ordering to arrive at the final motif node label **P**{pattern number}**O**{occurrence number}**N**{chronologic occurrence index}. This is reflected in step 3, where occurrence 2 of pattern 1 is assigned index 3 instead of 2.

In all cases, the time interval is stored as node metadata, and determines what that node's parent(s) will be.

3.3 STG Examples

In the following examples, we use hierarchical spectral clustering from the MSAF toolkit for segmentation [1, 9], and Hsiao et al's BPS-motif discovery algorithm for motifs [14]. To better visualize the STG's structure, first consider the following subgraph taken from an STG generated for the first movement exposition of the piano transcription of Beethoven's Symphony No. 1, Op. 21 in Figure

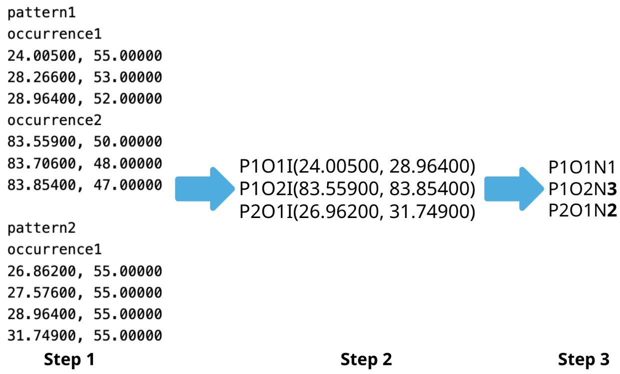

Figure 2. Parsing Motifs

3.[2] The nodes are taken from levels 4 and 5 of the segmentation sub-hierarchy. Node S3L5N2's time interval is

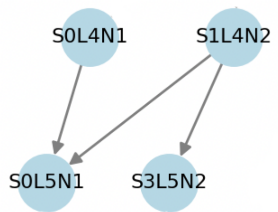

Figure 3. STG Subgraph Example

a total subset of S1L4N2's, while S0L5N1's time interval starts in S0L4N1's, but ends in S2L4N2's. The entire STG generated for this piece, unifying hierarchical formal segmentation with disjoint motif detection, is shown in Figure 4. The segmentation nodes are easily visible, and the motif nodes form the crowded bottom layer.

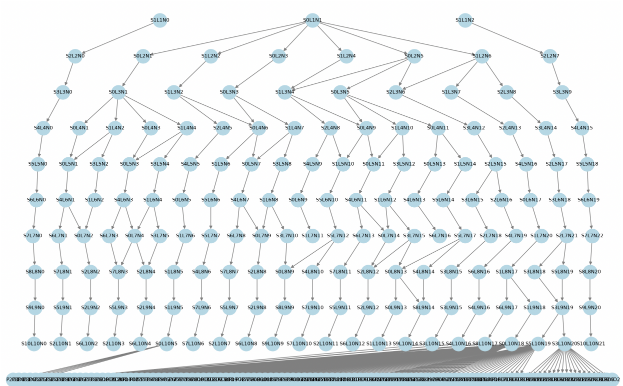

Figure 4. STG Subgraph Example

4. SYNTHESIS

4.1 Distance Metric

4.2 Stochastic Optimization

There remain some issues with this approach that we are continuing to investigate. We are still unsure whether the adaptive nature of the proposal distribution undermines the Markov property of the chain, and therefore convergence of the algorithm. Furthermore, all MCMC algorithms are guaranteed convergence in the infinite limit, but we are unsure whether it is possible to establish when the algorithm converges in practice. For these reasons, we might shift our approach to other optimization strategies, such as manipulating the vector embeddings produced by graph neural networks.

[2] MIDI file is from the LOP database, which can be found here: https://qsdfo.github.io/LOP/database

5. CONCLUSIONS AND FUTURE WORK

In this paper we have presented the semantic temporal graph, a unified meta-representation of complete musical structure that encapsulates the both the intrinsic compositional hierarchy and the temporal relationships linking the elements of adjacent levels. Such a model is essential to comprehending the fusion of the fundamental components that constitute the piece's core architecture, something cannot be done by considering each level in isolation. The STG thus is a vital contribution to the holistic cognition and analysis of musical form.

