Automatic Music Structure Analysis: Segmentation and Beyond

Topic Reivew

Qingyang (Tom) Xi

advisor: Dr. Brian McFee

Feburary 2022

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Music Technology

at

New York University

Steinhardt School of Culture, Education, and Human Development

Department of Music and Performing Arts Professions

Introduction

Musical structure is a central matter in the human musical experience, and has received attention from many fields, including Music Theory, Music Cognition, and Music Informatics. Many different methods and styles exist to analyze and communicate the structure of musical works, each analyzing different representations of the musical work, and aiming towards different goals and facilitate different tasks. For example, a Schenkerian analysis of a musical work will often start with the score, focuses on identifying a distilled structure in the form of a harmonically supported falling diatonic line called the Ursatz, and prioritizes communicating the tonal organization of the work for performers or listeners; while an online music streaming service provider might want to perform an analysis on their song catalogue that picks out the most appropriate 30 seconds of any track in the catalogue for consumer previewing, a music structure related task called music thumb-nailing. These different analysis goals and priorities contribute to the subtleties and complexities associated with analyzing musical structures.

Recently, in the field of Music Information Retrieval (MIR), automatic Music Structure Analysis (MSA) has seen considerable development. However, while musical structures can be intricate and deeply nuanced, the MSA tasks in the MIR society are still largely confined to audio segmentation, which merely partitions an audio excerpt into non-overlapping labeled segments. Furthermore, the development of data-driven automatic methods is stymied by subjectivity and bias associated with different annotations and annotators who hold different, but often equally valid, interpretations that reflect their own priorities.

In any case, I believe that successful auditory segmentation is a necessary precursor for music appreciation, understanding, and communication. However, merely identifying the audio segmentation boundaries is far from a satisfactory structural analysis. Output format, system design, and techniques used for MSA are all heavily influences by the context and use cases under which the analysis would be useful. Different tasks like cover song detection, music thumb-nailing, or automatic music generation will directly impact the selection of features, methods, and models for MSA. Across input modalities and downstream tasks, an emerging theme is to produce an explainable segmentation of the musical work that encapsulate much relevant knowledge.

Need for Study

While automatic MSA has seen robust development in the past two-decade, current automatic MSA algorithms still suffers from limited utility, and their outputs lack explainability in general. Current methods cannot resolve ambiguity or disagreement in annotation, and they are unable to meaningfully incorporate multiple interpretations or perspectives. Furthermore, explainability and specificity of MSA outputs are inhibiting their application and utility from impacting more real-world scenarios and use cases. Currently, a typical automatic MSA pipeline generally involves 3 steps: 1) select and compute relevant musical features for each granular time point of the audio or musical score under analysis; 2) compare the features obtained for each time point with every other time point, and collect the result in the Self-Similarity Matrix (SSM); and 3) perform post-processing of the SSM to generate a labeled segmentation of the musical work. My hope is to incorporate two new elements into the typical MSA pipeline to address the lack of explainability and utility of current automatic MSA methods.

Intended Contributions

I would like to improve on the explainability and utility of MSA segmentation outputs by incorporating two ideas:

First, I would like to consider the relationships between any pair of segments. Inspired by Neo-Riemannian theories, a branch of Transformational theory, I wish to shift the focus towards the transition between musical segments, as opposed to their individual identities. While typical MSA algorithms segment music on a larger, formal scale, this idea of focusing on relationships between different segments on a smaller time scale is better articulated by Lewin (2011) in his Transformational theory. In Lewin's case, relationships between musical segments (chords in this case) are called transformations, and take on structural importance that precedes individual chord's position in a tonal hierarchy. Shifting attention away from "the identities of" to "the relationships between" different segments will create a network of relationship that all possible member segments can occupy. One of the frequently referenced graphs in Neo-Riemannian theory is the Tonnetz, where different relationships between triads are represented by pairs of triangular faces in different orientations. More generally, by connecting related musical segments based on their relationships, a graph much like the Tonnetz can be formed to inform the relationships between the segments. Such graphs represent underlying topologies that detail the relationships between segments. This means that by introducing relationships, the previously monotonous segment boundaries could encode more information about the transition.

