Music Structure: finding internal connections for music generation

Huiran Yu

December 2022

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Roger B. Dannenberg, Chair
Daniel Sleator

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science.

Keywords: Stuff, More Stuff

Abstract

Melody prediction is an essential research focus in computer music, aiming to predict melody terms given musical context. It can help people understand how human structures a piece of music to predict the upcoming notes while listening, and also contribute to the melody generation task in automatic composition. Nowadays, most studies only focus on developing new methods to model the musical sequence. However, constructing effective quantitative criteria to measure models' behavior still lacks its demanded attention. In our research, we offer an information entropy metric to test the ability of the standard models, then further combine musical theory with the models to see if we could get better outcomes.

We first established the metric to measure the capability of the baseline models. Each model generated the probability distribution of the terms in the sequence, and we calculated the average information entropy throughout the whole melody. we discovered that stronger models would be more likely to generate lower entropy, which means music is more predictable under these models. We also found models trained on the whole dataset and those trained within the particular song showed drastic differences.

After setting up the baseline, we designed another model recognizing periodic occurrences of notes and pattern, which incorporated music characteristics of fixed phrase length and periodic repetition. This simple model makes satisfying predictions, and with an ensemble strategy considering the entropy value and confidence of each model, we combined the new model with the statistic model reducing the prediction error from 7.5% to 6.5%, eliminated 13% failed cases.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Melody prediction is an essential research focus in computer music, aiming to predict melody terms given musical context. It can help people understand how human structures a piece of music to predict the upcoming notes while listening, and also contribute to the melody generation task in automatic composition. Nowadays, most studies only focus on developing new methods to model musical sequences. However, constructing effective quantitative criteria to measure models' behavior still lacks its demanded attention. In our research, we offer an information entropy metric to test the ability of the standard models, then further combine musical theory with the models to see if we could get better outcomes.

Cross-entropy is a measure of the difference between two probability distributions. It is commonly used in machine learning as a loss function that can be minimized in order to train a model to make more accurate predictions. The cross-entropy loss is calculated by taking the negative logarithm of the predicted probability for the correct class and then summing over all classes. This loss function can be used for a variety of tasks, including classification, sequence prediction, and density estimation. In general, minimizing the cross-entropy loss can help improve the performance of a machine learning model by making its predictions more accurate.

Information entropy is a measure of the uncertainty associated with a random variable. It is a fundamental concept in information theory, and is commonly used to quantify the amount of information that is contained in a message or signal. The information entropy of a random variable is defined as the expected value of the negative logarithm of the probability of the variable. This measure is used to determine the minimum number of bits required to encode a message, and can be used to compare the information content of different messages. In general, the more uncertain or random a variable is, the higher its information entropy will be.

We first established the metric to measure the capability of the baseline models. Each model generated the probability distribution of the terms in the sequence, and we calculated the average information entropy throughout the melody. we discovered that stronger models would be more likely to generate lower entropy, which means music is more predictable under these models. We also found models trained on the whole dataset and those trained within the particular song showed drastic differences.

After setting up the baseline, we designed another model recognizing periodic occurrences of notes and patterns, incorporating music characteristics of fixed phrase length and periodic repetition. This simple model makes satisfying predictions, and with an ensemble strategy considering the entropy value and confidence of each model, we combined the new model with the statistic model reducing the prediction error from 7.5% to 6.5%, eliminating 13% failed cases.

Structure information is vital in music analysis and composition, as it highly relates to people's music perception. In practice, repetition is a primary indicator of structure[25], creating internal references and coherence. In recent years, deep learning models have rapid development in music generation. However, structure information is vastly ignored in deep learning model[35][19] design and evaluation, making it hard for models to generate long-term, self-coherent music. Therefore, establishing a formal foundation of structure in music is highly demanded.

