Automatic Analysis and Influence of Hierarchical Structure on Melody, Rhythm and Harmony in Popular Music

Shuqi Dai, Huan Zhang, and Roger B. Dannenberg

Computer Science Department, Carnegie Mellon University
shuqid@cs.cmu.edu, huanz@andrew.cmu.edu, rbd@cs.cmu.edu

Abstract. Repetition is a basic indicator of musical structure. This study introduces new algorithms for identifying musical phrases based on repetition. Phrases combine to form *sections* yielding a two-level hierarchical structure. Automatically detected hierarchical repetition structures reveal significant interactions between structure and chord progressions, melody and rhythm. Different levels of hierarchy interact differently, providing evidence that *structural hierarchy* plays an important role in music beyond simple notions of repetition or similarity. Our work suggests new applications for music generation and music evaluation.

Keywords: Music Structure, Music Understanding, Structure Analysis, Multi-level Hierarchy, Music Segmentation, Pattern Detection, Repetition, Music Similarity

1 Introduction

Form and structure are among the most important elements in music and have been widely studied in music theory. Music structure has a hierarchical organization ranging from low-level motives to higher-level phrases and sections. These different levels influence the organization of other elements such as harmony, melody and rhythm, but these influences are not well formalized. MIR research has developed techniques for detecting music segmentation and repetition structures, but hierarchy is often ignored (Dannenberg & Goto, 2009; Paulus, Müller, & Klapuri, 2010). A fundamental question is whether higher levels of hierarchy are essentially just larger groupings or whether different levels play different roles. If the latter is true, then a better representation and understanding of hierarchy should be useful for prediction, generation, analysis and other tasks. Long-term structure in music is also a recent topic in music generation with deep learning, and attention models such as the Transformer (Vaswani et al., 2017; Huang et al., 2018) seem to improve results. While this suggests some data-driven discovery of structure, results are hard to interpret and, for example, it is not clear whether hierarchy plays a role.

We began our study by developing a method to identify low-level structure in popular songs. Our approach identifies phrases with repetition of harmony,

melody and rhythm. Next, we discovered a simple way to infer higher-level structure from this phrase-level structure. Beyond viewing structure as mere repetition, we show that chord progressions, melodic structures and rhythmic patterns are all related to music structure, and there are significantly different interactions at different levels of hierarchy. Our main contributions are: 1) a novel algorithm to extract repetition structure at both phrase and section levels from a MIDI data set of popular music, 2) formal evidence that melody, harmony and rhythm are organized to reflect different levels of hierarchy, 3) data-driven models offering new music features and insights for traditional music theory. We believe that this work is important in highlighting roles that structure can play in music analysis, music similarity, music generation, and other MIR tasks.

Section 2 discusses related work, and Section 3 presents our phrase-structure analysis method. Section 4 describes general properties of structures we found, while Section 5 explores the relationships between structures and harmony, melody and rhythm, respectively. Finally, Sections 6 and 7 present discussion and conclusions.

2 Related Work

Computational analysis of musical form has long been an important task in Music Information Retrieval (MIR). Large-scale structure in music, from classical sonata form to the repeated structure in pop songs, is essential to music analysis as well as composition. Marsden (2010) implemented Schenkerian analysis and applied it to Mozart variations. Hamanaka, Hirata and Tojo (2014) describe a tool for Generative Theory of Tonal Music (GTTM) analysis that matches closely the analyses of musicologists. Allegraud et al. (2019) use unsupervised learning to segment Mozart string quartets. Go, Ryo, Eita and Kazuyoshi (2019) perform structural analysis using homogeneity, repetitiveness, novelty, and regularity. Our work builds on the idea of extracting structure by discovering repetition.

Identifying hierarchical structure is likely to play a role in music listening. Granroth-Wilding (2013) employs ideas from Natural Language Processing (NLP) to obtain a hierarchical structure of chord sequences. Marsden, Hirata and Tojo (2013) state that advances in the theory of tree structures in music will depend on clarity about data structures and explicit algorithms. Jiang and Müller (2013) propose a two-step segmentation algorithm for analyzing music recordings in predefined sonata form: a thumb-nailing approach for detecting coarse structure and a rule-based approach for analyzing the finer substructure. We present a detailed algorithm for segmenting music into phrases and deriving a higher-level sectional structure starting with a symbolic representation.

