Transformer Decoder を用いたポピュラー音楽における メロディからの伴奏生成

1235057　沖 貴司　【 ソフトウェア検証・解析学研究室 】

Popular Music Accompaniment Generation from Melody Using Transformer Decoder

1235057　OKI, Takashi　【 Software Verification and Analysis Lab. 】

1　はじめに

近年，音楽配信サービスや動画サービスの発展により，音楽が身近なものとなっている．また，作曲用ソフトウェア等も発展しており，作曲に関する高度な専門知識の無い人が，音楽鑑賞だけでなく作曲にも取り組む機会が増えている．

作曲では，多くの場合にメロディと伴奏を制作する必要がある．メロディの制作は，普段から音楽を聴く中で身につけた感性によって行われる場合も多く，音楽的な知識があまり無い人でも充分に制作は可能である．一方，伴奏の制作は，ハーモニーやリズムパターン等を考慮する必要があり，高度な専門知識や経験が無ければ困難である．よって，コンピュータによる伴奏生成に関する多くの試みがなされている．

音楽自動生成のアプローチとして，ルールに基づく手法や機械学習に基づく手法等があるが，本研究では機械学習に基づく伴奏生成について扱う．

近年の機械学習を用いた音楽自動生成の研究では，ニューラルネットワークによるシーケンスモデルの一つである Transformer[1] を用いることが有効なアプローチとして注目されている．Transformer は，図 1 に示すように，入力シーケンスを解釈するエンコーダーと，出力シーケンスを構成するデコーダーから構成される．Transformer を用いたメロディからの伴奏生成 [4] では，メロディをエンコーダー，過去の伴奏をデコーダーに入力して生成が行われている．

一方，近年では Transformer のエンコーダーもしくはデコーダーのみを用いた多くのモデルが提案され，様々な自然言語処理タスクで最先端の結果を残した．入力を伴う生成でも，図 2 のように Transformer のデコーダー（以下，Transformer Decoder と呼ぶ）を用いて優れた性能を出している [3]．この結果から，シーケンスを扱うタスクにおいて，Transformer より，その一部の構造を用いる方が優れた結果を出せることが考えられる．よって，メロディからの伴奏生成においても，Transformer Decoder を用いることで優れた結果を出す可能性が考えられるが，現在その研究はなされていない．

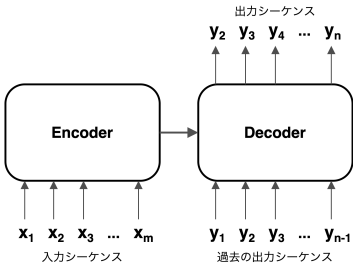

図 1　Transformer を用いた生成の概略図

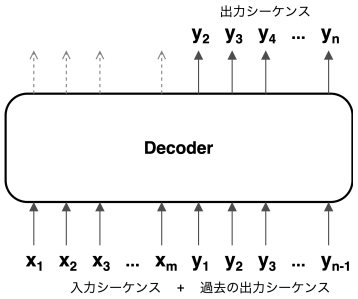

図 2　Transformer Decoder を用いた生成の概略図

そこで，本研究では Transformer Decoder のみを用いた生成手法を実装し，生成された楽曲の品質が Transformer を用いた場合よりも優れている可能性を発見した．この手法の有効性を示すために主観的評価による比較実験を行った．

2　提案手法

Transformer Decoder は，過去のシーケンスから次のトークンを出力する処理を繰り返す自己回帰モデルである．本手法では，メロディ及び伴奏シーケンスをこの順序で連結し，一続きのシーケンスにする．伴奏生成は，メロディシーケンスを過去のシーケンスとしてモデルに入力し，それに続くトークンを繰り返し出力することで行う．

また，機械学習による音楽自動生成では，音楽情報のシーケンスへの変換方法が結果に大きく影響することが知られている．音楽情報のシーケンスへの変換方法として，REMI を独自に拡張したものを用いる．REMI[5] は，ポピュラー音楽におけるピアノ曲を生成するために

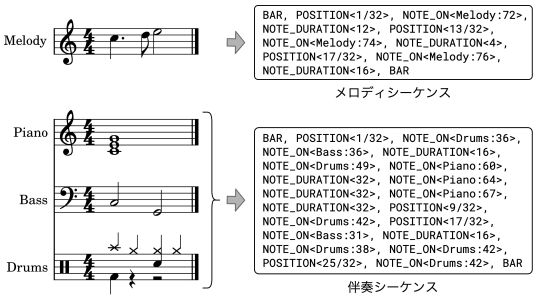

図 3　楽譜からシーケンスへの変換例

提案された変換方法である．REMI で主に使用されるトークンは以下の通りである．

- BAR — 小節の始まりを表す．
- NOTE_ON<p> — p 番目のピッチの音の立ち上がりを表す．
- POSITION<t/T> — NOTE_ON<p>トークンの直前に現れる．NOTE_ON<p>トークンによって音を立ち上げるタイミングが，小節内での位置 t ($1 \leq t \leq T$) であることを表す．ここで，T は 1 小節の時間分解能を示す．
- NOTE_DURATION<t> — NOTE_ON<p>の直後に現れる．NOTE_ON<p>トークンによって立ち上がる音を長さ t だけ持続することを表す．

