Vol.2021-MUS-131 No.43
Vol.2021-SLP-137 No.43
2021/6/19

汎用敵対的生成ネットワークにより生成された旋律画像の補正

鈴木 大河[1,a)]　横山 真男[2]

概要：画像を利用した自動作曲の分野では，音楽の生成に特化した敵対的生成ネットワークを用いた手法が提案されている．この手法により生成された画像は，音高を示す点に濃淡が付き滲んだものとなる可能性がある．この問題は音高の決定が出来ないという問題を発生させる．本研究では，汎用敵対的生成ネットワークを用いて生成された旋律画像に対し，音高の濃淡に閾値を付け加重平均を取ることで，生成された画像の特徴を損なうことなく鮮明な音源にする手法を提案する．また，単純な GAN モデルを利用し生成された画像に対して補正を行い，主観評価を行う．

1. はじめに

近年，敵対的生成ネットワーク (Generative Adversarial Networks:GAN)[1] による画像の生成が多くの分野で行われている．GAN には多くのモデルが存在し，対応付けされた画像を利用した画像の生成や [2]，画像の連続した変化である動画の生成にも応用できる [3]．音の連続した変化が音楽であることから音楽分野も例外ではなく，旋律を音高と時間の 2 方向に展開し画像化することで GAN の仕組みを利用した音楽の生成が可能である [4][5]．先行研究は出力画像の鮮明化と調性の認識や横長画像の出力など音楽への最適化が行われている．しかし，二値化された旋律画像を学習し生成された画像は，音高を示す点に濃淡がつくことから鮮明ではなく滲むことがあり，この画像の滲みは画像を直接 MIDI 形式に変換する際の弊害となる．一方で，GAN を利用した場合に限らず，音楽の自動生成の研究では短区間の生成が行われることが多い．再帰型ニューラルネットワーク (Recurrent Neural Network:RNN) を利用した手法により，ある程度の長さの旋律を生成すると同様のパターンの繰り返しになることも多く，音楽的な正しさと興味深さを両立させることは困難である．

そこで，本研究は汎用敵対的生成ネットワークにより生成された旋律画像を旋律概形 [6] とみなし，旋律画像の特徴を損なうことなく補正する手法を検討した．本手法は GAN の仕組みそのものに手を加えることはなく，生成された画像の音高の濃淡に加重平均をとり加工を行うことで，汎用的な生成器を応用した音楽の生成を行うことが期待できる．

[1] 明星大学大学院 情報学研究科 情報学専攻
[2] 明星大学 情報学部 情報学科
[a)] 20mj005@stu.meisei-u.ac.jp

2. 関連研究

2.1 旋律概形

土屋ら [6] は，音楽の非専門家を対象とし旋律概形を利用した作曲手法を提案した．旋律概形は，各音の音高と音価が陽に表現されず，旋律を 1 本の曲線で示したものである．旋律概形は旋律の音高の時系列に対しフーリエ変換を行うことで抽出できる．また，抽出された旋律概形は，隠れマルコフモデルを用いて連続した音から各音階における正しい音名を推定し旋律に変換する．主観評価から，旋律概形には旋律の大まかな特徴を反映することが可能であると示された．本研究では，旋律概形の概念を部分的に利用した画像を GAN を用いて生成する．また，生成された画像に対し，旋律概形を模した作曲手法と同様の変換を行うことで，音階構成音に基づいた旋律が生成される．

2.2 音楽生成を行う GAN

Cunningham ら [4] は GAN を利用した音楽生成モデルである MidiNet を提案した．複数種類のモデルを含み，8 小節の旋律か旋律と和音の対を生成することができる．このモデルは，MIDI ファイルを小節単位に分割しているため，細かい音価を表現した生成が可能となる．専門家への主観評価で，RNN を利用した Google Magenta の MelodyRNN と比較して同様の旋律の繰り返しが少ない出力がされていることから，提案手法による出力がより興味深いと示されている．一方でこのモデルは前後の小節に依存した生成を行う影響から，8 小節を通した場合に音楽的な正しさがな

