Unsupervised Melody Segmentation Based on a Nested Pitman-Yor Language Model

Shun Sawada
Future University Hakodate
Hokkaido, Japan
b1012046@gmail.com

Kazuyoshi Yoshii
Kyoto University, Japan
yoshii@kuis.kyoto-u.ac.jp

Keiji Hirata
Future University Hakodate
Hokkaido, Japan
hirata@fun.ac.jp

Abstract

We presents unsupervised melody segmentation using a language model based on a non-parametric Bayesian model. We adapt unsupervised word segmentation with a nested Pitman-Yor language model (NPYLM) used in the field of natural language processing to musical note sequences. Treating music as a language, we aim to extract fundamental units, similar to "words" in natural language, from symbolic musical note sequences using a data-driven approach, the NPYLM. We assume musical note sequences generated by the probabilistic model, integrate a note-level n-gram language model and motif-level n-gram language model, and extract fundamental units (motifs) from them. This enables us to conduct melody segmentation, obtaining a language model for the segments, directly from a musical note sequence without annotation. We discuss the characteristics of this model by comparing the rules and grouping structure of a generative theory of tonal music (GTTM).

1 Introduction

In general, a melody is considered to be time series data of notes with various properties such as pitch and duration. We call this time series data a musical note sequence. In the field of musical information retrieval (MIR), the task of melody segmentation, that is, division of a musical note sequence into meaningful units such as motifs and phrases, is one of the most important and fundamental tasks. Melody segmentation, the division of a musical note sequence into meaningful units such as motifs and phrases, is one of the most important and fundamental tasks in the field of musical information retrieval (MIR). Motifs are considered to be one of the most important and fundamental units of music (Lerdahl and Jackendoff, 1996). If we are able to divide a musical note sequence into appropriate motifs, these motifs can then be used in various tasks such as analyzing a musical structure, automatic composition, and representation learning via "motif embedding" (Hirai and Sawada, 2019).

There are two types of conventional melody segmentation method: rule-based (Lerdahl and Jackendoff, 1996; Cambouropoulos, 2001; Temperley, 2004), and statistic based using properties of musical data (Lattner et al., 2015; Pearce et al., 2010). Although supervised melody segmentation methods have also been proposed (Hamanaka et al., 2017), the cost of producing annotation data of sufficient quality and quantity is enormous. Further, the interpretation of the motifs is subjective and can vary from one annotator to another.

In this study, we aim to extract fundamental units, like "words" in natural language, from symbolic musical note sequences using an unsupervised data-driven approach. It is not known how many notes a motif is made up of, and there are theoretically an infinite number of possible motifs. Therefore, we have to use the vocabulary-free n-gram model instead of the conventional n-gram model, which requires motifs to be defined as a vocabulary in advance. Specifically, we apply an unsupervised word segmentation method, a nested Pitman-Yor language model (NPYLM) (Mochihashi et al., 2009), to a musical note sequence. A sentence in natural language (e.g., English) consists of a combination of words. A word in a natural language consists of a combination of characters. If we think of a character in natural language as equivalent to a musical note and a sentence as equivalent to a note sequence of some length, we can think of the note sequence as consisting of combinations of motifs with units corresponding to words in natural language.

There have been studies that apply the Pitman-Yor language model to music. A hierarchi-

cal Pitman-Yor language model (HPYLM) (Teh, 2006) is an n-gram language model by Pitman-Yor process which is a generalization of a Dirichlet process. A variable-order Pitman-Yor language model (VPYLM) is an extension of a HPYLM that makes it possible to learn an appropriate context length n of an n-gram. Yoshii and Goto (2011) and Nakano et al. (2015) apply a VPYLM to the chord progression. It is thereby possible to learn an appropriate n-gram length for each chord.

2 Unsupervised melody segmentation using nested Pitman-Yor language model

The musical note sequence $\mathbf{s}$ can be expressed as $\mathbf{s} = s_1 s_2 \cdots s_N$ using musical notes s. When the motif is designated as $\mathbf{m}$, melody segmentation is to obtain the motif sequence $\mathbf{s} = \mathbf{m}_1 \mathbf{m}_2 \cdots \mathbf{m}_M$. N is the length of the musical note sequence and M is the number of motifs in the musical note sequence.

