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. For example, a Schenkerian analysis of a musical work prioritizes communicating the tonal organization of the work for performers or listeners; while an online music streaming service provider might want to pick out the most appropriate 15 seconds of any track in their catalogue for consumer previewing. These different priorities contribute to the subtleties and complexities associated with analyzing musical structures.

While musical structure can be communicated via many different formats, a common approach for performing Music Structure Analysis (MSA) in the Music Information Retrieval (MIR) community is to segment the musical work into non-overlapping labeled segments. These segments can be the verses, bridges, choruses of a pop song; they can also be different chords in a chord progression. With the emergence of several public datasets of music segmentation annotations in the past decade, data-driven methods have consistently replaced classical methods to become state of the art in accuracy. However, the development of data-driven MSA methods is stymied by data biases and conflicting annotations created by annotators who hold different, but often equally valid, interpretations that reflect their own priorities.

While other MSA tasks like pattern discovery and other representations of musical structure exist, this work will focus on the segmentation task, because I believe successful auditory segmentation is a necessary precursor for music appreciation and communication. That being said, the status quo of merely identifying the audio segmentation boundaries leave much to be desired of a structure analysis. While musical structures can be intricate and deeply nuanced, a labeled segmentation presents only a highly compressed view, and

more explanations as to why a model decided to transition to a new segment, as to merely when, is desired. More interpretable segmentation models could be achieved, for example, by showing how individual musical features influence the model's overall segmentation, or qualify the decision for any of the section transitions by providing similar instances in training data as evidence for each. Explainable segmentation produced by more interpretable MSA models could help explain inter-annotator disagreement, and help with potential adoption of the model for more user specific use case scenarios.

Need for Study

While automatic MSA has seen robust development in the past two decades, automatic MSA tools still have not yet seen heavy user adoption outside of academia, and their outputs lack explainability in general. 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. Current methods cannot resolve ambiguity or disagreement in annotation, and they are unable to meaningfully incorporate multiple interpretations or perspectives. Furthermore, the lack of explainability and specificity of MSA outputs are inhibiting their application and utility from impacting more real-world scenarios and use cases.

Intended Contributions

My hope is to incorporate two new elements into the typical MSA pipeline to address the lack of explainability of current automatic MSA methods.

First, I would like to develop methods to understand how each individual feature and the structure thereof influences the final segmentation, and work towards recognizing 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. In the case for segmentation boundaries, an opportunity for explainability might be to show how much each individual feature contributes to a certain decision, or perhaps how the structures of the individual feature themselves influence the overall predicted structure. While traditionally, a choice is made to present a single structure (potentially out of a family of hierarchically consistent structures), I would like to promote the view that counterpoints between multiple concurrent structures (that are not necessarily hierarchically consistent) 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.

Secondly, 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. In Neo-Riemannian theories, the segments are typically harmonies, and the representation for different types of segment transitions are often facilitated by topologies induced by the relationship between consecutive segments. A favorite such topology 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 relationship between segments. By recognizing and identifying different types of segment transitions based on inter-segmental relationship, models would be at a better position to provide supporting examples from their training data that can justify their decisions. It is also remarkable how this idea of relationship between elements is closely related to the concept of kernel learning in mathematics.

Research Question

- How can incorporating topological relationships and analyzing different musical features independently help improve the explainability of automatic music structure analysis?

Sub Questions

- Can analysis of individual musical features separately as well as together help explain predicted segmentation boundaries by a MSA model?
- 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 sub-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 segmentation 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 segmentation can be analyzed further either individually or comparatively, and can be used creatively.

I'll summarize briefly two important analysis frameworks that produces labeled segmentation in the Music Theory literature: Schenkerian Analysis (Schenker, 2001), and the Generative Theory of Tonal Music (Lerdahl & Jackendoff, 1983). 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 emphasizes 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 elevates the importance of rhythm and metrical position when compared to Schenker. Instead of relying on a preconceived concept of canonical Ursatz, GTTM employ cognitive principles to generate the hierarchical

structure from individual notes, from the bottom up (Lerdahl & Jackendoff, 1983). In other words, while both employing hierarchical structure, GTTM and Schenkerian Analysis differ by their biases: GTTM is surface driven, where Schenkerian is concept driven.

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. Even though good segmentation could theoretically help each of these tasks, meaningful incorporation would require an improvement in explainability. By providing human understandable explanations to why a model produces a certain segmentation in terms of which features and training data are more influential, adapting the model to different use case scenarios becomes possible by adjusting features and training data to steer the model's behavior.

