E-040

出力結果に強弱を伴ったオーケストレーションを行う自動編曲システム

An automatic arrangements system for orchestra that output with dynamics.

奥村静† 西浦良太† 土屋 誠司‡ 渡部 広一‡
Shizuka Okumura Ryota Nishiura Seiji Tsuchiya Hirokazu Watabe

1. はじめに

楽曲は音楽業界に留まらず，映画やドラマの効果音・サウンドトラックや飲食店の BGM 等，幅広い業界において需要がある．しかし楽曲制作を誰でも気軽に始められる技術が進出してきた現代社会であっても，やはりオーケストラ等の本格的な楽曲制作のためには専門知識と幅広い経験が必要である．既にこの分野を担う様々な生成 AI が存在しているが，その中には強弱の要素を楽曲生成においてノイズであると捉え，出力結果に強弱を伴わないものがある．筆者は強弱を楽曲構成における重大な要素の一つだと捉えており，精度向上のために排除するべきではないと考える．そこで本稿ではディープラーニングを用いて強弱を含んだオーケストレーションを行う自動編曲システムを提案する．ディープラーニングモデルには自然言語処理分野で一定の評価を得ている Transformer を採用した．編曲を行えるようメロディと伴奏の関係性を学習し，楽曲を壮大にする AI を実現する．上記のモデルでオーケストレーションを実現できるモデルの完成を目指す.

2. 関連技術

2.1 MIDI

MIDI(Musical Instrument Digital Interface)は，楽器や音楽機器とコンピュータや他の楽器との間で音楽情報をやり取りするための通信プロトコルである.

2.1 Music Transformer

Music Transformer[1]は音楽生成のためのニューラルネットワークであり，Transformer モデル[2]の音楽版と言える．Music Transformer は MIDI 形式の楽曲を学習データとして使用し，メロディと和音を同時に生成することができる.

2.2 Multitrack Music Transformer

強弱を含めたいくつかの MIDI 情報を排除することで短いシーケンス長を保ち，多様な楽器を扱える新しいマルチトラック音楽表現である．既存のシステムと比較して同等の性能を達成し，推論速度およびメモリ使用量の削減を実現している．これを実現するために新たな MIDI トークン化手法を提案している.

3. 先行研究

本研究は，鎌田凌輔氏による「ディープラーニングを用いたオーケストレーションを行う自動編曲システム」[3]を先行研究としている.

3.1. 先行研究の概要

作曲の AI である MMT を編曲に適応させており，最適な楽器選択ができるように学習を行なった上でオーケストレーションを行う自動編曲システムに改良している．定性的評価では BLEU スコアを用いており，一般的に高精度とされている 41 というスコアを示している.

3.2. 先行研究の問題点

楽曲生成において Tempo(BPM)と Velocity（強弱）の 2 要素はノイズになると判断し，学習の段階でこれらを省いている．よって全ての楽曲の全てのパートが一曲を通して同じ Tempo，音の大きさで出力される．Tempo の違いは音のタイミングや長さである程度表現できるのに対して，強弱は他の要素では表現できない独立した要素である．先行研究の手法では楽曲を生成する上で重大な意味を持つ要素が欠落している事が問題点として挙げられる.

4. データセット

4.1. Symbolic Orchestral Database (SOD)

Symbolic Orchestral Database[4,5]はオーケストラで構成された MIDI データが 5742 曲格納されたデータセットである． 90%を train data，10%を valid data としている.

4.2. Lakh-MIDI Dataset (LMD)

Lakh-MIDI[6]データセットは 176,581 曲の重複排他された 1980 年代-2010 年代の洋楽を中心とした有名楽曲の MIDI ファイルを含んでいる． 98%を train data，2%を valid data としている.

4.3 評価用データの作成

評価用データとしては，筆者自身が 10 曲を作曲した．学習データと全く異なるデータでテストすることが可能となる．加えてそれぞれ筆者なりの編曲結果も用意することで，出力結果と比較検討できるようにした.

5. 提案手法

5.1. Pre-training Transformer

本稿で提案した Transformer モデルに音楽表現を学習させる手法としては，LMD の楽曲データを用いて教師なし学習を行った．この手法は大規模言語モデルの事前学習に使用されるものである．大量の音楽データをトークンとして読み込むことで，次の単語を統計的に予測するタスクを繰り返し行い，トークン間の出現確率の組み合わせを学習することができるようになる.

5.2. Fine-tuning Transformer

ファインチューニングでは楽曲内の一旋律に対して他の旋律の対応を学習させる．これによってメロディに対する伴奏を生成することができ，オーケストラのような壮大な伴奏を生成できると考えている．学習データにはオーケストラ楽曲の MIDI が含まれた SOD を用いた．MIDI データから入力データとして 1 トラックずつ取り出し，元の楽曲を教師データとすることで伴奏の生成を実現できるようにする．この作業によって生成した入力データと教師データの組み合わせは 24,962 セットとなった.