In our recent works, we explored the structure of repetition in music and measured its relationship with entropy and cross-entropy on a Chinese pop song dataset POP909[34] and an American folk song dataset PDSA[2]. We extracted note and rhythm patterns from these datasets, and surprisingly, the patterns that appear in every single song are significantly different from the distribution over the whole dataset. The patterns in one song are more song-specific and build up the entire song from multiple levels of repetition. The entropy result indicated by the variable-order Markov chain[6][7] corresponds to the patterns' statistics. The variable-order Markov chain is a model capable of simultaneously capturing repetition patterns of multiple lengths. The model reaches lower entropy and cross-entropy when predicting a phrase based on other parts of the same song rather than on other songs from the dataset of the same genre. We also find that the predictivity of a song over time represented by the cross-entropy value has a well-managed paradigm in human-composed music pieces. All the observations indicate that to help machines generate coherent, structured music, we need to use methods that are aware of local information and vocabulary and can have better arrangements of the building blocks according to the potential structure of the generated music. These results have been submitted to ISMIR 2022 and got accepted.[8]

Inspired by the fact that the repetition structure in music has already been clearly discovered when using the variable-order Markov model which is entirely statistical without music rules and theories involved, in the next step, we are going to build rule-based predictors for music generation to see whether they will further contribute to the predictability of music. For example, a pattern recognizer that can spot repeated structures in the melody will possibly lower the uncertainty of prediction, or structure-related predictors that take the hierarchy of the song into account should do better when it comes to variations of different sections. We will also conduct formal evaluations of entropy and cross-entropy on these predictors.

This thesis focus on the

Structure is fundamental to music, as seen in the focus on form and analysis in music theory [4, 14, 15], music segmentation [1, 10, 30], structure analysis [13, 17, 25] and chorus detection [26] in MIR research, and recently in the attention given to long-term dependencies in music sequence generation using deep learning techniques [8, 19]. As a basic indicator of music structure, repetition contains important music information. Music relies heavily on repetition to create internal references, coherence and structure.

Unfortunately, the nature of repetition and structure in music is still not well understood, and much remains to be explored with music information retrieval techniques. For example, while music theory may suggest that songs have distinctive motives, our work quantifies and generalizes this notion. We will use "structure" to refer broadly to organizing principles in music, which are generally hierarchical and include sections, phrases and various kinds of patterns. A primary generator of structure is repetition, which includes not just music content within repeat signs but also approximate repetitions at different time scales.

In music generation, many researchers rely on deep learning models to capture music structure and organizing principles implicitly from data. However, repetition, especially long-term repetition structure, does not seem to emerge automatically in deep music generation. Deep learning is a promising direction, but such research should include evaluations where we can assess the successes and failures of new approaches. Moreover, some may argue that we do not need deep learning models to learn prevalent repetitions in music: we can produce repetition simply by generating phrases to the desired length and pasting them into a template. However, we will see that phrase structure, song structure, and other elements of music are intertwined, making this simple approach unable to reproduce characteristics of actual songs. Thus, we need a better understanding of repetition if we want machines to compose or even just listen to music in a more human way.

We aim to use formal models to explore music repetition and structure. By characterizing structural information in music, we can discover new principles of music organization and propose new challenges and evaluation strategies for music information retrieval and music generation.

An essential aspect of repetition structure is hierarchy. We use objective data analysis to support the existence and significance of multiple levels of hierarchy in popular music. We also present a number of results that show strong interactions between structure and pitch, rhythm, harmony, entropy and cross-entropy. Simply stated, structure can help predict pitch (or rhythm, harmony, etc.) and pitch (or rhythm, harmony, etc.) can help predict structure. These findings, in turn, challenge and inform research on deep learning to model hierarchical music structure.

Another important effect of repetition is that song-specific vocabulary of rhythm and pitch patterns is limited relative to what would be expected from the entire dataset. This vocabulary serves both to unify multiple phrases within a song and distinguish songs from others.

The main contribution of this work is a better understanding of the nature of repetition in popular music. We will see that repetition exists in many forms and at different levels of hierarchy. We offer ways to quantify music repetition structure, especially as it relates to pitch and rhythm, often by measuring entropy or cross-entropy. We also reveal striking differences from a structural perspective through case studies on recent deep music generation models. These and other findings offer challenges as well as opportunities for deep-learning music generation and suggest new formal music criteria and evaluation methods.

After describing related work in the next section, we discuss music repetition and structure in Section **??**, how it relates to deep music generation in Section **??**, and we present conclusions in Section **??**.

Chapter 2

Related Works

Repetition is a key element of music structure. Repetition is one of the three commonly used principles for segmenting music, along with novelty at segment boundaries and homogeneity within segments [25]. People have developed a variety of segmentation and section detection methods based on repetition with acoustic features[10, 26]. Repetition becomes especially useful in segmenting symbolic music or lead sheet representations where timbre and texture may be lacking [7].

