

拍節関係を取り入れた音楽生成モデルの提案

名越 崇晃^{1,a)} 北原 鉄朗^{1,b)}

概要:

本研究では、音楽生成モデルにおける生成品質を向上させるため、新たな損失関数を提案する。この損失関数は一小節 64 ティックという拍数制約に基づき設計され、Transformer アーキテクチャに適用される。しかし、この手法はリズム的な要素を主眼としており、リズム的な整合性を重視した設計となっている。提案モデル「Metric-Oriented Rhythmic Transformer for Melodic Generation (MORTM)」は META MIDI DATASET を用いた実験で、生成結果のリズム構造と旋律進行の自然さの向上を目指した。

1. はじめに

音楽生成は、コンピュータによる自動作曲技術として古くから研究されてきた分野である。ディープラーニング以前の音楽生成研究では、ルールベースのアルゴリズムや確率モデルが主流であった。たとえば、Markov 連鎖を用いた楽曲生成 [1] や、確率グラフモデルによるメロディ生成 [2] などが挙げられる。これらの手法は理論に基づく生成が可能である一方、複雑な依存関係の処理が難しいという課題があった。

近年、深層学習技術の進展に伴い、音楽生成の手法は大きく変化した。特に、Transformer モデルはその優れた長期依存関係の処理能力から、音楽生成においても有望なアーキテクチャとして注目されている [3], [4]。さらに、相対位置エンコーディングの導入により、従来の位置エンコーディングよりも柔軟な表現が可能となり [5]、長い楽曲構造の生成をより効果的に行えるようになった。

しかし、既存の音楽生成モデルは、主にクロスエントロピー損失関数を使用し、生成されたデータと訓練データの間の数値的な誤差を最小化する手法に依存している [6], [7]。このような損失関数では、リズム制約的な正当性や構造が十分に考慮されていないため、生成された音楽が和声やリズムの一貫性を欠き、音楽的に不自然な結果となる場合がある [8], [9]。

また、前回の研究では、ジャムセッションにおける即興演奏の一環として、メロディを自動生成する AI システムの基礎的検討を行った [10]。その主な焦点は、Transformer モデルを用いたメロディ生成の初期的な可能性の評価にあっ

た。一方、本研究では、リズム制約に基づく損失関数の導入を通じて、生成音楽の理論的整合性を高めることを目指している。

本研究では、リズム制約に基づいた損失関数を提案し、この損失関数を用いることで音楽生成モデルの理論的整合性を向上させることを目指す。提案手法は、生成された音楽が一小節あたりの拍数や繰り返し構造といったリズム制約の基本的要素を遵守するよう設計されている。具体的には、相対位置エンコーディングを搭載した Transformer エンコーダーアーキテクチャを採用し、META MIDI DATASET を用いた評価実験を通じて、生成モデルが理論的に整合性のある音楽を生成する能力を検証する。

2. 目的

本研究の目的は、音楽生成モデルにおける生成品質を向上させるため、リズム制約を損失関数に組み込む新たな手法を提案することである。具体的にはリズム制約に基づき、生成された音楽の一小節あたりの拍数が適切であるかを評価可能な損失関数を設計する。

これにより、前回の研究で作成した生成モデルでは考慮されていなかったリズム制約的な制約を明示的にモデルに組み込み、生成音楽のリズム、旋律進行において整合性を向上させることを目指す。

従来の音楽生成モデルは、クロスエントロピー損失のみに基づいて学習を行うため、拍数やリズム構造が不自然になることがあった。本研究では、リズム制約を損失関数に組み込むことで、拍数の正確性や旋律の一貫性を担保し、音楽的に自然な結果を実現する。

¹ 日本大学 文理学部 情報科学科

^{a)} nagoshi@kthrlab.jp

^{b)} kitahara@kthrlab.jp

3. 提案手法

本研究では、リズム制約を考慮した損失関数と Transformer ベースの音楽生成モデルを提案する。提案手法は、従来のクロスエントロピー損失に加え、リズム制約的な制約を考慮した損失項を導入することで、生成される音楽がより理論的整合性を持つことを目指す。本節では、提案する損失関数およびモデルのアーキテクチャについて詳述する。