Segmentation of music audio is a common MIR task with a substantial literature. Dannenberg and Goto (2009) survey audio segmentation techniques based on repetition, textural similarity, and contrast. Barrington, Chan and Lanckriet (2009) perform audio music segmentation based on timbre and rhythmical properties. However, MIDI has the advantage of greater and more reliable rhythmic information along with the possibility of cleanly separating melody.

Many chord recognition algorithms exist, e.g. Masada and Bunescu (2018) use a semi-Markov Conditional Random Field model. References to melody extraction from MIDI can be found in Zheng and Dannenberg (2019) who use maximum likelihood and Dynamic Programming. Rolland (1999) presents an efficient algorithm for spotting matching melodic phrases, which relates to our algorithm for segmentation based on matching sub-segments of music. Lukashevich (2008) proposes a music segmentation evaluation measure considering over- and under-segmentation. Collins, Arzt, Flossmann and Widmer (2013) develop a geometric approach to discover inexact intra-opus patterns in point-set representations of piano sonatas. Our work introduces new methods for the analysis of multi-level hierarchy in MIDI and investigates the interplay of structure with harmony, melody and rhythm.

3 Phrase-level Structure Analysis

We introduce a novel algorithm based on repetition and similarity to extract structure from annotated MIDI files. Given input consisting of a chord and melody sequence for each song together with its time signature (obtained from MIDI pre-processing), the algorithm outputs a repetition structure. In this section, we will introduce the design motivation, structure representation, details of the algorithm and some evaluation results.

3.1 Motivation and Representation

We represent the structure of a song with alternating letters and integers that indicate phrase labels and phrase length in measures (all boundaries are bar lines). We indicate *melodic phrases* (where a clear melody is present, mostly a vocal line or a instrument solo) with capital letters and *non-melodic* phrases with lower-case letters. For example, `i4A8B8x4A8B8B8X2c4c4X2B9o2` denotes a structure where `A8` and `B8`, for example, represent different repeated melodic phrases of length 8 measures. The `B9` indicates a near-repetition of the earlier `B8`, but with an additional measure. In addition, `i` indicates an introduction with no melody and `o` is a non-melody ending. `X` and `x` denote extra melodic and non-melodic phrases that have no repetition in the song. (The first and second occurrence of `X2` in the structure do *not* match. We could have labeled them as `D2` and `E2` but `X2` makes these non-matching phrases easier to spot.) Non-melodic phrases such as `c` often refer to a transition or bridge, while `X` usually indicates non-repeating phrases or just inserted measures.

Songs do not have unique structures. Consider a simple song with measures *qrstuvwxqrst*. Here, matching letters mean repeated measures, based on overall similarity of chords, rhythm onset times, and a melodic distance function.

We assume that shorter descriptions are more "natural" (Simon & Sumner, 1968). Therefore, we model structural description as a form of data compression, e.g. we can represent this song more compactly as `ABA` where `A`=*qrst* and `B`=*uvwx*. This description requires us to represent 3 phrase symbols (`ABA`) plus

the descriptions of `A` (*qrst*) and `B` (*uvwx*) for a total of 3 phrases and 8 constituent measures. The description length here is $h \cdot 3 + g \cdot 8$, where g and h are constants that favor fewer phrases and more repetition, respectively. We manually tuned the settings to $h = 1.0$ and $g = 1.3$ after experimenting with the training data. In comparison the representation `A`=*qrstuvwxqrst* has a description length of $h \cdot 1 + g \cdot 12$, which is longer and therefore not as good. Extending this idea, we define *Structure Description Length* (SDL) for a song structure Ω consisting of one or more repetitions of phrases from the set P as

$$SDL(\Omega) = h \cdot |\Omega| + g \cdot \sum_{\forall p \in P} avglen(p) \tag{1}$$

where $avglen(p)$ is the average length of instances of phrase p. (Recall that matching phrases need not be exactly the same length.) Since there are often many possible structure descriptions, SDL allows us to automatically select a preferred one.