Repetition also plays an important role in music expectation and prediction [20, 24]. Studies of repetition and structure are common in Music Psychology [22]. For example, listening experiments with reordered Classical and Popular music have shown that listeners are rather insensitive to restructuring, but these results are subtle and somewhat ambiguous [29]. Music form and structure, including repetition, is also a major focus of Music Theory [4, 21, 32].

There are many deep learning models for music generation [3, 11, 19, 27, 28], however, capturing repetition and long-term dependencies in music still remains a challenge. One mainstream approach is to model distribution of music via an intermediate representation (embedding), such as Variational Auto-Encoders (VAE) [28, 35], Generative Adversarial Networks (GANs) [37] and Contrastive Learning [16, 35]. Due to their fixed-length representation and short-length output, it is difficult to exhibit long-term structure. Another popular trend is to use sequential models such as LSTMs and Transformers [19, 27, 33] to generate longer music sequences, but they still struggle to generate repetition and coherent structure on long-term time scales. Some recent work introduces explicit structure planning for music generation, which shows that using structure information leads to better musicality [6, 8, 23].

Current evaluation methods for music generation rarely consider repetition and structure. Deep music generation systems [18, 19] use objective metrics such as negative log-likelihood, cross-entropy and prediction accuracy to compare generated music with ground-truth human-composed music. But these metrics do not precisely correspond to human perception and are not reliable for musicality. Another trend is using domain-knowledge [5] and musical features [9, 12, 31, 36] such as pitch class, pitch intervals, and rhythm density to evaluate music statistically. However, most of them ignore even short patterns, and none evaluate music structure. In contrast, we offer quantitative and objective methods to evaluate music repetition and structure.

Chapter 3

Baseline Models: Variable-order Markov Chain

3.1 Model Deifinitions

3.1.1 Histogram

3.1.2 Dth-order Markov Model

Define Σ as a finite alphabet. Given a sequence $q_1^n = q_1 q_2 \cdots q_n$, $q_i \in \Sigma$, we would like to learn the probability distribution $\hat{P}(s_n|s_1^{n-1})$ for all $s_n \in \Sigma$. Here, s_1^{n-1} represents the prediction context.

Suppose the outcome at current place is only dependent on D previous observations. Then,

$$\hat{P}(s_n|s_1^{n-1}) = \hat{P}(s_n|s_{n-D}^{n-1}) \tag{3.1}$$

In practice, when D becomes larger, this model suffers from the problem of data sparsity, and cannot fully use the subsequence repetitions which are shorter than D in the training data for prediction.

3.1.3 Expectation Networks

3.1.4 Variable-order Markov Model

1. Prediction by Partial Match (PPM)

   To solve the problems of the fixed-order Markov model, we would like to merge Markov models of different orders into one prediction model. We introduce the concept of "escape probability" to merge When we cannot find the higher-order prefix in the training data, we "escape" to the lower order where the length of the prefix will be one shorter.

   We use Prediction by Partial Match (PPM) to implement the variable-order Markov model.

The formalized expressions are

$$\hat{P}(s_n|s_{n-D}^{n-1}) = \begin{cases} \hat{P}(s_n|s_{n-D}^{n-1}), & s_{n-D}^{n} \in \text{ training set} \\ \hat{P}(s_n|s_{n-D+1}^{n-1})\hat{P}(escape|s_{n-D}^{n-1}), & \text{otherwise} \end{cases} \tag{3.2}$$

where:

$$\hat{P}(\sigma|s) = \frac{N(\sigma|s)}{\sum_{\sigma' \in \Sigma(s)} N(\sigma'|s) + |\Sigma(s)|)} \tag{3.3}$$

$$\hat{P}(escape|s) = \frac{|\Sigma(s)|}{\sum_{\sigma' \in \Sigma(s)} N(\sigma'|s) + |\Sigma(s)|} \tag{3.4}$$

Here, $N(\sigma|s)$ is the number of the symbol σ that appears after the context s; $\Sigma(s)$ is the set of symbols that appear after the context s.