3.1 MORTM の目的

本節では、提案モデルである **Metric-Oriented Rhythmic Transformer for Melodic Generation (MORTM)** の目的、およびその利点について説明する。MORTM の目的は、与えられた簡単な旋律を基に、自然でリズム制約に整合した続きの旋律を生成することである。具体的には、過去の旋律データを学習し、その中から規則性やパターンを導き出すことで、楽曲を生成することを目指している。本研究で提案するモデル「MORTM」は、従来の Transformer を基盤とし、以下の3点で特徴付けられる。

3.1.1 相対位置エンコーディングを導入したアテンション機構

Transformer モデルは畳み込みや再帰構造を持たないため、入力トークンの順序情報を明示的に付加する必要があった。そこで位置エンコーディングには従来の絶対位置エンコーディングに加え、相対位置エンコーディングを採用し、よりも柔軟な時間的関係を学習することを可能にした [5]。相対位置エンコーディングはシーケンスの長さに依存しないため、モデルの一般化能力が向上する。また、トークン間の相対的な距離情報を強調するため、局所的な繰り返しを引き起こしやすい可能性があり、これにより音楽のリズムや音符間の関係や、繰り返し構造を自然に捉えることができる。

3.1.2 エンコーダーのみを使用

従来の Transformer アーキテクチャからデコーダーを排除し、エンコーダーのみを使用することで、計算コストを削減しながら効率的に音楽トークンの依存関係を学習する。

3.1.3 オリジナルトークナイザーの設計

音楽データのトークン化には以下のオリジナルトークンを導入し、リズム制約的要素を明示的に表現する：

- **P トークン**：音高 (Pitch)
- **D トークン**：音の長さ (Duration)
- **S トークン**：音同士の距離 (Spacing)
- **M トークン**：小節の区切り (Measure)
- **< S_SEQ > トークン**：シーケンスの開始
- **< E_SEQ > トークン**：シーケンスの終了

これにより、楽曲の時間的構造や小節の単位が明示的にモデルへ入力される。

3.2 リズム制約を考慮した損失関数

3.2.1 一小節 64 ティックに基づく損失

一小節は 64 ティックに分割され、生成される音楽の **D トークン** (音の長さ) および **S トークン** (音同士の距離) の合計が 64 ティックに収まる必要がある。この制約を定量的に評価するため、以下の損失関数を提案する。

$$L_{theory} = e^{k \cdot |x-n|} - 1$$

ここで、

- x は生成されたトークン列のティック数合計
- n は正解値である 64 ティック
- k は損失のスケールを調整するハイパーパラメータ
例えば、生成結果が 44 拍の場合、 $|x-n| = 20$ ($64-44$) となり、この値を損失関数に代入することで誤差が大きいくほど指数関数的に損失が増加するよう設計されている。

3.2.2 小節単位での平均損失

提案手法では、生成される音楽を 8 小節ごとにシーケンスを区切って学習を行う。各小節ごとに前述の損失を計算し、その平均値を最終的な損失として使用する。

$$L_{total} = cw \cdot L_{CE} + rw \cdot \frac{1}{N} \sum_{i=1}^N L_{theory,i} \log L_{theory,i}$$

ここで、

- L_{CE} は従来のクロスエントロピー損失
- $L_{theory,i}$ は第 i 小節におけるリズム制約的損失
- N は 1 シーケンス内の小節数 (8 小節)
- cw と rw はエポックごとに変化する重み

このアプローチにより、従来の学習性能を維持しつつ、リズム制約的に整合性のある楽曲生成を実現する。

3.3 モデルの学習と出力形式

提案モデルは、次のトークンをロジット形式で出力することで、トークン列の予測を行う。学習時には 8 小節単位で区切られたトークン列を対象に、損失関数 L_{total} を用いて最適化する。

3.4 損失関数のスケジューリング機能

この損失関数には、スケジューリング機能を導入している。学習の初期段階ではクロスエントロピー損失に重み付けを行い、基本的な楽曲構造を学習させる。エポックが進むにつれ重みである cw と rw がリズム制約的損失のほうに傾くようになっている。これにより、最初はクロスエントロピーにより正解ラベルとの誤差を無くすようにし、次第にリズム制約に焦点を当てるようにしている。

