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. 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 organization. Individual notes constitute the bottom of the hierarchy, followed by harmony, rhythmic patterns, motives, phrases, and finally large scale sections. Together, this 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, 15], which looks for disjoint repeating musical patterns. More recently, researchers have also developed avenues for harmonic [16], functional harmonic [17], and melodic [18–20] 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) [21–23], 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. 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.

To address the first gap, in Section 4, we develop the notion of a *semantic temporal graph* (STG), a k-partite directed acyclic graph (DAG) where 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 5 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.

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 [24] and unsupervised [12, 25] 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].

MSA algorithms also extend to motif detection, harmonic analysis, and melody extraction. Motif extraction algorithms, which search for disjoint, repeating, and possibly overlapping patterns in a piece, generally fall into three categories: string-based approaches, where music data is represented as a one-dimensional pitch sequence repeated patterns are detected with sub-string matching; geometry-based approaches, which represents music data as multidimensional point sets

, where music data is represented as multidimensional point sets [14], as well as with string-based methods, such as by the Variable Markov Oracle, which uses a string-based method in which input features are symbolized and repeated suffixes are located within the time series [**?**].

harmony (NN)

melody (salamon paper + NN approaches)

As for a model unifying prototype graph

3. FORMATS AND PARSING

3.1 MIREX Standard Formats

3.2 Parsing

4. ABSTRACT REPRESENTATION

4.1 Semantic Temporal Graph

5. SYNTHESIS

6. CONCLUSIONS AND FUTURE WORK

7. 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. 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

[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.-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.

[16] 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

[17] ——, "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

[18] 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.

[19] 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

[20] 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.

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

[22] 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

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

[24] 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

[25] 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