2. Exclusion Mechanism

   When we escape to the suffix of the context s, it is no longer necessary to consider the symbols that have already appeared after the s as part of the alphabet, because we have already known that the target symbol σ will never be part of these symbols, and they can be excluded from the probability calculation. With the exclusion mechanism, if we mark the set of the excluded symbols as ϵ, the formula (3.2)-(3.4) will turn into:

$$\hat{P}(s_n|s_{n-D}^{n-1}, \epsilon) = \begin{cases} \hat{P}(s_n|s_{n-D}^{n-1}, \epsilon), & s_{n-D}^{n} \in \text{ training set} \\ \hat{P}(s_n|s_{n-D+1}^{n-1}, \epsilon \cup \Sigma_{n-D}^{n-1}))\hat{P}(escape|s_{n-D}^{n-1}, \epsilon), & \text{otherwise} \end{cases} \tag{3.5}$$

$$\hat{P}(\sigma|s, \epsilon) = \frac{N(\sigma|s)}{\sum_{\sigma' \in \Sigma(s)/\epsilon} N(\sigma'|s) + |\Sigma(s)|)} \tag{3.6}$$

$$\hat{P}(escape|s, \epsilon) = \frac{|\Sigma(s)|}{\sum_{\sigma' \in \Sigma(s)/\epsilon} N(\sigma'|s) + |\Sigma(s)|} \tag{3.7}$$

3.2 Foreground, Background and

Here we define two terms: foreground and background. The foreground information is the contents within the same specific song as the predicted sequence, and the background is the set of other songs with in the dataset. In practice, we randomly shuffled the dataset and separated it into training set and testing set for convenience. The training set is used to train the background model and every song in the testset is trained as the foreground when testing a sequence within the song.

3.3 The confidence of the prediction

When predicting a sequence, we combine the probability outcome from the two models. The baseline combination is linear combination with a mixing ratio α:

$$P_{final}(\sigma|s) = (1-\alpha)P_{foreground}(\sigma|s) + \alpha P_{background}(\sigma|s) \tag{3.8}$$

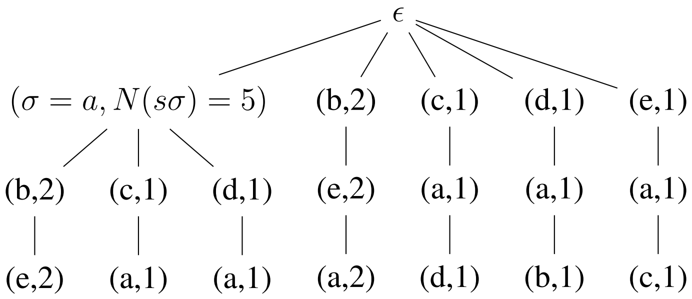

Figure 3.1: Tree constructed by PPM algorithm. abeacadabea

To make fully use of the predicting model, we introduce a confidence parameter C computed from the number of instances used to calculate the probability distribution:

$$C(P(\sigma|s)) = 1 - \frac{1}{\sum_{\sigma' \in \Sigma_s} N(\sigma'|s) + 1}, \sigma \in \Sigma(s) \tag{3.9}$$

When the number of instances equals to zero, the confidence will be zero too. As the number increases, the confidence C will approach one. To make better balance between the two models, we only calculate the confidence of the foreground model since the number of training instances in the background will be significantly larger than the foreground, and the confidence will be so close to one that makes little difference. The merging formula with the confidence parameter is:

$$P_{final}(\sigma|s) = C(P_{foreground}(\sigma|s))(1-\alpha)P_{foreground}(\sigma|s) \tag{3.10}$$

$$+(1 - C(P_{foreground}(\sigma|s))(1-\alpha))P_{background}(\sigma|s) \tag{3.11}$$

Chapter 4

Case Study: Bar-cycle Model

4.1 Definition of the bar-cycle model

One simple but significant feature of music, especially pop music is that the contents of music often repeat after some number of measures, and the repeat period is generally the power of two. The reason behind this is that the music phrases tend to have length of power of two measures, and the the contents are likely to repeat itself throughout the music piece. This characteristics deeply related to the repetition structure in music.

We model this bar-cycle phenomenon into a time-position conditioned first-order Markov model. Suppose we have a pitch sequence $S = [(t_1, s_1), (t_2, s_2), \cdots, (t_N, s_N))]$, where t_i is the onset time of the note, s_i is the pitch of the note. Then,

$$P(s_i|S_1^{i-1}) = P(s_n|t_{i-1}, s_{i-1}) = P(s_i|\hat{t}_{i-1}, s_{i-1}), \tag{4.1}$$

Where $\hat{t}_{i-1} = t_{i-1} \mod \text{len}(cycle)$. Specially, $P(s_0) = P(s_0|\hat{t}_0, \epsilon)$. In our experiments, we tested the model with cycle length of one measure, two measure and four measures. The onset time of the notes are measured in 16-th note.