ⓒ 2021 Information Processing Society of Japan

Vol.2021-MUS-131 No.43
Vol.2021-SLP-137 No.43
2021/6/19

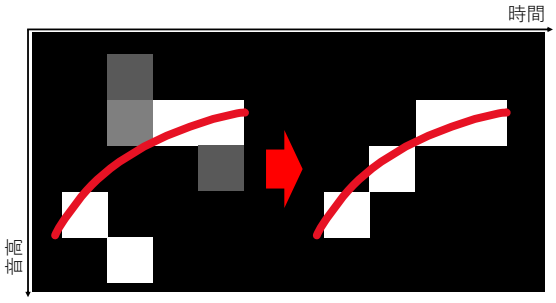


図 1 補正イメージ

く，予期せぬ音が発生するため人工的に生成された音楽である印象が強いことが示された．

Cheng ら [5] は，音楽の自動生成を行う Temporal GAN(TGAN) モデルを提案した．このモデルは，旋律と対応する和音のデータセットから，数小節単位の旋律と和音の対を生成するビートモデルと，生成された旋律と和音の対を時間軸方向へ順序付けする音楽感モデルの 2 種類により構成される．ビートモデルは，出力された画像を分析し音楽的な正しさを理解した出力がされた．一方で，音楽感モデルは音楽感モデルのディスクリミネータに対し，ビートモデルにより生成された楽曲と人間により作成された音楽的に正しい例と悪い例を入力した場合に，log 損失が大差がないことから時間軸方向の音楽的な正しさをディスクリミネータが理解したとは言えないことを示した．

本研究は，先行研究と同様の音価を持つ旋律の生成を行わず，旋律概形を模した画像を GAN により生成し補正を行う．最終的に補正した旋律画像を変換することで，音楽的に正しい旋律の生成を試みる．

3. 提案手法

本研究は，汎用敵対生成ネットワークにより生成された旋律画像の補正手法を提案する．提案手法のイメージを図 1 に示す．黒と灰色の箱が音を示し，曲線が想定される旋律概形のイメージを指す．左側が補正前，右側が補正後を表現する．補正対象の画像を生成するモデルは音高と時間方向への展開を行い，白色で表現した旋律部分と黒色の背景に二値化した画像を学習および生成する．画像は左上を原点とし，下方向に MIDI ノートナンバーによる音高情報を，横方向に時間情報を格納する．同一時間に複数の音高が確認された場合，音高の平均を出力値とする．この時，最低閾値以外の閾値を設け，その閾値を越えた場合には音高平均に同じ音高を加える．これにより，より白い音高に推定されやすくすることを期待する．

各時刻 t における補正された音 $CorrectedNote_t$ は式 (1) を用いて導出される．

$$CorrectedNote_t = \frac{1}{C_{sum}} \sum_{m=0}^{119} greyscale_m \cdot C \qquad (1)$$

この時，C は補正値を，C_{sum} は C の合計値を，$greyscale$ は時刻 t における音高の濃度を示し，$greyscale_m$ は時刻 t における MIDI ノートナンバー m の音高濃度を示す．C_{sum} は全ての要素の C が 1 である場合に要素数に等しくなる．補正値 C は，$greyscale$ に応じてより大きな値を取ることにより，単純な平均と比較して濃度の高い音高に寄せた音高が算出される．また，補正値である C は，$greyscale$ から式 (2) を用いて導出される．

$$C = \begin{cases} 0 & (0 \leq greyscale < G_{min} \text{のとき}) \\ 1 & (G_{min} \leq greyscale < G_1 \text{のとき}) \\ 2 & (G_1 \leq greyscale < G_2 \text{のとき}) \\ \cdots & \\ n-1 & (G_{n-2} \leq greyscale < G_{n-1} \text{のとき}) \\ n & (G_{n-1} \leq greyscale \text{ のとき}) \end{cases} \qquad (2)$$

この時，G は灰色の濃さの閾値を示し，G_{min} は認識する最低濃度の閾値を，n は閾値の個数を，$G_1 \sim G_n$ はそれぞれの閾値を示す．G_{min} と G を調整することにより，補正を行わない出力や単純な平均の算出などを行うことができる．音高は MIDI ノートと対応していることから 0 から 119 までの整数であるため，小数点以下は四捨五入する．

補正された画像は最終的に MIDI ノートに変換される．この時，生成された画像を旋律概形とみなし，先行研究 [6] を応用することで音楽的な正しさを考慮した変換を行う．