Secondly, I would like to recognize concurrently unfolding structures. Traditional MSA only produces a single segmentation that represents the structure of the audio excerpt in question. It is well known that the segmentation is heavily influenced by the choice of musical features, and different structures surface accordingly. While traditionally, a choice is made to present a single structure, I would like to promote the view that counterpoints between multiple concurrent structures are equally (if not more) commonplace. Take for example a rhythmic guitar strumming passage over a jagged harmonic progression, interesting structures can be learnt by analyzing either just the rhythmic aspect, or just the chord progression. The counterpoint between the rhythmic and harmonic structures should not be obfuscated by merging the two.

To summarize, I wish to improving the explainability and specificity of automatic MSA on two separate fronts: 1) by incorporating topological relationships that's inherent in musical elements, and 2) by recognizing concurrently unfolding structures and allow for possibility of multiple valid and distinct interpretations or annotations of the same piece of music.

Limitations and Delimitations

The availability of annotated music data limits this study by dictating the genre of audio recording and types of structure that gets analyzed, and perhaps imitated by a data-driven MSA model. Furthermore, these data also dictate the objective evaluation that is typical in assessing a data-driven algorithm.

I'd like to delimitate my research by focusing on aspects of automatic MSA that aims to solve specific problems, for example to facilitate music generation. This research will also not focus deeply on the cognitive aspects of how human perceive musical structures.

Research Question

- Can incorporating topological relationships and analyzing different musical features independently help improve the explainability and utility of automatic music structure analysis? If so, how?

Sub Questions

- Can analysis of individual musical features separately as well as together help explain subjectivity and interpretations in different annotations of musical structure?

- Can we design systems to analyze different musical features of a musical work in order to understand the individual effects and contributions of each musical dimension?

- Can these usb-analysis on individual musical dimensions help explainability of the overall structure produced by the models, or help adapt their utilities to specific situations?

- When analyzing either individual musical features or an aggregate, can considering topological relationships on both the frame level and the section level be used to help specify more nuanced parallelism in different musical aspects?

- How can we establish or discover musically meaningful topologies, both on the frame level and the section level, that will aid analysis and improve explainability of model analysis output?

Literature Review

Autonomous methods to analyze music structure have seen a robust stream of development in the past two decades. Summarizing works that focus on analyzing musical structures either from symbolic representations like MIDI files (Janssen, De Haas, Volk, & Van Kranenburg, 2013), or from signal-based representations like audio recordings (Nieto, 2015; Nieto et al., 2020; Paulus, Müller, & Klapuri, 2010) have detailed state of the art and limitations of automatic MSA.

Goals of Music Structure Analysis

In its most general sense, music structure analysis aims to understand how the constituent parts of a piece of music relate to each other and fit together as a whole. Therefore, when faced with either musical scores or audio recordings, discovering meaningful segments and segmentations of the input media have emerged as a universal step in conducting MSA (Janssen et al., 2013; Nieto et al., 2020). Depending on the context and goal, these segments can be of different time-scales: a single note or an entire verse in a pop song can both be appropriate segments for different goals; they can be non-overlapping as in the two scenarios described above, or they can overlap as in the case of finding segments that are similar to a specific motif or pattern. These segments and segmentations can be analyzed further either individually or comparatively, and can be used creatively.

As a survey to goals of important traditional MSA methods and how they relate to segments, I'll summarize briefly four dominant analysis frameworks in the Music Theory literature: Schenkerian Analysis (Schenker, 2001), the Generative Theory of Tonal Music (Lerdahl & Jackendoff, 1983), the Set-Theoretical framework (Forte, 1973), and neo-Reimmanian analyses (Cohn, 1998). Schenkerian analysis assumes that there is one overarching formal organization called the Ursatz that unifies the pitch and harmonic content of a piece of music, guided by a rather rigid structural formula that emphases the falling diatonic line supported by common practice functional harmony. Here the goal of the analysis is to show a hierarchical segmentation of the symbolic representations (pitches stripped of their duration in this case), while the resulting structure is expected to conform to canonical forms (Schenker, 2001). Similarly hierarchical, the Generative Theory of Tonal Music (GTTM) rejects the idea of preconceived forms like Schenker's Ursatz, and instead employ cognitive principles to generate the hierarchical structure from individual notes, from the bottom up (Lerdahl & Jackendoff, 1983). While both employing hierarchical structure, GTTM and Schenkerian Analysis differ by their biases: GTTM is surface driven, where Schenkerian is conceptual driven. Moving away from from the idea of hierarchical segmentation, both Set-Theory and Neo-Reimmanian theory place more emphases of their analysis goals on finding out important relationships between musical elements. Where GTTM and Schenkerian approaches promote position finding, Set-Theory and Neo-Reimmanian approaches promote pattern matching. In fact, much of the automatic techniques for symbolic music analysis has the specific task of discovering segments of musical patterns, as opposed to the audio segmentation task that is typical for automatic signal-based analysis (Janssen et al., 2013).