Chapter 5

Experiments

5.0.1 Dataset

We used two datasets: POP909 and PDSA in our experiments.

POP909 is a Chinese pop song dataset which contains 879 songs in total. The songs ranging from ???? (time) and they are segmented into phrases by [cite]. The dataset is split into training set, validation set and test set of size 529, 175, 175 respectively.

PDSA[2] is the abbreviation of ”Public Domain Songs of American???”, which is a dataset contains 258 public domain American pop songs, folk songs and general classical pieces. The dataset is split into training set, validation set and test set of size 156, 51, 51 respectively.

5.0.2 Experiment Settings

Instead of predicting the target sequences autoregressively with only prefixes training the foreground model, we included both prefix and suffix sequences as the training data. This is based on the consideration that in practice, people like to hear music for multiple times, which means they will already have the impression of the whole picture of the song before they expect the next note to come. From another point of view, different from composing a music piece from scratch, structure and repetition analysis requires the information of the whole song to see the internal connections.

We removed all the repeated phrases to reduce the effect of pure memorization and focus more on the relationships of building materials throughout the song. Then, the pitch sequences were separated into 8-note chunks in each song to make full use of the notes within the same phrase while training the foreground model.

The order of the Markov model and the variable-order Markov model is set to 8. We used the mixing ratio α from 0 to 1, with the step of 0.1. All the notes in the datasets are transposed to C major and are described as scale degree, which means we have seven different types of note in total.