3.2 Data Pre-processing

We use a Chinese Pop song MIDI data set consisting of 909 manual transcriptions of audio (Wang et al., 2020). We use key and chord labels from audio in combination with labels automatically derived from MIDI, resolving differences with heuristics to improve the labeling. Our MIDI transcription files have a melody track, simplifying melody extraction, and we quantize melodies to 16^{th}s.

3.3 Algorithm Design

Here, we present an overview of our data and analysis algorithm for repetition-based phrase identification. Due to space limitations, we have posted a more detailed description at `cs.cmu.edu/~music/shuqid/musan.pdf`.

Given a song consisting of melody, a chord analysis, and a time signature, our goal is to determine the phrase structure with the shortest structure description length (SDL). The algorithm consists of: (1) Find all matched phrase pairs (repetitions) of equal-length and non-overlapping song segments of length 4 to 20 measures. (2) Merge matching pairs into sets of matching phrases. If we view each of the phrases in our matched phrase pairs as a node in an undirected graph and add edges between the phrases that are matched, then finding all the sets is equivalent to finding all maximal cliques in this sparse undirected graph. We call these *phrase sets*. (3) Find the best structure minimizing *SDL*. The problem is equivalent to the maximum weighted clique problem in the undirected graph. Since this is an NP-complete problem, we combine dynamic programming, A* search, and heuristics to create a good solution with reasonable efficiency.

Since even long songs have only hundreds of measures and the number of phrase sets grows roughly linearly with song length, computation is feasible. Our full algorithm correctly produced 92% of the human-labeled structures. The average run time of each song on a laptop with a 2.3GHz 8-Core Intel Core-i9 and a 64GB-2667MHz-DDR4 RAM is 345 s, but for 80% of the songs, the average run time is only 21 s.

4 Hierarchical Structure Exploration

In this section, we characterize the lower-level phrase structure and the higher-level section structure we found in our data set.

4.1 Phrase-level Structure statistics

What portion of the song is covered by repetition structure? Figure 1 shows that in most songs, repeated melodic phrases cover 50% to 90% of the whole song.

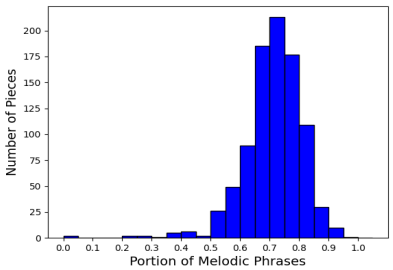

Fig. 1: Distribution of proportion of repeated melodic phrases.

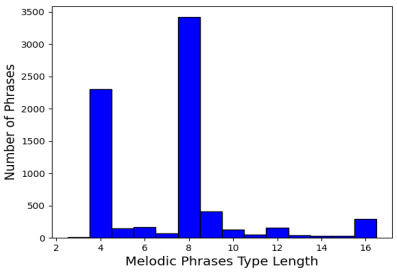

Fig. 2: Distribution of melodic phrase length.

Figure 2 shows the distribution of different phrase lengths among phrases. The majority of melodic phrases have 4 or 8 measures (but we consider longer, higher-level sections below).

4.2 Higher-level Sectional structure

The importance of multi-level hierarchy in music is firmly established. Structure in traditional forms ranges from sub-divisions of beats to movements of symphonies. We looked for automatic ways to detect structures above the level of our "phrases," which are based on repeated patterns. One indication of higher-level structure is the presence of non-matching (`X`) and non-melodic phrases that partition the song structure. In our analysis, successive non-melodic phrases and `X` phrases with total lengths of more than two measures indicate the boundaries of high-level *sections*. For example, the song with structure analysis `i4A8B8x4A8X2B8B8c4c4B9o2`, after removing separator phrases `i4`, `x4`, `c4c4`, and `o2`, has three sections: `A8B8`, `A8X2B8B8` and `B9`. For lack of more standard terminology, we refer to our low-level repetition segments as *phrases* and these higher-level segments as *sections*.