More recently, MSA has been proposed to facilitate a range of different audio applications by Nieto et al. (2020), ranging from automatic music generation to music visualization. However the widespread utility of automatic MSA is hindered by the poor explainability of system outputs and difficulty in transferring knowledge between system designed for different tasks.

Segmentation Principles and Methods

Paulus et al. (2010) identified three main principles when segmenting music: homogeneity, novelty, and repetition. Later on, Sargent, Bimbot, and Vincent (2011) employed a fourth principle: regularity of segments. The first three principles proposed by Paulus et al. can be captured by a standard tool in MSA: the self-similarity matrix (SSM)

(Foote, 1999). The SSM records the similarity between any two point in time of some relevant musical feature, resulting in a square matrix. The size of the matrix is determined by the temporal granularity of the small segment possible, typically beats in audio-based MSA, and smaller in symbolic MSA. From the computed SSM of a piece of music, homogeneity will show up as blocks, novelty will show up as edges, and repetition will show up as diagonal lines. Variations of SSM like the lag matrix (Goto, 2003) and the cross recurrence matrix (Serra, Serra, & Andrzejak, 2009) also exist and both improve discovery of repetitions.

Ultimately, the structure discovered by SSM is entirely conditioned on the type of musical feature that is being examined. While hand-crafted musical features reflecting domain knowledge are still effective (McFee & Ellis, 2014), features in the form of self-supervised deep neural network embeddings are start to become widespread (McCallum, 2019; Salamon, Nieto, & Bryan, 2021).

While the SSM provide information on homogeneity, novelty, and repetition of different musical features, it still requires post-processing in general to produce segments. Segments can be produced by convolving a checkerboard kernel along the diagonal of the SSM (Foote, 2000). The relationship and similarities amongst the obtained segments and be further analyzed using techniques such as 2DMFC (Nieto & Bello, 2014). Serrà, Müller, Grosche, and Arcos (2014) achieves segmentation without convolution by shearing the SSM matrix and computing temporal flux. Alternatively, McFee and Ellis (2014) treats the segmentation problem as a clustering of musical point clouds, and applies spectral clustering using the SSM as the laplacian.

Subjectivity and Hierarchy

One of the major challenges for automatic MSA is that musical structures are highly subjective and ambiguous (J. B. L. Smith, 2014). Part of the ambiguity in MSA comes from the hierarchical nature of musical structures, where multiple levels of segmentation of different granularity exist concurrently, much like different levels of a Schenkerian analysis.

McFee, Nieto, Farbood, and Bello (2017) introduces a L-measure that compares all levels of two hierarchical segmentations together. More recently, the L-measure is refined and extended in works by Kinnaird and McFee (2021). Beyond hierarchical ambiguity, multiple different interpretation often exists (Serrà et al., 2014; Wang, Mysore, & Dubnov, 2017), which further complicates automatically evaluating MSA outputs.

Explainability and Independent Musical Dimensions

Related to the subjectivity of human structure analysis is the explainability of the outputs of automatic MSA systems. While explainability of model outputs can be hard to measure quantitatively, J. B. Smith and Chew (2013) propose to study the relevance of individual musical features (e.g., rhythm, timbre, and harmony) have on the overall human production of structure annotations. In that study, Smith and Chew posit that attentiveness to different musical dimensions may explain why two listeners disagreed about a piece's structure, and reiterate the fact that not all musical dimension contribute equally to a structure annotation of musical work.

In fact, paying attention to multiple musical dimensions independently and concurrently are highly valued by music theorists, and is termed Multivalent analysis. Webster (2009), in his book chapter on Multivalent analysis, was able to demonstrate the possibility of different musical dimension each making coherent, yet often different and contrapuntal, structures. Webster emphasizes the need to be context sensitive in performing Multivalent analysis, and concurs with the view that prioritizing different musical dimensions is a major source of subjectivity in interpretation. More recently, younger music theorists like Brody (2016) and Yust (2018) have both embraced Multivalent analysis and achieved new musical insight by looking at separable musical dimensions both individually and concurrently. Based on these trends in both MIR and music theory, I would argue that factoring musical works into separate dimensions and study their individual structure and their contribution to the overall structure is a key to improve automatic MSA system's explainability. By factoring out the analysis into musically meaningful dimensions, this approach should be promising for adapting MSA systems to different use case scenarios as well.