4. プロトタイプシステムの実装

提案手法の基本的な性能を評価するために，簡易的な単純 WGAN-GP[7] を利用した試験用画像生成システムを実装した．このシステムの学習画像は，ヤマハミュージックデータショップ[*1]の MIDI ファイルを加工したものを利用した．使用楽曲は，1989 年から 1999 年までのオリコンシングルチャートの上位 10 曲，合計 110 曲のサビ部分の旋律を利用した．旋律画像は MIDI ファイルを音高と時間の 2 方向に展開し画像化する．画像の背景色は黒，ノートがある場合の色は白を用いる．MIDI ファイルを画像にするにあたり，音価の情報は無視し全て同一とする．縦方向は MIDI ノートナンバーに合わせ 120 ピクセル，横方向は音の数のピクセル数の画像を生成する．この画像を縦方向に 8 ピクセル拡張し，時間軸を横方向に引き延ばすことで全画像を縦 128 ピクセル横 128 ピクセルに統一する．旋律画像の形式を図 2 に示す．

提案手法の補正値 C の影響を比較するため，プロトタイプシステムの補正値 C の暫定値として次の 3 種類を用意する．

C_a　音高濃度の加重平均から音高を算出
C_b　音高の平均から音高を算出
C_{max}　最も色の濃い音高を算出

[*1] https://yamahamusicdata.jp/(2021.5.18 最終アクセス)

ⓒ 2021 Information Processing Society of Japan

Vol.2021-MUS-131 No.43
Vol.2021-SLP-137 No.43
2021/6/19

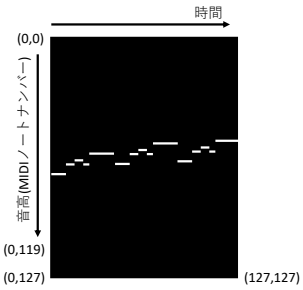


図 2 旋律画像の形式

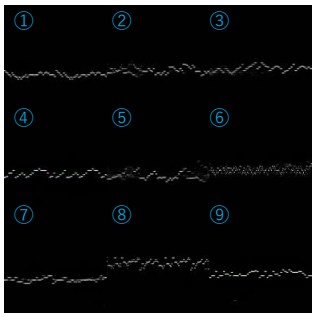


図 3 補正前の画像

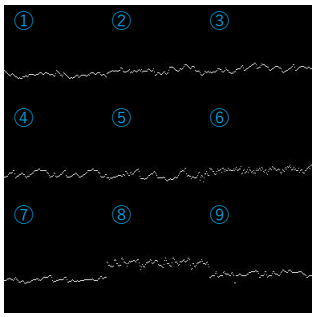


図 4 C_a による補正後の画像

C_a の暫定値を式 (3) に，C_b の暫定値を式 (4) に示す．

$$C_a = \begin{cases} 0 & (0 \leq greyscale < 25 \text{ のとき}) \\ 1 & (25 \leq greyscale < 50 \text{ のとき}) \\ 2 & (50 \leq greyscale < 100 \text{ のとき}) \\ 3 & (100 \leq greyscale < 150 \text{ のとき}) \\ 4 & (150 \leq greyscale < 200 \text{ のとき}) \\ 5 & (200 \leq greyscale \text{ のとき}) \end{cases} \tag{3}$$

$$C_b = \begin{cases} 0 & (0 \leq greyscale < 50 \text{ のとき}) \\ 1 & (50 \leq greyscale \text{ のとき}) \end{cases} \tag{4}$$

今回は主観評価で音源の評価を行わないため，音楽的な正しさを考えず画像の y 座標を直接 MIDI ノートと扱い MIDI ファイルとして出力する．

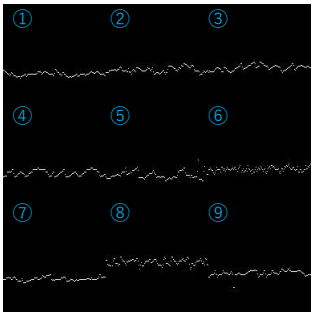
5. 主観評価

提案手法の基本的な性能を調査するために，プロトタイプシステムにより生成された補正前の画像と補正値 C_a により補正された画像の 2 枚を比較し，筆者が主観評価を行った．また，補正値の影響を調査するために，C_a により補正された画像と，補正値 C_b と C_{max} により補正された画像を比較した．補正前の画像を 9 枚並べたものを図 3 に，C_a による補正を行った画像を図 4 に，C_b による補正を行った画像を図 5 に，C_{max} による補正を行った画像を図 6 に示す．