このように、提案手法は従来の Transformer モデルの強みを活かしつつ、リズム制約を考慮した損失関数を組み込むことで、生成される音楽の品質を上げている。

4. 評価実験

本節では、提案手法の有効性を検証するために行った実験について述べる。実験デザイン、使用データセット、評価指標、および実験結果の分析方法について詳述する。

4.1 実験デザイン

提案手法の性能を評価するため、特定の音階（スケール）に基づき生成モデルを初期化し、その続きの旋律を生成させる。サンプル 1 には繰り返しが多く、同じ音高を多く使用した旋律を 8 小節与えている。サンプル 2 にはサックスソロのような旋律を与えており、繰り返し構造が比較的少なく、様々な種類の音高を使用している。

4.2 使用データセット

実験には **META MIDI DATASET** を使用する。これは広範な音楽ジャンルをカバーする MIDI 形式のデータセットであり、リズム制約的な評価に適している。

4.3 評価指標

提案手法の性能評価には、以下の指標を使用する。

- **拍数の正確性**：一小節あたりのトークンの合計が 64 ティックに収まっているかを評価する。
- **音の出現度分布**：生成された音の分布をヒストグラムで可視化し、バランスや偏りを確認する。
- **旋律の構造**：譜面を用いて生成された旋律の視覚的構造を評価する。

以上の評価実験により、提案手法がメロディの質の向上に貢献しているかを示す。

5. 結果

本節では、提案手法によって生成された旋律に関する実験結果を示す。

5.1 音高出現頻度のヒストグラムによる評価

音高の分布をヒストグラムで可視化し、旋律における音の使用傾向を分析した。

5.1.1 出現頻度の偏り

- 図 4 では、A $\sharp$ が最も頻繁に使用され、次いで D $\sharp$ 、F、C が主要な音として登場していた。一方で、G や G $\sharp$ など一部の音はほとんど使用されておらず、音高の分布に偏りが見られるが、多様性の一定の確保が確認された。
- 図 3 では、A $\sharp$ と C が最も高頻度で使用されており、旋律の基盤を形成していた。一方で、D や G $\sharp$ などの音は頻度が非常に低く、音高の分布に偏りが顕著であった。

5.1.2 分布のバランス

- 図 4 では、音高の分布において一定の多様性が観察され、旋律進行に変化をもたらしている。

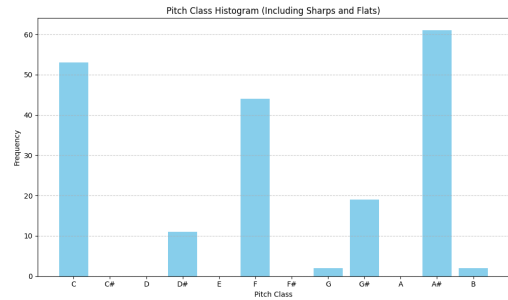


図 1 生成された旋律の音高分布（サンプル 1）

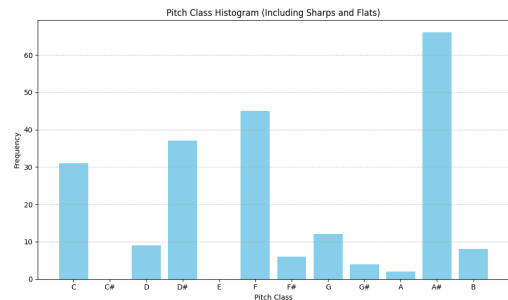


図 2 生成された旋律の音高分布（サンプル 2）

- 図 3 では、A $\sharp$ と C への依存が顕著であり、音楽的多様性が制限される可能性が示された。

5.1.3 音高の多様性

- 両サンプルとも、主要音（A $\sharp$ 、C、D $\sharp$ など）が繰り返し使用される傾向が見られた。ただし、特定の音に過度に集中することで旋律が単調になりやすい点が示唆された。