Furthermore, we have introduced a novel approach rooted in stochastic optimization for deriving an STG representative of a dataset of music. By considering the label-aware edit distance between two STG as a cost function to minimize over the corpus, we are able to describe an application of the Metropolis Hastings algorithm that allows us to construct the "centroid" STG $g*$ for the corpus that minimizes this cost by construction. To our knowledge, this is the inaugural presentation of a method for discerning the overarching structure of an entire dataset, rather than just individual pieces.

However, due to uncertainties we have regarding the convergence of the algorithm, we are interested in exploring graph neural networks as an alternative approach in the near future. We also will add the remaining layers of the compositional hierarchy to the STG; namely, rhythmic contour, functional harmonic contour, and melodic contour, and derive the canonical STG for the relevant dataset using these more robust STGs.

6. REFERENCES

[1] O. Nieto, "Discovering structure in music: Automatic approaches and perceptual evaluations," Ph.D. dissertation, New York University, 2015. [Online]. Available: https://www.proquest.com/openview/09f67403121bcbc7d2ee431985bf0568/1

[2] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*, vol. 1, 2000, pp. 452–455 vol.1.

[3] J. Serrà, M. Müller, P. Grosche, and J. L. Arcos, "Unsupervised music structure annotation by time series

structure features and segment similarity," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1229–1240, 2014.

[4] C.-i. Wang and G. J. Mysore, "Structural segmentation with the variable markov oracle and boundary adjustment," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 291–295.

[5] O. Nieto and T. Jehan, "Convex non-negative matrix factorization for automatic music structure identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013.* IEEE, 2013, pp. 236–240. [Online]. Available: https://doi.org/10.1109/ICASSP.2013.6637644

[6] O. Nieto and J. P. Bello, "Music segment similarity using 2d-fourier magnitude coefficients," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 664–668.

[7] B. McFee, O. Nieto, M. M. Farbood, and J. P. Bello, "Evaluating hierarchical structure in music annotations," *Frontiers in Psychology*, vol. 8, 2017. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2017.01337

[8] B. McFee and D. P. W. Ellis, "Learning to segment songs with ordinal linear discriminant analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014.* IEEE, 2014, pp. 5197–5201. [Online]. Available: https://doi.org/10.1109/ICASSP.2014.6854594

[9] B. McFee and D. Ellis, "Analyzing song structure with spectral clustering," in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, H. Wang, Y. Yang, and J. H. Lee, Eds., 2014, pp. 405–410. [Online]. Available: http://www.terasoft.com.tw/conf/ismir2014/proceedings/T073_319_Paper.pdf

[10] C. Hernandez-Olivan, S. R. Llamas, and J. R. Beltrán, "Symbolic music structure analysis with graph representations and changepoint detection methods," *CoRR*, vol. abs/2303.13881, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.13881

[11] C. Finkensiep, M. Haeberle, F. Eisenbrand, M. Neuwirth, and M. Rohrmeier, "Repetition-structure inference with formal prototypes," in *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, A. Sarti, F. Antonacci, M. Sandler, P. Bestagini, S. Dixon, B. Liang, G. Richard, and J. Pauwels, Eds., 2023, pp. 383–390. [Online]. Available: https://doi.org/10.5281/zenodo.10265305

[12] M. Buisson, B. McFee, S. Essid, and H. C. Crayencour, "Learning multi-level representations for hierarchical music structure analysis." in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, P. Rao, H. A. Murthy, A. Srinivasamurthy, R. M. Bittner, R. C. Repetto, M. Goto, X. Serra, and M. Miron, Eds., 2022, pp. 591–597. [Online]. Available: https://archives.ismir.net/ismir2022/paper/000071.pdf

[13] S. Dai, H. Zhang, and R. B. Dannenberg, "Automatic analysis and influence of hierarchical structure on melody, rhythm and harmony in popular music," *CoRR*, vol. abs/2010.07518, 2020. [Online]. Available: https://arxiv.org/abs/2010.07518