5.1 考察

5.1.1 補正前の画像 (図 3) の出力

補正前の画像は大半が滲んでおり，何らかの補正が必要であると考えられる．補正前の画像の 2 枚目と 5 枚目に関



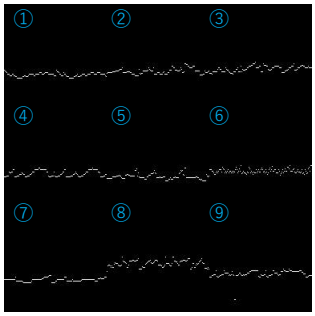


図 5 C_b による補正後の画像　**図 6** C_{max} による補正後の画像

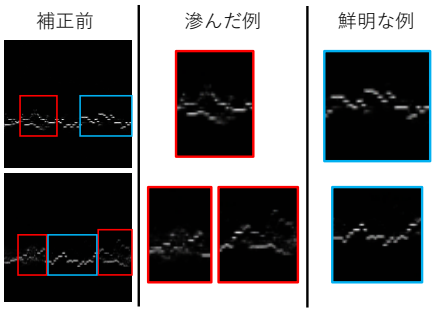


図 7 補正前の画像の滲む例・鮮明な例

して，滲んだ出力がされた例と比較的鮮明な出力がされた例を図 7 に示す．

音高は一定の範囲に固まっていることから，旋律に多く用いられる音域が出力されていると考えられる．出力された 9 枚の画像の音高で，最も高い音は MIDI ノートナンバーで 112，次に高い音が 92，最も低い音は 7，次に低い音が 58 となった．最低と最高を除いた画像内の平均最高値は 88.7，同様に平均最低値は 70.1 で，これは中央の B♭ から 1.5 オクターブの範囲内であり，多くの楽曲で旋律となる範囲であると考えられる．最高音と最低音に関係する部分で特筆される出力として 2 点挙げらる．補正前の画像の 8 枚目では最高値が 74，最低値が 58 となることから旋律としては低い出力がされている．補正前の画像の 9 枚目では，明確な外れ音である出力された画像中の最高音が確認され，この音は全ての補正例で強く影響を受けている．この 2 点は補正による修正が困難であると考えられる．

また，同音が続く区間が複数見られるが，学習画像を生成する段階で音価を無視しているため，これは同音が連続して現れる区間であり音価を表現するものではない．学習画像として利用した画像の多くの場合は，音価を無視したことで比較的鮮明が学習画像を生成することに成功した．一方で，長い旋律を利用した場合，学習画像を生成する段階で音価を無視する影響からか，旋律を時間軸方向に圧縮された影響で激しい音高の上下動と滲みが見られた．これに影響された出力例として 6 枚目が挙げられ，全体的に小刻みに音高が動いている．この出力は比較的長い旋律による影響を受けたと考えられる．

一方で，音高が 1 ピクセルの差で連続する区間も見られ

ⓒ 2021 Information Processing Society of Japan

Vol.2021-MUS-131 No.43
Vol.2021-SLP-137 No.43
2021/6/19

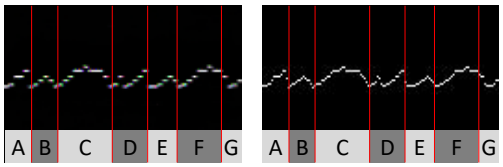


図 8 4 枚目の区間比較 (左から補正前，C_a による補正))

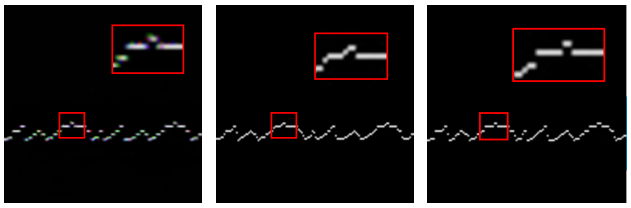


図 9 C 区間の比較 (左から補正前，C_a による補正，C_b による補正)