When the musical note sequence $\mathbf{s} = s_1 s_2 \cdots s_N$ is given, unsupervised melody segmentation is considered as the problem of finding the motif sequence that maximizes the probability $p(\mathbf{m}|\mathbf{s})$ of the motif sequence $\mathbf{s} = \mathbf{m}_1 \mathbf{m}_2 \cdots \mathbf{m}_M$ obtained by dividing the note sequence. The $p(\mathbf{m}|\mathbf{s})$ can be computed by the language model. The model must calculate probabilities for every possible segmentation of the motif to perform a melody segmentation. Using an n-gram language model with a note-level Pitman-Yor process, we can give probabilities for all possible motif segmentations and thus compute the likelihood of the motifs. We can sample the word segmentation on the basis of this probability.

2.1 Modeling of melody using nested Pitman-Yor language model

In this section, the melody is modeled using a NPYLM that is an n-gram language model based on a hierarchical Pitman-Yor (PY) process. The PY process is a stochastic process that generates a discrete probability distribution G, which is similar to a probability distribution G_0 ($G \sim PY(G_0, d, \theta)$). When we have a uni-gram distribution of motifs G_1, the bi-gram distribution G_2 of motifs will be similar to G_1. Therefore, we can generate G_2 from a PY process of base measure G_1 ($G_2 \sim PY(G_1, d, \theta)$). The uni-gram motif distribution G_1 can be generated as $G_1 \sim PY(G_0, d, \theta)$. The NPYLM is a hierarchical language model in which the note-level HPYLM is embedded as a base measure of the motif-level HPYLM. For details, see (Teh, 2006).

2.2 Unsupervised melody segmentation and training language model

A straightforward method of melody segmentation is to repeat Gibbs sampling, where every note is sampled with the probability of being a motif boundary, and the language model is updated in accordance with the results of that sampling. We used a sentence-wise Gibbs sampler of word segmentation using efficient dynamic programming (Mochihashi et al., 2009). Sampling a new segmentation, we update the NPYLM by adding a new sentence in accordance with the new segmentation. By repeating this process for all musical pieces in a random order, the melody segmentation and language model are alternately optimized.

The musical note sequence is divided into motifs as follows. $\alpha[t][k]$ is the probability of note sequence $s_1 \cdots s_t$ with the final k characters being a motif.

$$\alpha[t][k] = \sum_{j=1}^{t-k} p(s_{t-k+1}^{t} | s_{t-k-j+1}^{t-k}) \cdot \alpha[t-k][j] \tag{1}$$

where $\alpha[0][0] = 1$ and $s_i^j = s_i \cdots s_j$. $p(s_{t-k+1}^{t} | s_{t-k-j+1}^{t-k})$ is obtained by the language model. If $\alpha[t][k]$ can be obtained, we can sample a motif backward. The length of the motif k is sampled from the end of the note sequence to its beginning in accordance with the forward probability $\alpha[t][k]$ (backward sampling). For details, see (Mochihashi et al., 2009).

2.3 Representation of musical note sequence

Musical note sequences can be represented in a number of ways depending on which attributes are used. In this paper, we assume that a musical note in music is like a character in natural language. A melody is considered to be time series data of notes with the properties of various pitches and durations. Therefore, the following representation of musical note sequences with pitch and duration is used.

Pitch-class sequence

A pitch-class sequence considers a melody to be a sequence of pitch classes. The role of each pitch class is assumed to be the same in each key. For

example, the note C in the key of C major and the note D in the key of D major are the same in the sense that they are both the tonic for their key. For this reason, we transpose all the keys to the key of C in advance.

In pitch-class sequences, the octave is ignored, the sharp and the flat are not distinguished, and 12 different symbols are used. There is a total of 13 symbols: 12 symbols for pitch class and 1 for rests.

Pitch-interval sequence

A pitch-interval sequence considers the melody as a sequence of differences between the pitch of the previous note and the current note. We define a pitch-interval sequence on the basis of the assumption that the melody is given meaning by the relative difference in pitch to the previous notes The Implication-Realization (I-R) model (Narmour, 1990), a music theory that classifies and analyzes melodies, ives an abstract of the melody by focusing on the relationship between the pitches of the notes. The intervals are considered up to two octaves above and below ($-24 \leq d_t \leq 24$). Therefore, the resulting number of symbols is 50 (49 + rest symbol).