本手法では，NOTE_ON トークンに，NOTE_ON<Melody:p>，NOTE_ON<Piano:p>のように各楽器名も含めることで，複数の楽器を区別して扱えるように拡張する．図 3 に本手法における楽譜からシーケンスへの変換例を示す．

モデルの学習及び比較実験には，The Lakh MIDI Dataset v0.1 の Clean MIDI subset[2] をデータセットとして使用する．これは，多くのポピュラー音楽を含む，約 17000 個の MIDI ファイルの集合である．これらの MIDI ファイルから，すべての楽器が充分に演奏されている範囲を 8 小節ずつ切り取って使用する．なお，楽曲によって伴奏として演奏されている楽器は様々であるが，本研究では学習の容易さの観点から，使用する伴奏楽器をピアノ，ベース及びドラムに限定する．

3　比較実験

比較実験のために Transformer 及び Transformer Decoder モデルを実装した両モデルの構成は，自己注意層の数を 4，自己注意層の次元数を 256，自己注意ヘッドの数を 4，埋め込みの次元数を 256，全結合層の次元数を 1024，Transformer の扱うシーケンスの最大長を 512，Transformer Decoder の扱うシーケンスの最大長を 1024（メロディ及び伴奏シーケンスを連結して扱うため，Transformer の 2 倍のシーケンス長が必要である）とし，パラメータの数がある程度等しくなるように設定した．また，生成方法として，多くの先行研究で用いられている確率的サンプリング法を用いた．

これらの設定で生成した伴奏の品質を比較するために，5 名に対してリスニングテストを行った．リスニングテストのために，同じメロディに対して上記 2 モデルで生成した音源を 1 セットとし，30 セットを作成した．被験者は，各セットにおいて 2 種類の音源をランダムな順序で聴き，それぞれの音源に対して “ハーモニーが心地よい”，“リズムが統一されている”，“まとまりがある”，“総合的に好き” という観点で，5 段階のリッカート尺度による評価を行った．

表 1　5 段階尺度を 1〜5 点で点数付けした場合の平均点と Wilcoxon の符号付き順位検定の p 値．左の列が Transformer Decoder の平均点，中央の列が Transformer の平均点，右の列が p 値を示す．

	TD scores（提案手法）	T scores	p-value
ハーモニーが心地よい	**3.44**	2.91	7.67e−7
リズムが統一されている	**3.64**	3.19	3.85e−5
まとまりがある	**3.39**	2.85	3.99e−5
総合的に好き	**3.32**	2.83	2.79e−7

表 2　セットごとの勝ち数（合計セット数 = 150）．左の列が Transformer Decoder の方が評価が良かったセット数，右の列が Transformer の方が評価が良かったセット数を示す．

	TD wins（提案手法）	T wins
ハーモニーが心地よい	**74**	26
リズムが統一されている	**68**	35
まとまりがある	**69**	26
総合的に好き	**69**	27

表 1 は 5 段階尺度を 1〜5 点で点数付けした場合の平均点と Wilcoxon の符号付き順位検定の p 値を示し，表 2 はセットごとの勝ち数を示す．これらの結果から，Transformer Decoder を用いた方が，より好まれていることが分かる．

4　まとめ

Transformer Decoder と独自に拡張した REMI を用いたメロディからの伴奏の自動生成を提案した．リスニングテストを通して，従来の Transformer を用いた生成手法との比較によってその有効性を示した．

今後の課題として，Transformer Decoder による生成が Transformer を用いた場合よりも優れた結果を出した要因を解析することや，一般的な長さの楽曲を生成するために，さらに長期的なシーケンスの生成を行えるようにすることが考えられる．

参考文献

[1] Ashish Vaswani et al. “Attention Is All You Need,” NIPS 2017, 2017.
[2] Colin Raffel. “Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching,” PhD Thesis, Columbia University, 2016.
[3] Tom B. Brown et al. “Language Models are Few-Shot Learners,” CoRR, 2020.
[4] Yi Ren et al. “PopMAG: Pop Music Accompaniment Generation,” CoRR, 2020.
[5] Yu-Siang Huang, Yi-Hsuan Yang. “Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions,” CoRR, 2020.