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 points 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 smallest segment possible, with finer granularity generating a larger matrix. Depending on application, the temporal granularity can be fine or coarse. 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. Figure 1 shows 3 different SSM of a 40 seconds excerpt of Tchaikovsky's Dance of the Sugar Plum Fairy. This excerpt starts with 3 short related phrases elaborating a dominant pedal, before ending with the famous solo celeste arpeggios that leads to recapitulation. Different features give rise to different structures, and that can be seen in the difference between

SSMs of different musical features in figure 1. 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.

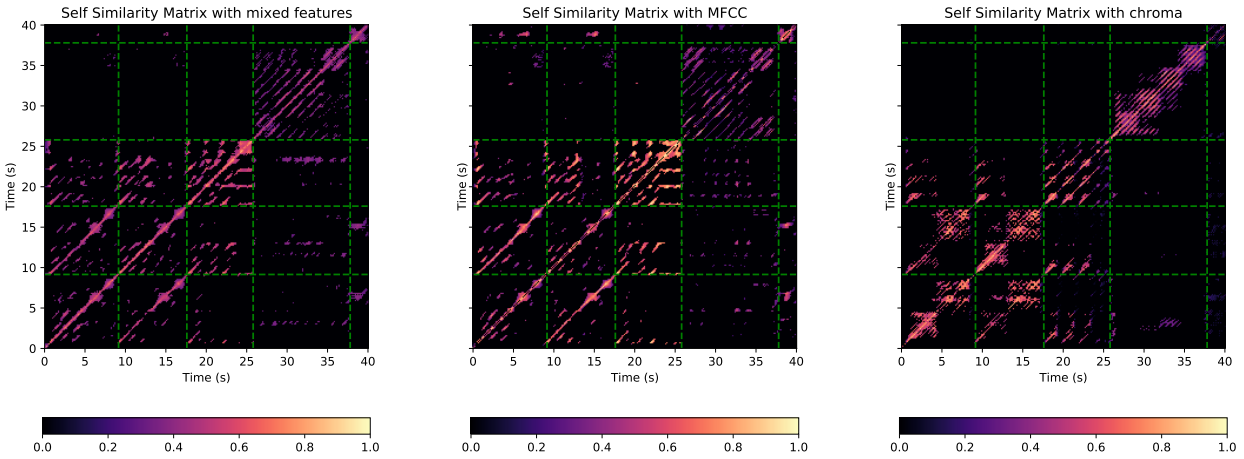

Figure 1

Self Similarity Matrices of an excerpt of the Dance of the Sugar Plum Fairy by Tchaikovsky calculated by using different features. Section transitions are depicted by green dashed lines.

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), which generates a novelty function that could be post-processed, for example by peak picking, to generate segment transitions. 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 infused with temporal connectivity 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 are part of the same family of hierarchically consistent segmentation, 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, which would greatly eliminate the effect of inter-annotator discrepancy due to a difference of segments temporal scale. Beyond hierarchical ambiguity, multiple different interpretation often exists (Serrà et al., 2014; Wang, Mysore, & Dubnov, 2017), which further complicates the desirable behavior of a MSA model when faced with competing interpretations.

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, Molnar (2019) points out several meaningful ways to provide explainability to data-driven models. To summarize, these methods either look at how each individual feature influences the overall prediction, or they look at whether the model can associate key supportive or counterfactual data points to particular decisions, or they could involve understanding the model's internal states or by picking inherently interpretable models like decision trees.

There has been attempts to improve MSA model explainability by looking at individual features. 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 studying their individual structure and their contribution to different structural interpretations 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.

Methodology

I intend to use mixed methods to explore my research questions. Typically, research on MIR tasks like segmentation follow four steps: 1) prepare a suitable dataset, 2) evaluate baseline system on the dataset, 3) propose a new system design, and 4) evaluate the proposed system. However, 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, and then by gaining a more nuanced look at the segment transitions themselves.

To understand the how different musical features or dimensions influence the perception 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 whether that feature is useful to include in the analysis based on its individual structure. 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 more nuanced relationships between identified musical elements and segments.

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.

With a plan to employ both quantitative metrics and qualitative analysis, the goal for this proposed work is to explore the potential impact that individualized attention to different musical dimensions and more nuanced inter-segment relationships can have on improving MSA model explainability.

References

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

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).

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

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.

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

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.

Molnar, C. (2019). *Interpretable machine learning*.

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.

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.