We found that most of the songs have two or three sections (Figure 3), and each section typically has 1 to 6 phrases (Figure 4). Over 90% of songs have two or three distinct *phrases* with melody (e.g. `A`, `B`, ...). Within each section in the song, there are typically one to three distinct melodic phrases (Figure 5).

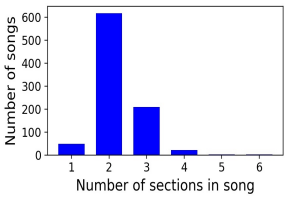

Fig. 3: Distribution of the number of high-level sections in a song.

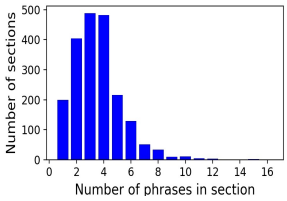

Fig. 4: Distribution of the number of phrases in a section.

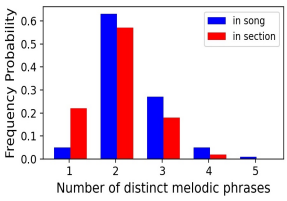

Fig. 5: Distribution of the number of distinct melodic phrases.

Data further shows that 20% of sections are exact repetitions of the previous section in terms of phrases; 29% of the successive sections repeat a suffix of the previous section (e.g. `AAB AB`) while 18% repeat a prefix (e.g. `ABB AB`).

5 Interactions with Segment Structure

We could have used any number of ways to form higher-level structure (sections), but our choice is supported by the finding of interactions between sections, melody, harmony and rhythm that do not occur so strongly at the phrase level, suggesting that the section structure is not just an arbitrary construction.

In Figure 6, we show probabilities of different harmonies at different locations in phrases and sections in major mode. We are much more likely to see a I at the beginning of a phrase and at the end of a section. I and V chords are more popular at the ends of phrases (about equally). We expected to see a predominance of I chords at the ends of phrases, but as the last two categories reveal, the V is a more common ending *within* a section, while the I chord is more common at the *end* of a section. Here, we see significant interactions not only between structure and harmony but between *different structural levels*. We evaluate the significance of these differences by assuming a null hypothesis of equal probability everywhere (the *background* category in Figure 6) and using one-tailed unpaired t-tests. All the test results are significant ($P < .0001$).

Chord *transitions* at the ends of phrases or sections proved to be significantly different from general transition probabilities at other positions in the phrase. For example in major mode, 58% of progressions at the *end* of the section are V–I (Authentic Cadence in music theory). Transition probabilities from V–I at the end of phrase, end of phrase in the middle of a section, and end of section are 0.89, 0.84 and 0.94, while the average transition probability in all other positions is only 0.47.

Phrase and section structures also influence the distribution of melody pitches. Figure 7 shows probabilities of different melody pitches at different locations in phrases and sections, counting only pitches in the context of a I chord in the major mode. Scale degree $\hat{1}$ in melody tends to occur at the *end* of sections, but not at the *start* or *middle* of phrases. While scale degree $\hat{3}$ is common in the start of phrases, but not at the end of sections.

We are also interested in the duration and rhythm of the melody. The distribution of note length at the beginning and middle of phrases is about the same as