5.2 譜面による評価

生成された旋律を譜面形式で可視化し、以下の観点で評価を行った。8 小節目までが与えられたサンプルであり、残りの小節が MORTM が生成した旋律になる。

5.2.1 リズムの整合性

- 図 4 では、音符が連続的かつ滑らかに配置され、小節単位でリズムが整然としていることが確認された。具体的には、一小節 64 ティックの制約が全体を通じてほぼ満たされており、リズム制約がモデルに適切に反映されている。
- 図 3 では、音符の配置が一定の間隔で繰り返され、リズムの規則性が際立っていた。ただし、音符の密度が高いため、全体のリズム表現に柔軟性が不足している可能性が示唆された。

5.2.2 旋律進行の自然さ

- 図 4 では、MIDI ノートが広範囲にわたり使用され、音高が徐々に下がる傾向が確認された。この結果、楽曲全体における終止感が強調され、自然な進行が表現されていることが分かった。
- 図 3 では、音高が主に中音域（60–80 の範囲）に集中

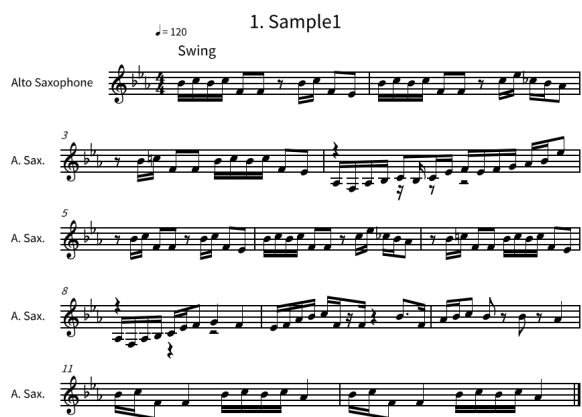


図 3 生成された旋律の譜面 (サンプル 1)



図 4 生成された旋律の譜面 (サンプル 2)

し、旋律の動きが限定的で単調な印象を与える結果となった。

5.2.3 繰り返し構造

- 図 4 では、音高の動きに一定の規則性がありつつも多様性が保たれ、音楽的な一貫性を維持している。
- 図 3 では、同じ音高が繰り返される傾向が強く、繰り返し構造が強調される一方で、旋律の多様性が制限される結果が確認された。

6. 考察

提案した音楽生成モデル「Metric-Oriented Rhythmic Transformer for Melodic Generation (MORTM)」の性能を、生成された旋律の譜面および音高分布ヒストグラムの

分析を通じて検証した。

6.1 譜面の分析

譜面からは、提案モデルがリズム制約を守りながら自然な旋律を生成できていることが確認できた。

6.1.1 リズム構造の整合性

1 小節 64 ティックのリズム制約が全体にわたって守られており、生成された音符の配置は時間的に整然としていた。この結果は、リズム制約を損失関数に明示的に組み込むアプローチの有効性を示している。

6.1.2 旋律進行の自然さ

図 4 では音高の動きに多様性があり、旋律進行が滑らかかつ音楽的であった。一方で、図 3 は音高の変化が単調であり、規則性が強調された結果となった。この差異は、生成モデルが異なる入力条件や損失関数の重み付けに応じて旋律の表現を調整できる。

6.1.3 音高分布ヒストグラムの分析

ヒストグラムの結果は、生成モデルがスケールや調性に基づく規則性を学習し、それを旋律生成に反映していることを示している。

6.1.4 主要音高の頻出

図 1 および図 2 のいずれにおいても、特定の音高（例: A#, C）が高頻度で使用されており、これが旋律の調性と安定性を構成している。また、これらの音高の頻出は楽曲全体の一貫性を高める要因として機能している。

6.1.5 音高分布の偏り

図 1 では一部の音高（例: A#）に使用頻度が集中しており、旋律が単調になる可能性が示唆される。一方で、図 2 は音高の分布が広がりを持ち、音楽的多様性が向上していた。これは、損失関数の設計が生成結果の多様性に直接影響を及ぼすことを示している。