Topological Relationships

One attractive idea in Neo-Reimannian theory is the incorporation of geometry and topologies like the Tonnetz to explain musical similarity and distances of elements. Besides tonnetz, other topological spaces like the ones introduced by Callender, Quinn, and Tymoczko (2008) can also be used to compute similarity metrics, or facilitate descriptions of relationships. Such ideas has already seen their applications in automatic symbolic music analysis (Harte, Sandler, & Gasser, 2006), and the dissertation by Tralie (2017) also showcase the potential of rigorously incorporating topological relationships has on increasing the explainability and utility of audio-based MSA. In his book, Yust (2018) also dedicated 2 chapters on how graph and geometry can be used to represent and reason about temporal structures of different musical dimensions. These trends inspires me to incorporate topological and geometrical relationships in musical analyses, and I believe this added view point of how musical elements transition from one to another in time will be valuable to provide better model output explainability and improve adaptivity to different use cases.

Methodology

Typically, for researches aiming to improve a particular MIR task like MSA, one would often follow these four steps: 1) prepare a suitable dataset, 2) evaluate a existing baseline system on the dataset, 3) propose a new system design, and 4) evaluate the proposed system. While it is straightforward to evaluate the accuracy of MSA outputs against human annotations, it is not immediately clear how to measure explainability and utility of automatic MSA systems quantitatively. Nevertheless, inspired by other works on MSA explainability, I would start to answer my research questions by first exploring the potential of treating musical dimensions individually.

In order to understand the how different musical features or dimensions influence musical structures, it seems a productive first step to study the influences and impacts individual musical features have on human produced music structure annotation in the forms of labeled segments. To understand the individual and combined effects of different musical features, I propose training a model that is capable of scanning the SSM of a particular musical feature and predicting how much contribution this particular feature would have on the overall structure analysis. Such model can be trained with datasets that are already available. Pending on the discovery of this pilot study, I would further investigate whether annotator's attentiveness to different musical dimensions can be predicted via the style and statistics of their structure annotations.

In addition to factoring out separable musical dimensions, I would also consider a more nuanced relationships between identified musical elements and segments by incorporating topology and geometry. These spacial concepts for musical elements are being explored by music theorist and MIR researchers alike, and I believe would be an important tool for improving MSA model explainability and utility. I plan to evaluate the utility of my proposed system by assessing the system's adaptability to different circumstances, and the extent of controllability and variation that it is capable of.

It may seem more straightforward to evaluate the explainability of model output via human subjective studies, but due to the current situation at large, I'm deliberately avoiding this method in order to avoid potential obstacle and uncertainties in progressing my research. Instead, I would like to evaluate model explainability by proxy via the model's ability to emulate different annotators, whose annotations and styles are different. The idea here is that if the proposed model has some mechanism (attentions for different dimensions) to explain why different annotators generate different annotations for the same musical work, then when the model outputs its own structure analysis predictions, one can look at the state or parameter of that mechanism to gain an increased understanding of what drove the model to these decisions.

References

Brody, C. (2016). Parametric interaction in tonal repertoires. *Journal of Music Theory*, *60*(2), 97–148.

Callender, C., Quinn, I., & Tymoczko, D. (2008). Generalized voice-leading spaces. *Science*, *320*(5874), 346–348.

Cohn, R. (1998). Introduction to neo-riemannian theory: a survey and a historical perspective. *Journal of Music Theory*, 167–180.

Foote, J. (1999). Visualizing music and audio using self-similarity. In *Proceedings of the seventh acm international conference on multimedia (part 1)* (pp. 77–80).

Foote, J. (2000). 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, pp. 452–455).

Forte, A. (1973). *The structure of atonal music* (Vol. 304). Yale University Press.

Goto, M. (2003). A chorus-section detecting method for musical audio signals. In *2003 ieee international conference on acoustics, speech, and signal processing, 2003. proceedings. (icassp '03).* (Vol. 5, p. V-437). doi: 10.1109/ICASSP.2003.1200000

Harte, C., Sandler, M., & Gasser, M. (2006). Detecting harmonic change in musical audio. In *Proceedings of the 1st acm workshop on audio and music computing multimedia* (pp. 21–26). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/1178723.1178727` doi: 10.1145/1178723.1178727

Janssen, B., De Haas, W. B., Volk, A., & Van Kranenburg, P. (2013). Discovering repeated patterns in music: state of knowledge, challenges, perspectives. In *Proc. of the 10th international symposium on computer music multidisciplinary research* (Vol. 20, p. 74). Marseille, France.