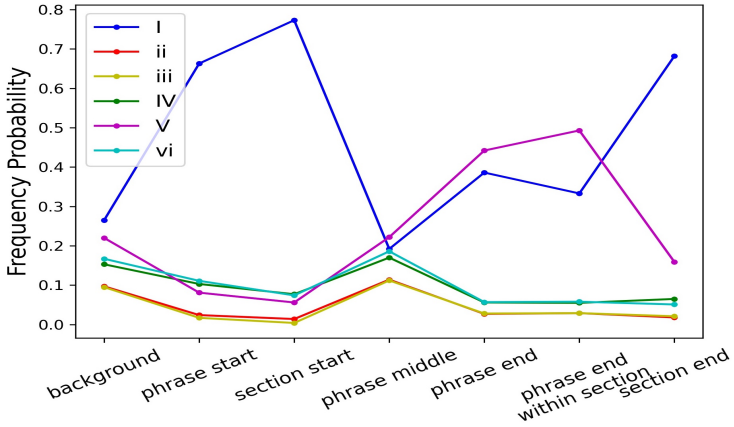

Fig. 6: Chord frequency probability at different level of structure in major mode. X-axis represent different locations in phrase and sections. *Background* means no location constraint, for comparison.

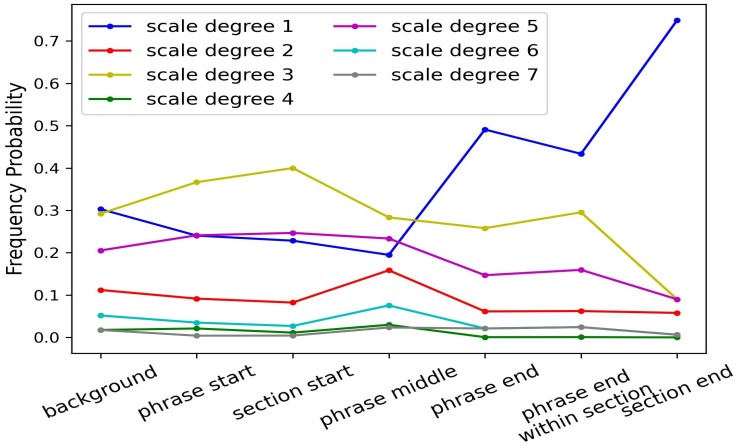

Fig. 7: Melody pitch distribution probabilities conditioned on I chord at different levels of structure in major mode.

the overall distribution, consisting mostly of short notes. In contrast, the phrase endings mostly consist of longer notes. We also observed a difference in phrase endings depending on position. For example, only 6.4% of whole-or-longer notes occur at the ends of phrases in the *middle* of a section, while 72% occur at the ends of phrases at the *end* of a section.

We also discovered two measures of phrase harmonic structure that correlate with year of composition. (Our data spans 7 decades of music.) Although space limits a complete discussion, we found cross-phrase similarity decreases with date (contrast between song sections increases) ($P < .01$), and phrase complexity (in an information theoretical sense) *increases* with date, indicating generally longer phrases and more variety of chord types ($P < .002$).

6 Discussion and Future Direction

The data-driven analysis results in this paper show that music elements such as harmony, melody and rhythm behave differently at different positions relative to the hierarchical music structure. These music-structure-related features support many aspects of traditional music theory. For example, in our analysis, half cadences are more often seen at the ends of phrases, but only in the middle of sections, consistent with the music theoretic concept that a half cadence calls for continuation. It is worth noting that the phrase structure extraction algorithm is fully based on repetition and similarity without using any knowledge of these music concepts. Thus, our approach forms a good test for music theory and existing domain knowledge.

The algorithms we proposed for extracting hierarchical repetition structures from MIDI files have a high accuracy of 93% compared to human labeling, and can be used to analyze other MIDI data sets. Our findings can guide music imitation or generation and can also be used to evaluate whether songs follow structural conventions. Notice that in the phrase-level structure analysis algorithm, parameters are manually tuned, but perhaps they could be adjusted automatically according to different styles of music.

Future work might strive to learn more about variations between similar phrases and how contrasting phrases are constructed. It would also be interesting to compare other data sets, including non-pop music. We have only begun to look for interactions between structure, melody, harmony and rhythm, and these initial results show this to be a promising research direction. The idea that structural tendencies change over time is also promising.

Our results with Chinese pop music are consistent with basic concepts of Western music theory, so we suspect that similar results would be obtained with Western pop music. Still, it would be interesting to conduct a comparative study with Western pop songs. Future work might also investigate more robust indicators of sections. It seems that the non-melodic phrases we use to detect sections are not present in all styles. Consider a repeated form such as `AABA|AABA`. There might be ways to identify these higher-level sections which are not separated by non-melodic phrases.