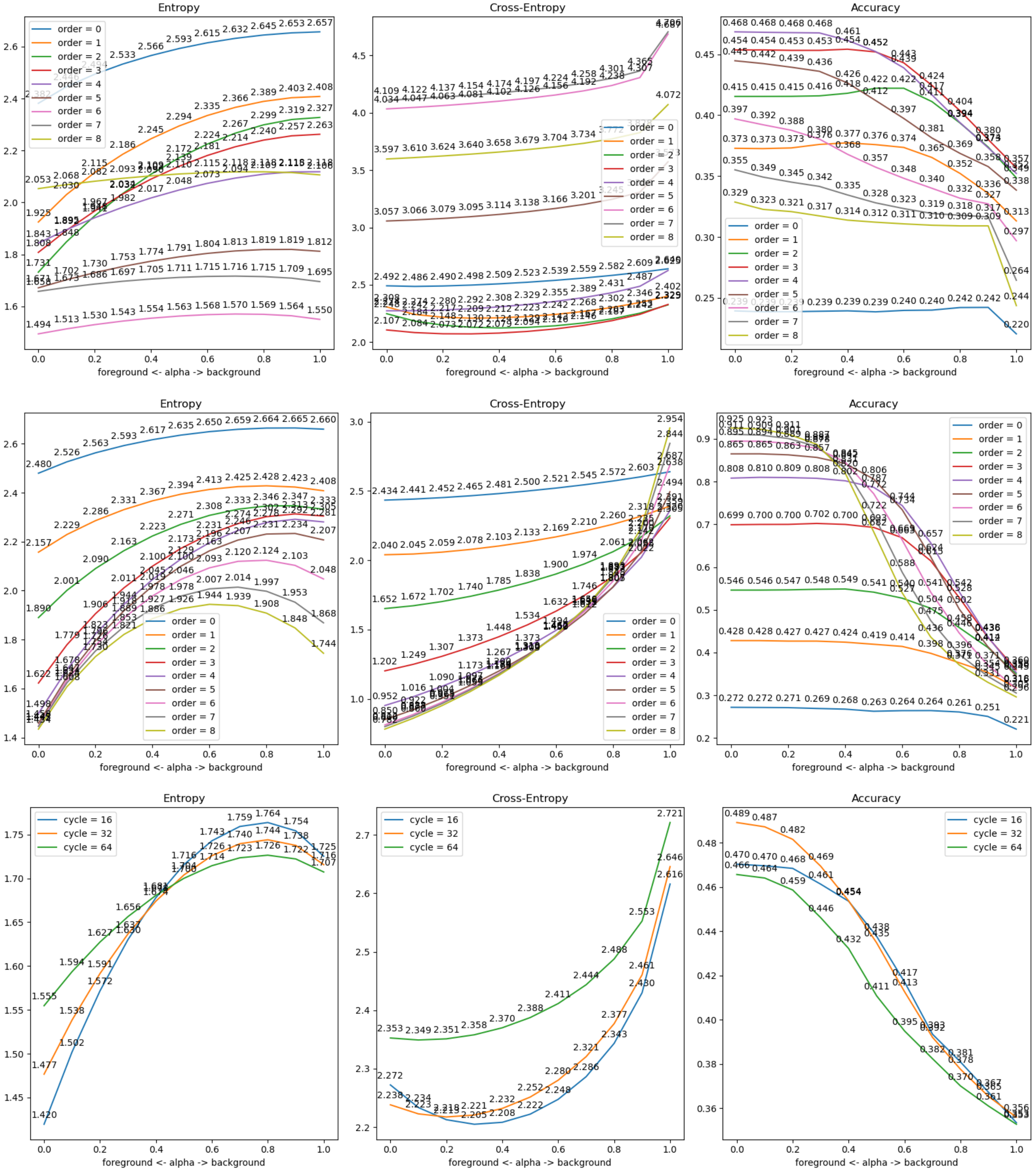

Figure 5.1: POP909

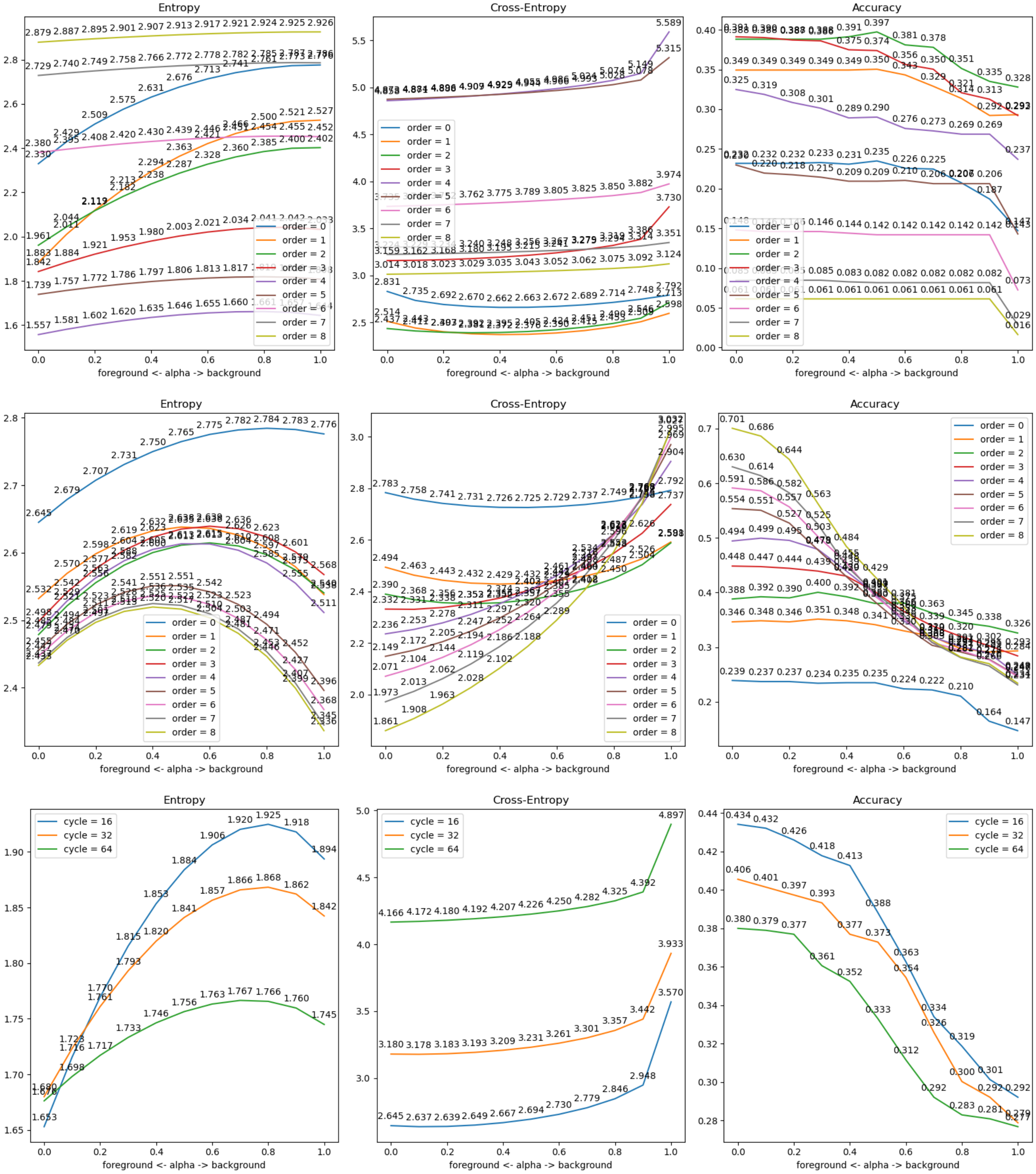

Figure 5.2: PDSA

5.0.3 Prediction result of the single models

We used Markov model, variable-order Markov model and bar-cycle model to predict the two datasets respectively, and also calculated the entropy and cross-entropy of the distribution. The results are given in figure ???. Generally, the variable-order Markov model outperformed the original Markov model and the bar-cycle model, while the first-order bar-cycle model has similar performance as ???-order variable-order Markov model. In the following sections, we are going to analyze the performance of the variable-order Markov model and the bar-cycle model in detail.

Model size difference.

Prediction result analysis of the variable-order Markov model

First of all, the prediction accuracy of the 8th order variable Markov model reaches the prediction probability of 92...%, indicating that

Success cases:

Failed cases:

Prediction result analysis of the bar-cycle model

Success cases:

Failed cases:

5.0.4 Combining bar-cycle model with variable-order Markov model

We can see from the previous analysis that these two different approaches actually represent different aspects of repetition behavior of music. Then another question will be whether there are any ensemble methods to further decrease the error rate based on the results of the two models.

Here, we selected *, *, * of the variable-order Markov model and the entropy and confidence of the distributions generated from the two models as features to predict which distribution should we choose out of the two models. This is a classical binary classification problem and can be solved by an SVM model.

Chapter 6

Conclusion

Bibliography

[1] P. Allegraud, L. Bigo, L. Feisthauer, M. Giraud, R. Groult, E. Leguy, and F. Levé. Learning sonata form structure on mozartś string quartets. *Transactions of the Int. Society for Music Information Retrieval Conf.*, 2, Dec. 2019. 1