Duration sequence

The duration sequence is defined as a sequence of durations focusing only on the duration of the notes in a melody. The durations are limited to the length from a thirty-second note up to two whole notes. We are also able to represent dotted notes and tuplets of each note, from thirty-second notes up to whole notes. Rests are treated as a specific symbol with the meaning of a rest.

Compound-representation sequence

The three sequences introduced in the previous section can be combined with one another to form compound representations. First, we combine the pitch-class sequence and duration sequence. Corresponding symbols from the pitch-class sequence and the duration sequence are combined to form a compound representation. We call this the pitch-class and duration sequence (P-D sequence). Second, we combine the pitch-interval sequence and duration sequence. Similarly combining their respective symbols, we form the pitch interval and duration sequence (I-D sequence) Third, the combination of the pitch-class sequence and pitch-interval sequences is called pitch-class and interval sequence (P-I sequence) Finally, we label the combination of all three sequences (pitch-class, pitch interval, and duration) as the P-I-D sequence.

3 Evaluation

In this section, we discuss the characteristics of the melody segmentation obtained with NPYLM, comparing them with the rules and grouping structure of a generative theory of tonal music (GTTM).

3.1 Experimental conditions

To investigate the characteristics of the melody segmentation obtained with NPYLM, we calculate the F-measure for the segments using the ground truth of the grouping structure and each of the rules of the GTTM, although the grouping structure of the GTTM is not necessarily the best for a language model. In this experiment, 300 songs of the GTTM database (Hamanaka) were used as learning data (302 phrases). This dataset consists of monophonic melodies of classical music composed by multiple composers. The total number of notes in the training data set was 12,343, and the average number of notes for each song was 40.9.

The grouping structure of the GTTM represents the cognitive grouping of music experts as they listen to the musical pieces. The sub-rules of the GTTM, the grouping preference rules (GPR), can indicate the candidate boundaries of a group. Each rule does not necessarily coincide with the GTTM grouping structure, but each one mechanically calculates possible boundaries. We compare the segments of the proposed method with the GTTM rules related to the representations of the notes described in Section 3. Specifically, we use GPR 2a, 2b, 3a, and 3d. Given four notes n_1, n_2, n_3, and n_4, each GPR draws a grouping boundary if the relationship between n_2 and n_3 satisfies the following conditions: $rest_{i-1} < rest_i \; and \; rest_i > rest_{i+1}$, where $rest_i$ is the time interval from the beginning to the end of the note **(GPR 2a)**; $ioi_{i-1} < ioi_i \; and \; ioi_i > ioi_{i+1}$, where ioi_i is the inter-onset interval **(GPR 2b)**; $interval_{i-1} < interval_i \; and \; interval_i > interval_{i+1}$, where $interval_i$ is the pitch interval **(GPR 3a)**; $len_{i-1} = 0 \; and \; len_i \neq 0 \; and \; len_{i+1} = 0$, where len_{i-1} is the difference of the duration **(GPR 3d)**;

3.2 Experimental results and discussion

Table 1 shows the F-measure of each representation of the musical note sequence. The row (a)

	Representations			GPR 2a	GPR 2b	GPR 3a	GPR 3d	Grouping
	Pitch	Interval	Duration	(Rest)	(ioi)	(Interval)	(Length)	**Structure**
(a)	✓			8.7	21.6	27.7	9.3	29.4
(b)		✓		6.4	21.9	22.6	8.0	23.6
(c)			✓	6.8	18.7	24.3	17.0	**34.2**
(d)	✓	✓		6.3	18.6	24.9	7.7	21.3
(e)	✓		✓	6.8	21.0	24.6	13.1	28.1
(f)		✓	✓	9.6	19.0	21.8	14.6	24.2
(g)	✓	✓	✓	10.6	17.1	19.5	11.7	21.7

Table 1: F-measure of each representation of the musical note sequence.

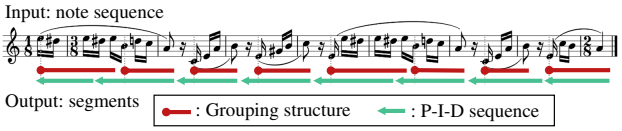

Figure 1: A segmentation result for Bagatelle "Für Elise" WoO.59 (Ludwig van Beethoven) in the GTTM database (No. 3).