5.3. モデルの全体像

全体の流れとしては，システムにメロディを入力するこ

† 同志社大学大学院理工学研究科
‡ 同志社大学理工学部インテリジェント情報工学科

とでファインチューニングを行った Transformer によってオーケストラ規模にまで編曲を行う．基本的な構造は Transformer と同様であるが，MIDI のトークン化手法のみ強弱の要素を付け加えた Multitrack Music Transformer のトークン化手法を用いている．ディープラーニングモデルには，自然言語処理の分野で一定の評価を得ている BERT[7]を採用した．音楽は構造的な情報（小節，拍子など）や多様な情報（テンポ，楽器，ピッチ）を含むため，自然言語処理の事前学習技術をシンボル音楽に単純に適応するだけでは，効果が少ないと示されている．本研究では Music BERT[8]で有効性を示された Bar-level mask strategy や Octuple MIDI encoding を採用することで，音楽理解に特化した事前学習を行う．パラメータによっては，原曲の情報を一部反映しているものの生成結果が和音の羅列にとどまり，楽曲としては成立しないものもあった．複数の設定で事前学習を繰り返した結果，楽曲としての完成度や原曲の再現性に違いが見られた．事前学習を繰り返した結果，最終的に batch size=1, accumulation steps=16，checkpoints=500,000 のモデルが，音楽的整合性と原曲の再現性の両面において最もバランスの取れた出力を示し，採用に至った．

5.4　強弱なしでの出力

本研究では比較検討用として，強弱を含めずに学習・生成を行った楽曲サンプルも用意した．

6. 評価手法

評価手法としては，5 年以上の音楽経験を持つ者を対象者として 10 曲のサンプルを用意し，3 つの項目について 10 段階で回答を得た．サンプルには強弱を含めて学習されたものが 6 曲と，強弱を含めずに学習されたものが 4 曲ランダムな順番で用意されており，この配分や順序は回答者に伝えていない．評価項目は「音楽的に豊かな強弱が反映されているか」「楽曲として成立しているか」「原曲の要素を適切に感じるか」の 3 つである．項目 1 については強弱が全く聞き取れなかった場合に 0 と回答してもらい，それ以外の場合は 1-10 の 10 段階でアンケートを行った．

7. 結果

16 人から得られた回答結果について回答者それぞれの平均値を算出したところ，点数のばらつきに大きな差があることが判明した．そこで異なるスケールを持つデータ間での比較検討を可能にするため，回答者ごとの評価傾向の違いを補正する Z スコア標準化を用いて回答者毎に平均が 0，標準偏差が 1 になるようデータを標準化した．標準化後のスコアについて，各設問における回答の平均値は以下の表 1 の通りである．

表 1　各設問における回答の平均値

	適切な強弱か	楽曲として成立しているか	編曲として適切か
強弱なし	-2.8	-2.8	-3.0
強弱あり	-3.0	-2.9	-3.0

また，3 つの設問に対する回答同士はどの組み合わせにおいてもそれぞれ正の相関関係にあり，設問 1 の評価が高ければ設問 2，3 も高い評価が得られた．そこで設問 3 つ分の回答を合計して比較することで，楽曲同士の精度が比較できた．

最高評価を得られた楽曲では 7 割程度の人が 10 段階中 8 以上の評価をつけたが，最低評価だった楽曲では 8 以上の評価をつけた人が 1 割であった．

8. 考察

表 1 より，数値が大きい方が高い精度を表しているが，比較すると強弱ありと強弱なしの評価結果で大差ない事が見て取れる．このことから，強弱の要素は楽曲生成において必ずしも精度を落とす原因であるとは言いきれない．しかし，楽曲生成の精度には楽曲ごとに大きなばらつきがある点に注意が必要である．また Z スコア標準化では平均値が 0 になるように標準化しているため，算出した値が負の数である点から全体的に生成の精度が芳しくないことが伺える．これらの原因としては，学習に長時間を要するため繰り返し施行することが難しく，学習の個体差が出てしまった事が挙げられる．実験最中に発生したサーバの故障により新しい環境になったことで，先行研究に比べてバッチサイズを落とす必要性がありパラメータを調整した．この際パラメータが最良化できなかったことも精度のばらつきの原因と考えられる．

9. おわりに

本研究ではディープラーニングを用いて強弱を含んだオーケストレーションを行う自動編曲システムを提案した．自然言語処理の要領で音の要素をトークン化し，Transformer 型を用いて学習及び生成を行なった結果，従来の作曲 AI と比較して表現力の加わった楽曲生成が可能となった．課題としては精度の不安定性が目立った．

参考文献

[1] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D.Hoffman, Monica Dinculescu, Douglas Eck. Music Transformer: Generating Music with Long-Term Structure. arXiv preprint arXiv:1809.04281.

[2]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.

[3]鎌田凌輔他, ディープラーニングを用いたオーケストレーションを行う自動編曲システム, 信学技報, vol.118, no.492, AI2018-53, pp.1-5, 2024 年

[4]Jean-PierreBriot, GaëtanHadjeres, and François-DavidPachet, “Deep learning techniques for music generation-a survey,” arXiv preprint arXiv:1709.06120,2017.

[5]Shulei Ji, Jing Luo, and Xinyu Yang, “A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions,” arXiv preprint arXiv:2011.06801.2020.

[6] Colin Raffel. Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching. Ph.D. thesis, Columbia University. 2016.

[7]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[8] Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, Tie-Yan Liu. Music BERT: Symbolic Music Understanding with Large-Scale Pre-Training. arXiv:2106.05630, 2021