[2] David Berger and Chuck Israels. *The Public Domain Song Anthology*. Aperio, Charlottesville, Mar 2020. ISBN 978-1-7333543-0-1. doi: 10.32881/book2. 1, 5.0.1

[3] Jean-Pierre Briot, Gaëtan Hadjeres, and François Pachet. Deep learning techniques for music generation. *Springer*, 10, 2019. 2

[4] William E. Caplan. *Classical Form: A Theory of Formal Functions for the Instrumental Music of Haydn, Mozart, and Beethoven, Revised Edition*. Oxford University Press, 2000. 1, 2

[5] Ching-Hua Chuan and Dorien Herremans. Modeling temporal tonal relations in polyphonic music through deep networks with a novel image-based representation. In *Proc. of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[6] Tom Collins and Robin Laney. Computer-generated stylistic compositions with long-term repetitive and phrasal structure. *Journal of Creative Music Systems*, 1(2), 2017. 2

[7] Shuqi Dai, Huan Zhang, and Roger B. Dannenberg. Automatic analysis and influence of hierarchical structure on melody, rhythm and harmony in popular music. In *in Proc. of the 2020 Joint Conference on AI Music Creativity (CSMC-MuMe 2020)*, 2020. 2

[8] Shuqi Dai, Zeyu Jin, Celso Gomes, and Roger B Dannenberg. Controllable deep melody generation via hierarchical music structure representation. In *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*, 2021. 1, 2

[9] Shuqi Dai, Xichu Ma, Ye Wang, and Roger B. Dannenberg. Personalized popular music generation using imitation and structure. *arXiv preprint arXiv:2105.04709*, 2021. 2

[10] R. B. Dannenberg and M. Goto. Music structure analysis from acoustic signals. *Handbook of Signal Processing in Acoustics*, 1:305–331, 2009. doi: 10.1007/978-0-387-30441-0_21. 1, 2

[11] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 2

[12] Anders Elowsson and Anders Friberg. Algorithmic composition of popular music. In *Proc. of the 12th Int. Conference on Music Perception and Cognition and the 8th Triennial Conf.*

of the European Society for the Cognitive Sciences of Music, pages 276–285, 2012. 2