6.1.6 補助音と経過音の役割

ヒストグラムでは頻度の低い音高（例: D, G#）も確認されており、これらは旋律に変化や動きを付加する補助的な役割を果たしていると考えられる。

6.2 提案モデルの評価

以上の結果から、提案モデルの利点と課題を以下のように整理する。

6.3 課題

図 1 で観察された音高分布の偏りは、旋律の単調さにつながる可能性がある。この問題には、損失関数の重み付けやデータ拡張の導入による改善が必要である。また相対位置のマルチヘッドアテンションにより、過度に同じ旋律が繰り返されてしまうことも課題としてあげられる。

6.4 総括

本研究の提案手法は、損失関数にリズム制約的要素を組み込むことで、以下のような成果を示した：

- **スケールの考慮**：与えられた旋律のスケールを保持し、自然な音高分布を実現.
- **リズムの整合性**：1小節 64 ティックの制約をほぼ満たし、時間的な構造の一貫性を担保.
- **旋律の自然な継続**：入力された旋律の特徴を保持しつつ、新たな旋律を自然に生成.

これにより、提案手法は音楽生成モデルとして理論的整合性と高い生成品質を両立し、実用的な応用可能性を示す結果となった.

7. まとめ

本研究では、音楽生成モデルにおけるリズム制約を考慮した新たな損失関数を提案し、これを導入した Transformer ベースの生成モデル **Metric-Oriented Rhythmic Transformer for Melodic Generation (MORTM)** を開発した. 提案手法は、従来のクロスエントロピー損失に加えて、一小節あたりの拍数や繰り返し構造といったリズム制約的な要素を直接損失関数に組み込むことで、生成された音楽の理論的整合性を向上させることを目指した. 提案モデルの性能を評価するために、META MIDI DATASET を用いた評価実験を実施した.

- **リズムの整合性**：1小節あたり 64 ティックのリズム制約をほぼ満たし、リズム制約に基づく適切な時間的構造を生成.
- **スケールの保持**：与えられた旋律のスケールを適切に維持し、音楽的に自然な旋律を生成.
- **旋律の自然な進行**：入力された旋律の特徴を反映しつつ、滑らかで一貫性のある旋律進行を実現.

一方で、本研究には以下のような課題も存在する.

- 提案手法の性能を多様なジャンルや楽器で検証する必要性.
- リズム制約以外の要素を考慮した損失関数の拡張.
- 過度な繰り返し構造の抑制

今後の研究としては、提案手法を拡張し、より幅広い音楽ジャンルや複数楽器を対象とした生成モデルの開発を行う. また、リズム制約以外の要素を考慮した損失関数の設計や、生成プロセスの効率化に向けた技術的改良にも取り組む予定である. これらの研究を通じて、提案手法の実用性をさらに高めることを目指す.

謝辞

本研究は、JSPS 科研費 23K24966, 24H00748 の助成を受けた.

参考文献

- [1] Darrell Conklin. Music generation from statistical models. *Proceedings of the AISB Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, pages 30–35, 2003.
- [2] David Cope. *Experiments in musical intelligence*. A-R Editions, 1996.
- [3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.
- [4] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, and Douglas Eck. Music transformer: Generating music with long-term structure. In *Proceedings of the International Conference on Learning Representations*, 2018.
- [5] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.
- [6] Li-Chia Yang, Shyh-Kang Chou, and Yi-Hsuan Yang. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. In *Proceedings of the International Society for Music Information Retrieval Conference*, 2019.
- [7] Jean-Pierre Briot, Gaetan Hadjeres, and François Pachet. *Deep learning techniques for music generation*. Springer Nature, 2020.
- [8] Gaëtan Hadjeres, François Pachet, and Frank Nielsen. Deepbach: a steerable model for bach chorales generation. In *Proceedings of the International Conference on Machine Learning*, pages 1362–1371, 2017.
- [9] Natasha Jaques, Shixiang Gu, Richard E Turner, and Douglas Eck. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control. In *Proceedings of the International Conference on Machine Learning*, pages 1645–1654, 2017.
- [10] 名越崇晃 and 北原鉄朗. ジャムセッションにおける合いの手の挿入に向けた旋律生成の検討. In *エンタテインメントコンピューティングシンポジウム (EC2024)*, pages 438–441. 情報処理学会, 2024.