[14] Y. Hsiao, T. Hung, T. Chen, and L. Su, "Bps-motif: A dataset for repeated pattern discovery of polyphonic symbolic music," in *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, A. Sarti, F. Antonacci, M. Sandler, P. Bestagini, S. Dixon, B. Liang, G. Richard, and J. Pauwels, Eds., 2023, pp. 281–288. [Online]. Available: https://doi.org/10.5281/zenodo.10265277

[15] C. Wang, J. Hsu, and S. Dubnov, "Music pattern discovery with variable markov oracle: A unified approach to symbolic and audio representations," in *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, M. Müller and F. Wiering, Eds., 2015, pp. 176–182. [Online]. Available: http://ismir2015.uma.es/articles/78_Paper.pdf

[16] G. Velarde, D. Meredith, and T. Weyde, *A Wavelet-Based Approach to Pattern Discovery in Melodies*. Cham: Springer International Publishing, 2016, pp. 303–333. [Online]. Available: https://doi.org/10.1007/978-3-319-25931-4_12

[17] T. Chen and L. Su, "Harmony transformer: Incorporating chord segmentation into harmony recognition," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, A. Flexer, G. Peeters, J. Urbano, and A. Volk, Eds., 2019, pp. 259–267. [Online]. Available: http://archives.ismir.net/ismir2019/paper/000030.pdf

[18] ——, "Functional harmony recognition of symbolic music data with multi-task recurrent neural networks," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 90–97. [Online]. Available: http://ismir2018.ircam.fr/doc/pdfs/178_Paper.pdf

[19] J. Salamon, E. Gomez, D. P. W. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.

[20] K. Kosta, W. T. Lu, G. Medeot, and P. Chanquion, "A deep learning method for melody extraction from a polyphonic symbolic music representation," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, P. Rao, H. A. Murthy, A. Srinivasamurthy, R. M. Bittner, R. C. Repetto, M. Goto, X. Serra, and M. Miron, Eds., 2022, pp. 757–763. [Online]. Available: https://archives.ismir.net/ismir2022/paper/000091.pdf

[21] Y.-H. Chou, I.-C. Chen, C.-J. Chang, J. Ching, and Y.-H. Yang, "Midibert-piano: Large-scale pre-training for symbolic music understanding," 2021.

[22] M. McCallum. (2017) Structural Segmentation. Music Information Retrieval Evaluation eXchange (MIREX). [Online]. Available: https://www.music-ir.org/mirex/wiki/2017:Structural_Segmentation

[23] T. Collins. (2017) Discovery of Repeated Themes & Sections. Music Information Retrieval Evaluation eXchange (MIREX). [Online]. Available: https://www.music-ir.org/mirex/wiki/2017:Discovery_of_Repeated_Themes_\%26_Sections

[24] Sutashu. (2010) Harmonic Analysis. Music Information Retrieval Evaluation eXchange (MIREX). [Online]. Available: https://www.music-ir.org/mirex/wiki/2010:Harmonic_Analysis

[25] D. Evans. (2021) Audio Melody Extraction. Music Information Retrieval Evaluation eXchange (MIREX). [Online]. Available: https://www.music-ir.org/mirex/wiki/2021:Audio_Melody_Extraction

[26] J. Wang, J. B. L. Smith, W. T. Lu, and X. Song, "Supervised metric learning for music structure features," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. van Kranenburg, and A. Srinivasamurthy, Eds., 2021, pp. 730–737. [Online]. Available: https://archives.ismir.net/ismir2021/paper/000091.pdf

[27] M. C. McCallum, "Unsupervised learning of deep features for music segmentation," *CoRR*, vol. abs/2108.12955, 2021. [Online]. Available: https://arxiv.org/abs/2108.12955

[28] H. Young, M. Du, and O. Bastani, "Neurosymbolic deep generative models for sequence data with relational constraints," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 37 254–37 266. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/f13ceb1b94145aad0e54186373cc86d7-Paper-Conference.pdf