[13] E. Nakamura G. Shibata, R. Nishikimi and K. Yoshii. Statistical music structure analysis based on a homogeneity-, repetitiveness-, and regularity-aware hierarchical hidden semi-markov model. In *Proc. of the 20th Int. Society for Music Information Retrieval Conf.*, 2019. 1

[14] Percy Goetschius. *Lessons in music form: A manual of analysis of all the structural factors and designs employed in musical composition*, volume 1. Library of Alexandria, 1904. 1

[15] Douglass Marshall Green. *Form in tonal music*. Holt, Rinehart and Winston, 1979. 1

[16] Gaëtan Hadjeres and Léopold Crestel. Vector quantized contrastive predictive coding for template-based music generation. *arXiv preprint arXiv:2004.10120*, 2020. 2

[17] M. Hamanaka, K. Hirata, and S. Tojo. Musical structural analysis database based on GTTM. In *Proc. of the 15th Int. Society for Music Information Retrieval Conf.*, 2014. 1

[18] Cheng-Zhi Anna Huang, Tim Cooijmans, Adam Roberts, Aaron Courville, and Douglas Eck. Counterpoint by convolution. In *Proc. of the 18th Int. Society for Music Information Retrieval Conf.*, Suzhou, China, 2017. 2

[19] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer. *arXiv preprint arXiv:1809.04281*, 2018. 1, 2

[20] David Huron. *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, Cambridge, MA, 2006. 2

[21] Olivier Julian and Christophe Levaux, editors. *Over and Over: Exploring Repetition in Popular Music*. Bloomsbury Academic, 2018. 2

[22] Elizabeth Hellmuth Margulis. *On Repeat: How Music Plays the Mind*. Oxford University Press, 2013. 2

[23] Gabriele Medeot, Srikanth Cherla, Katerina Kosta, Matt McVicar, Samer Abdallah, Marco Selvi, Ed Newton-Rex, and Kevin Webster. Structurenet: Inducing structure in generated melodies. In *Proc. of 19st Int. Conference on Music Information Retrieval Conf., ISMIR*, pages 725–731, 2018. 2

[24] Eugene Narmour et al. *The analysis and cognition of melodic complexity: The implication-realization model*. University of Chicago Press, 1992. 2

[25] O. Nieto, G. J. Mysore, C. C. Wang, J. B. L. Smith, J. Schlüter, T. Grill, and B. McFee. Audio-based music structure analysis: Current trends, open challenges, and applications. *Transactions of the Int. Society for Music Information Retrieval Conf.*, 3(1):246–263, 2020. 1, 2

[26] Jouni Paulus, Meinard Muller, and Ansii Klapuri. Audio-based music structure analysis. In *Proc. of the 11th Int. Society for Music Information Retrieval Conf.*, pages 625–636, 2010. 1, 2

[27] Christine Payne. Musenet. *OpenAI, openai.com/blog/musenet*, 2019. 2

[28] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In *Proc. of the International conference on machine learning*, pages 4364–4373. PMLR, 2018. 2

[29] Jonathan J. Rolison and Judy Edworthy. The role of formal structure in liking for popular music. *Music Perception: An Interdisciplinary Journal*, 29(3):269–284, 2012. 2

[30] Justin Salamon. Deep embeddings and section fusion improve music segmentation. *IEEE Signal Processing Letters*, 24(3):279–283, 2017. 1

[31] Bob L Sturm and Oded Ben-Tal. Taking the models back to music practice: Evaluating generative transcription models built using deep learning. *Journal of Creative Music Systems*, 2:32–60, 2017. 2

[32] Jay Summach. The structure, function, and genesis of the prechorus. *Music Theory Online*, 17(3), October 2011. 2

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[34] Ziyu Wang, Ke Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, Guxian Bin, and Gus Xia. Pop909: A pop-song dataset for music arrangement generation. In *Proc. of 21st Int. Conference on Music Information Retrieval Conf.*, 2020. 1

[35] Shiqi Wei and Gus Xia. Learning long-term music representations via hierarchical contextual constraints. In *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*, 2021. 1, 2

[36] Li-Chia Yang and Alexander Lerch. On the evaluation of generative models in music. *Neural Computing and Applications*, 32(9):4773–4784, 2020. 2

[37] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2