1T-06

音高遷移確率と画像濃度に着目した旋律画像の補正

鈴木大河 †　横山真男 ‡
明星大学情報学部

1　はじめに

近年，敵対的生成ネットワーク (Generative Adversarial Networks:GAN)[1] により，画像の生成が多くの分野で行われている．音楽は時間と音高の 2 次元展開することで画像化できる．このことから，畳み込みニューラルネットワークを用いた敵対的生成ネットワークにより音楽の生成が可能である．しかし，GAN は音階の前後関係を認識することが出来ず，出力された旋律は，出力区間の前後で音階が崩れてしまう問題がある [2]．

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，学習データから取得した音高遷移確率と画像濃度に着目し，生成された画像の特徴を損なうことなく補正する手法を提案する．

2　関連研究

旋律の自動生成手法には,「和声を軸にした旋律の自動生成」と「旋律の自動生成」の 2 種類が検討されてきた．和声を軸にする生成手法では，旋律の生成に利用する音を絞ることが可能となり，音楽理論に忠実で音階が崩れない旋律を生成することができる．しかし，旋律生成に入力として和声進行が必要であり，声構成音に制限された生成になる事から表現力に限りがある．本研究では，旋律を直接生成する手法を取る．

Yang ら [3] は GAN を利用した音楽生成モデルである MidiNet を提案した．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，MidiNet 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．

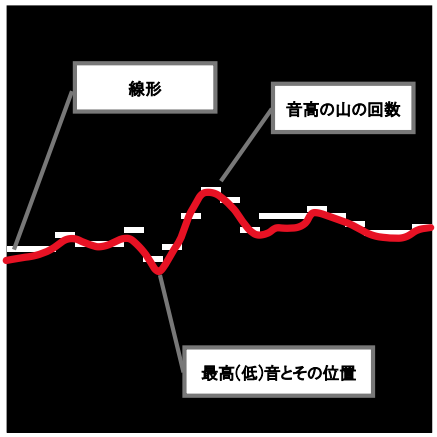

図 1: 画像化で表現可能な旋律特徴の例

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

本研究では，ソプラノ課題向けの旋律生成手法として，GAN により生成された旋律画像の補正手法を提案する．旋律を画像ベースで生成することにより表現可能と考えられる旋律特徴を図 1 に示す．旋律画像の生成には，画像生成モデルである単純な WGAN-gp を利用する．GAN により生成された旋律画像に対し，画像の濃度と学習データの音高遷移確率を利用した音高推定を行う．

3.1　学習対対象旋律の加工

本手法の旋律画像は，横軸に時間軸，縦軸に音高を取る．また，より長い区間の旋律の音高情報を含んだ旋律画像とするため，元の MIDI ファイル内の旋律が持つ音価は無視し圧縮する．画像化された旋律画像は，元データとして GAN の学習とディスクリミネータの評価に利用する．

Melody Image Correction focused on Pitch Transition Probability and Grayscale
†Taiga Suzuki (Meisei University)
‡Masao Yokoyama (Meisei University)

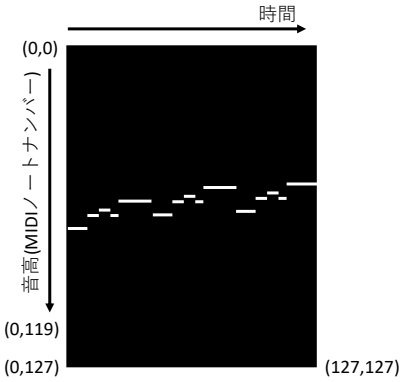

図 2: 旋律画像の形式

本システムの学習画像は，The Nottingham Database[4] から旋律を抽出し，MIDI 形式に変換したもの*を利用した．縦軸は音高である MIDI ノートナンバーを 0 から 127 まで格納し，横軸は時間軸とする．画像背景は RGB 値で (0,0,0) である黒を利用し，ノートが存在する区間を RGB 値 (255,255,255) で指定する．この画像を縦方向に 8 ピクセル黒色で拡張し，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．

3.2 音高推定値の算出

時刻 t における音高候補 $note_t$ は，画像のグレースケールの値から算出される G_{eval} と直前の時刻 $t-1$ で確定した音高空の遷移確率である P_{eval} を用いて推定される．

画像濃度からの補正値 G_{eval} は，式 (1) により算出される．$greyscale$ は時刻 t における音高の濃度を示す．

$$G_{eval} = \frac{1}{1 + exp(-(greyscale/100))} \qquad (1)$$

音高遷移確率からの補正値 P_{eval} は，決定済みの時刻 $t-1$ の音高 $note_{t-1}$ から，時刻 t の音高候補 $note_t$ への遷移確率を用いる．

各時刻 t における音高推定値 $Eval_t$ は式 (2) を用いて導出される．$Eval_t$ が最大となる音高が時刻 t における音高と推定される．

$$Eval_t = G_{eval} \cdot G_{weight} * P_{eval} \cdot P_{weight} \qquad (2)$$

補正値の算出イメージを図 3 に，GAN による出力画像と補正例を図 4 に，補正例の楽譜を図 5 に示す．

*このデータベースは ABC 記譜法により表現されているため，MIDI 形式に変換することなく画像化することが可能であるが，システムの汎用性を確保する観点から MIDI 形式への変換を行う．

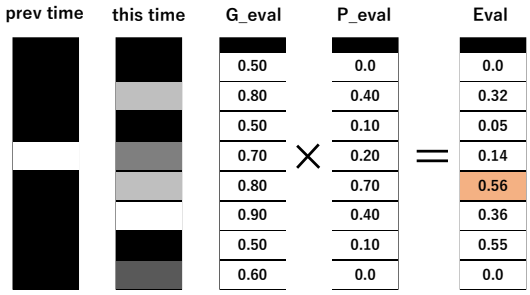

図 3: 補正値の算出イメージ

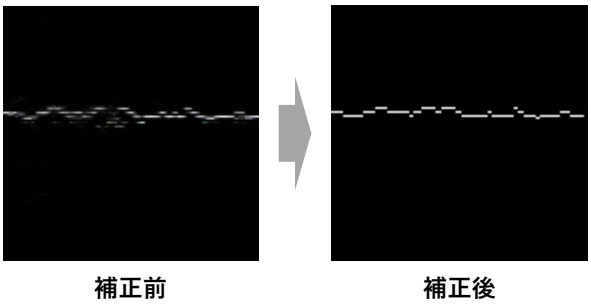

図 4: 出力・補正例

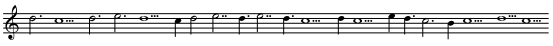

図 5: 補正例の楽譜

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] 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).

[3] Li-Chia Yang, Szu-Yu Chou, Yi-Hsuan Yang: 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).

[4] The ABC Music Project: The Nottingham Music Database，http://abc.sourceforge.net/NMD/ (最終アクセス:2022.01.06)．