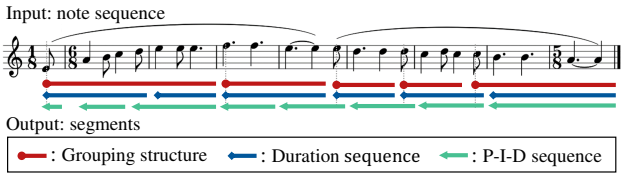

Figure 2: A segmentation result for Má Vlast Moldau (Bedřich Smetana) in the GTTM database (No. 60).

indicates the results of the pitch-class sequence, and the row (d) indicates the results of the P-I sequence. The F-measure for the grouping structure was highest when the duration sequence was used. Figures 1 and 2 show the segmentation results for musical pieces in the GTTM database (No. 3 and No. 60). The lines under the musical score indicate that the notes within the range of the line are in the same grouping.

Regarding GPR 2a, the F-measure was lower than that of the other rules, regardless of which representation was used. The current implementation considers rests to be a special type of note, so distinguishing whether the group boundary is after or before a rest is not possible (see Figure 2). Regarding GPR 3d, F-measure were higher when using a representation related to duration than when using the other representations.

The F-measure for the grouping structure was highest when the duration sequence was used. The grouping structure of the GTTM depends on its metrical structure, such as beats. When the duration sequence is input, we can obtain segments of the rhythmic pattern that occur frequently in a note sequence, because we focus only on the duration, ignoring the pitch completely. In fact, grouping boundaries were drawn more frequently at beat positions when using the duration representation than when using other representations, even though the representation did not explicitly include a metrical structure.

The grouping structure of the GTTM is not necessarily optimal for language models. However, depending on the application, we may have to consider giving information about motifs as prior knowledge and applying semi-supervised learning to obtain the expected melody segmentation. This NPYLM enables semi-supervised melody segmentation.

4 Conclusion

In this study, we performed unsupervised motif segmentation using a Nested Pitman-Yor Language Model. The resulting segments depend on which attributes are used for musical note representation. In the future, we will work on application tasks such as musical structure analysis and representation learning using the obtained segments to verify the usefulness of the segments obtained with the proposed model. We must also consider using other representations, e.g., using abstractions for melody such as I-R model or melodic contour, and explicitly incorporating the metrical structure.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP19J15634.

References

Emilios Cambouropoulos. 2001. The local boundary detection model (LBDM) and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 232—-235.

Masatoshi Hamanaka. Interactive GTTM Analyzer / GTTM Database Download Page. http://gttm.jp/gttm/ja/database/.

Masatoshi Hamanaka, Keiji Hirata, and Satoshi Tojo. 2017. deepgttm-iii: Multi-task learning with grouping and metrical structures. In *International Symposium on Computer Music Multidisciplinary Research*, pages 238–251.

Tatsunori Hirai and Shun Sawada. 2019. Melody2vec: Distributed representations of melodic phrases based on melody segmentation. *Journal of Information Processing*, 27:278–286.

Stefan Lattner, Maarten Grachten, Kat Agres, and Carlos Eduardo Cancino Chacón. 2015. Probabilistic segmentation of musical sequences using restricted boltzmann machines. In *International Conference on Mathematics and Computation in Music*, pages 323–334.

Fred Lerdahl and Ray S Jackendoff. 1996. *A generative theory of tonal music*.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, pages 100–108.

Tomoyasu Nakano, Kazuyoshi Yoshii, and Masataka Goto. 2015. Musical similarity and commonness estimation based on probabilistic generative models. In *2015 IEEE International Symposium on Multimedia (ISM)*, pages 197–204.

Eugene Narmour. 1990. *The analysis and cognition of basic melodic structures: The implication-realization model.*

Marcus T Pearce, Daniel Müllensiefen, and Geraint A Wiggins. 2010. *Melodic grouping in music information retrieval: New methods and applications*. Springer.

Yee Whye Teh. 2006. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992.

David Temperley. 2004. *The cognition of basic musical structures*. MIT press.

Kazuyoshi Yoshii and Masataka Goto. 2011. A vocabulary-free infinity-gram model for nonparametric bayesian chord progression analysis. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 645–650.