7 Conclusion

We believe this is the first study to analyze connections between different levels of music structure and the elements of harmony, melody and rhythm using a data-driven approach. We introduced a new hierarchical structure analysis algorithm. With it, we analyzed harmony, melody and rhythm in the context of multi-level structure. This work suggests there is still much to be learned about the role of structure in music, and that we can use hierarchical structure to inform future work on music style, analysis, evaluation and generation.

Our data set, annotations, and experimental results are released at: `https://github.com/Dsqvival/hierarchical-structure-analysis`.

References

Allegraud, P., Bigo, L., Feisthauer, L., Giraud, M., Groult, R., Leguy, E., & Levé, F. (2019, Dec.). Learning sonata form structure on mozartś string quartets. *Transactions of the International Society for Music Information Retrieval*, *2*.

Barrington, L., Chan, A. B., & Lanckriet, G. (2009). Dynamic texture models of music. In *2009 ieee international conference on acoustics, speech and signal processing* (p. 1589-1592).

Collins, T., Arzt, A., Flossmann, S., & Widmer, G. (2013). Siarct-cfp: Improving precision and the discovery of inexact musical patterns in point-set representations. In *Ismir* (pp. 549–554).

Dannenberg, R., & Goto, M. (2009). Music structure analysis from acoustic signals. In *Handbook of signal processing in acoustics* (Vol. 1, p. 305-331). Springer Verlag. doi: 10.1007/978-0-387-30441-0_21

Go, S., Ryo, N., Eita, N., & Kazuyoshi, Y. (2019). Statistical music structure analysis based on a homogeneity-, repetitiveness-, and regularity-aware hierarchical hidden semi-markov model. In *Proceedings of the international symposium on music information retrieval.*

Granroth-Wilding, M. (2013). *Harmonic analysis of music using combinatory categorial grammar* (Unpublished doctoral dissertation). Univ. of Pennsylvania.

Hamanaka, M., Hirata, K., & Tojo, S. (2014). Musical structural analysis database based on gttm. In *Proceedings of the international symposium on music information retrieval.*

Huang, C.-Z., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., ... Eck, D. (2018). Music transformer. *arXiv preprint arXiv:1809.04281*.

Jiang, N., & Müller, M. (2013). Automated methods for analyzing music recordings in sonata form. In *Ismir* (pp. 595–600).

Jiang, Z., & Dannenberg, R. (2019). Melody identification in standard midi files. In *Proceedings of the 16th sound & music computing conference* (p. 65-71).

Lukashevich, H. M. (2008). Towards quantitative measures of evaluating song segmentation. In *Ismir* (pp. 375–380).

Marsden, A. (2010). Recognition of variations using automatic schenkerian reduction. In *Proceedings of the international symposium on music information retrieval.*

Marsden, A., Hirata, K., & Tojo, S. (2013). Towards computable procedures for deriving tree structures in music: Context dependency in gttm and schenkerian theory. In *Sound & music computing conference.*

Masada, K., & Bunescu, R. C. (2018). Chord recognition in symbolic music: A segmental crf model, segment-level features, and comparative evaluations on classical and popular music. *ArXiv*, *abs/1810.10002*.

Paulus, J., Müller, M., & Klapuri, A. (2010). State of the art report: Audio-based music structure analysis. In *Proceedings of the international symposium on music information retrieval* (pp. 625–636).

Rolland, P.-Y. (1999). Discovering patterns in musical sequences. *Journal of New Music Research*, *28*(4), 334-350.

Simon, H., & Sumner, R. (1968). Pattern in music. In B. Kleinmuntz (Ed.), *Formal representation of human judgement* (p. 219-250). Wiley, New York.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Wang, Z., Chen, K., Jiang, J., Zhang, Y., Xu, M., Dai, S., ... Xia, G. (2020). Pop909: A pop-song dataset for music arrangement generation. *arXiv preprint arXiv:2008.07142*.