Kinnaird, K. M., & McFee, B. (2021). Automatic hierarchy expansion for improved structure and chord evaluation. *Transactions of the International Society for Music*

Information Retrieval, *4*(1), 81–92. Retrieved from `http://doi.org/10.5334/tismir.71` doi: 10.5334/tismir.71

Lerdahl, F., & Jackendoff, R. S. (1983). *A generative theory of tonal music.* MIT press.

Lewin, D. (2011). *Generalized musical intervals and transformations.* Oxford University Press, USA.

McCallum, M. C. (2019). Unsupervised learning of deep features for music segmentation. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 346–350).

McFee, B., & Ellis, D. (2014). Analyzing song structure with spectral clustering. In *15th international society for music information retrieval conference.*

McFee, B., Nieto, O., Farbood, M. M., & Bello, J. P. (2017). Evaluating hierarchical structure in music annotations. *Frontiers in psychology*, *8*, 1337.

Nieto, O. (2015). *Discovering structure in music: Automatic approaches and perceptual evaluations* (Unpublished doctoral dissertation). New York University.

Nieto, O., & Bello, J. P. (2014). Music segment similarity using 2d-fourier magnitude coefficients. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 664-668). doi: 10.1109/ICASSP.2014.6853679

Nieto, O., Mysore, G. J., Wang, C., Smith, J. B. L., Schlüter, J., Grill, T., & McFee, B. (2020). Audio-based music structure analysis: Current trends, open challenges, and applications. *Trans. Int. Soc. Music. Inf. Retr.*, *3*(1), 246–263. Retrieved from `https://doi.org/10.5334/tismir.54` doi: 10.5334/tismir.54

Paulus, J., Müller, M., & Klapuri, A. (2010). Audio-based music structure analysis. In *Proc. of the 11th int. society for music information retrieval conference.* Utrecht, Netherlands.

Salamon, J., Nieto, O., & Bryan, N. J. (2021). Deep embeddings and section fusion improve music segmentation. In *Proc. of the 22nd int. society for music information retrieval conf.* Online.

Sargent, G., Bimbot, F., & Vincent, E. (2011). A regularity-constrained viterbi algorithm and its application to the structural segmentation of songs. In *International society for music information retrieval conference (ismir).*

Schenker, H. (2001). *Free composition: New musical theories and fantasies* (E. Oster, Ed.). Pendragon Press. Retrieved from `https://books.google.com/books?id=NpchygEACAAJ`

Serrà, J., Müller, M., Grosche, P., & Arcos, J. L. (2014). Unsupervised music structure annotation by time series structure features and segment similarity. *IEEE Transactions on Multimedia*, *16*(5), 1229-1240. doi: 10.1109/TMM.2014.2310701

Serra, J., Serra, X., & Andrzejak, R. G. (2009). Cross recurrence quantification for cover song identification. *New Journal of Physics*, *11*(9), 093017.

Smith, J. B., & Chew, E. (2013). Using quadratic programming to estimate feature relevance in structural analyses of music. In *Proceedings of the 21st acm international conference on multimedia* (pp. 113–122). New York, NY, USA: Association for Computing Machinery.

Smith, J. B. L. (2014). *Explaining listener differences in the perception of musical structure* (Unpublished doctoral dissertation). Queen Mary University of London.

Tralie, C. J. (2017). *Geometric multimedia time series* (Unpublished doctoral dissertation). Duke University.

Wang, C.-i., Mysore, G. J., & Dubnov, S. (2017, October). Re-Visiting the Music Segmentation Problem with Crowdsourcing. In *Proceedings of the 18th International Society for Music Information Retrieval Conference* (p. 738-744). Suzhou, China: ISMIR. Retrieved from `https://doi.org/10.5281/zenodo.1415944` doi: 10.5281/zenodo.1415944

Webster, J. (2009). The concept of multivalent analysis. In W. E. Caplin, B. Pieter, & J. A. Hepokoski (Eds.), *Musical form, forms & formenlehre: Three methodological reflections* (p. 121–163). Leuven University Press.

Yust, J. (2018). *Organized time: Rhythm, tonality, and form.* Oxford University Press.