る．これは，補正を行った画像を MIDI ファイルへ変換する段階で音階などの音楽的な正しさを成立させるため，この段階での影響は少ないと考えられる．

5.1.2 補正された画像の検証

補正前と C_a による補正を行った画像 (図 4) を比較する．出力画像の 4 枚目を音高の変化で A から G の 7 区間に分割し，音高の上下動を比較した．全ての区間において音高の変化は概ね同一であると考えられる．B 区間，D 区間，E 区間では概形は同一であるが，より連続した線に補正された．線の端が滲んだことにより，平均値を算出する段階で前後に影響されたと考えられる．滲んだ部分は *greyscale* の値が非常に低いと考えられ，この影響が C 区間と F 区間のような同一音高から一部が上下動する区間で顕著に表れる．今回の補正では，補正対象とする濃度が比較的濃い C_b に対して C_a はより連続した線に補正された (図 9)．本研究では，出力された画像を線形概形とみなし最終的に再変換を行うため，連続した線に補正する C_a がより好ましいと考えられる．

補正前の画像の 5 枚目では前半と後半に大きく滲んだ部分が見られる．補正前の画像の最も明るい部分を音高として算出する C_{max} では大きな音高の変化が見られないのに対し，C_a による補正を行った画像ではある程度の変化がみられる (図 10)．この例では，前半部は C_a による補正が補正前の画像の見た目に即した補正だと考えられるが，後半は C_{max} による補正が見た目通りの補正を行っていると考えられる．

これらの問題は，補正値の G_{min} の調整と，補正値の閾値の段階の調整を行うことで違和感の解消を行えると考えられる．

6. 結論

本研究では，GAN を利用して生成された旋律画像に対し音高の加重平均を利用した補正手法の提案を行った．本手法を実装したプロトタイプシステムは，出力された画像の特徴を損なうことなく補正が行われることが確認された．補正値は暫定値を利用しているため，より多くのパターンによる出力を行いさらなる調整が必要である．また，音価を圧縮する影響から，長い旋律に寄せられた出力は時間方向の幅が足りず，1 ピクセル単位で大きく上下動する出力も見られた．音価を考慮した圧縮を行うか，学習対象となるデータを旋律概形そのものにするなど，出力対象とする画像そのものを改めて検討する必要がある．

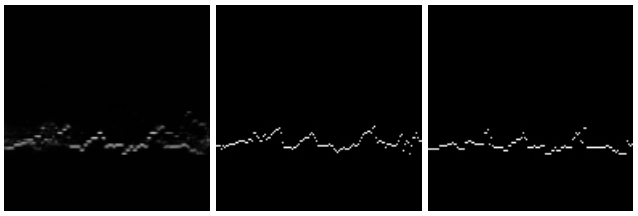


図 10 5 枚目の比較 (左から補正前，C_a による補正，C_{max} による補正)

今後は，pix2pix[2] を始めとした様々な GAN モデルによる出力画像への補正実験を行い，提案手法の汎用性を検証する．また，画像から MIDI ファイルの変換時に音楽理論に基づいた変換の実装を行い，出力された旋律の主観評価及び他の手法との比較実験を行う．

参考文献

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio: Generative Adversarial Nets, Advances in Neural Information Processing Systems, pp.2672-2680 (2014).

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros: Image-To-Image Translation With Conditional Adversarial Networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1125–1134 (2017).

[3] Aayush Bansal, Shugao Ma, Deva Ramanan, Yaser Sheikh: Recycle-GAN: Unsupervised Video Retargeting, Proceedings of the European conference on Computer Vision, pp.119–135 (2018).

[4] Sally Jo Cunningham, Zhiyao Duan, Xiao Hu, Douglas Turnbull: MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation, Proceedings of the 18th International Society for Music Information Retrieval Conference, pp.324–331 (2017).

[5] Ping-Sung Cheng, Chieh-Ying Lai, Chun-Chieh Chang, Shu-Fen Chiou, Yu-Chieh Yang: A Variant Model of TGAN for Music Generation, Proceedings of the 2020 Asia Service Sciences and Software Engineering Conference, pp.40–45 (2020).

[6] 土屋裕一, 北原鉄朗: 音符を単位としない旋律編集のための旋律概形抽出手法, 情報処理学会論文誌, Vol4, No54, pp.1302–1307 (2013).

[7] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron C. Courville: Improved Training of Wasserstein GANs, Proceedings of the 31st International Conference on Neural Information Processing Systems, pp.5769-5779 (2017).

ⓒ 2021 Information Processing Society of Japan