拡散モデルを用いた構造保持型スタイル変換によるピアノ演奏表現の個別化

Individualization of Piano Performance Expression through Structure-Preserving Style Transformation Using Diffusion Models

谷 悠輔，曽我 真人
Yusuke TANI, Masato SOGA
和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University
Email: s276139@wakayama-u.ac.jp

あらまし：**ピアノ演奏において、「感情表現」の習得は重要で困難な課題である**. 本研究では，学習者の演奏をベースに目標感情を反映させた「理想的な演奏例」の生成を目指し,拡散モデルを使用した個別学習者用の MIDI ファイルの生成手法を提案する.具体的には,シンボリック音楽生成に適した Diffusion-LM を採用し,感情ラベルを学習させた分類器との組み合わせによる生成制御と部分的ノイズ注入法を組み合わせた,スタイル変換アルゴリズムを実装した.実験の結果,分類器誘導により感情適合率は向上したが,その勾配の過大適用による旋律の崩壊や構造の破綻といった課題が確認された.
キーワード：Diffusion-LM，ピアノ学習支援, Classifier Guidance，スタイル変換

1. はじめに

ピアノ演奏の習得プロセスは，楽譜情報の再現である「技術的習得」とその打鍵に付与された「芸術的表現」の 2 段階に分けられる．特に後者は，中級者から上級者への最大の障壁となっている．音楽大学などの専門的な教育機関を除き，多くの一般学習者は週に一度程度の対面レッスンと家庭での独学で練習する．指導者が不足した現状では，学習者が自身の演奏を客観的に評価し，表現を改善することは困難である．そこで私は，本研究では，学習者自身の演奏データをベースとし，その構造を維持したまま特定の感情エッセンスを付与する「個別化された演奏例」の生成システムを提案する．

2. 提案手法

本システムは，大規模データセットから音楽的知識を抽出する「事前学習」と個別演奏を再構成する「スタイル変換」の 2 段階に分けられる．

2.1 事前学習

スタイル変換の基盤となる生成モデルおよび制御用分類器の構築を行う．

1.**データセットの前処理とセグメンテーション**：MAESTRO データセットを用い，各楽曲を 30 秒の固定長ウィンドウで分割するセグメンテーションを実施した．現在は実装の簡略化のため重複のない切り出しを採用した．

2.**音楽表現の定義(REMI 形式)**：MIDI データは，小節構造や発音位置を明示的に扱う REMI 形式でトークン化される．本研究では，語彙を「楽曲の骨格を成す構造情報群（Bar, Position, Pitch）」と「奏者の表情を規定する変化情報群（Velocity, Duration）」に分類し，スタイル変換時の制御核となるバイナリマスクを定義する．

3.モデルアーキテクチャ：生成モデルは TransformerNet を採用し，16 次元の連続的な埋め込み空間上で拡散プロセスを学習する．感情分類器は 3 層の軽量な Transformer エンコーダを構成する．生成モデルと埋め込み層の重みを共有する転移学習を行うことで，潜在空間における効率的な感情誘導を可能にしている．

2.2 5 段階スタイル変換アルゴリズム

学習者の演奏を目標感情へと誘導するため，潜在空間における「物理的制約」と「感情誘導」を統合した以下の 5 段階アルゴリズムを実装した．

1.**初期化と選択的摂動の付与**: 入力演奏x_0に対し，表情情報群にのみ選択的にt_{start}ステップ分(全ステップの 40%)のノイズを注入する．これにより，旋律を破壊することなく,表現の書き換えに必要な「編集の余地」を潜在空間上に確保する．

2.**感情ガイダンス**：逆拡散過程の各ステップにおいて，分類器からの出力の勾配を用い，潜在変数を目標感情yの方向へ誘導する．

3.**潜在空間の安定化**：勾配適用に伴うベクトルの発散を抑制するため，各ステップの終了時に潜在変数のノルムを 4.0 に強制再投影する．これにより，変数を「音楽的に意味のある埋め込み領域」の内側に拘束する．

4.**構造情報の動的固定(Infilling)**：毎ステップ，マスクを用いて構造情報群(音高・位置)を原曲の埋め込みベクトルで上書き(Infilling)する．これにより，微細な勾配の影響による旋律の変質を物理的に排除する．

5.**カテゴリ制約付きデコーディング**：最終段階のデコードにおいて，本来 Velocity であるべき位置に Pitch が配置される等の「属性の不整合」を完全に排除するため，同一カテゴリ内でのみ最適なトークンを選択する制約付きデコードを実施する．

3. 実験結果

事前学習において，生成モデル（Diffusion-LM）および感情分類器が, MAESTRO データセット における音楽的文脈と感情属性の相関を正しく獲得したかを確認した。

3.1 生成モデルの学習推移

Diffusion-LM の訓練は, 20,0000 ステップにわたり実施された．混合精度学習(FP16)の活用により，全工程は約 16.7 時間で完了した．

1.**ロスの収束**:訓練ロスおよび検証ロスは，ステップ数の進行に伴い順調に低下し, 20,0000 ステップ付近で十分に収束した．

2.**音楽的文脈の獲得**：検証データにおけるロスの低下により，モデルが訓練データの単純な暗記に陥ることなく，REMI トークンの接続規則や小節構造を汎用的に学習できていることが示唆された．

3.2 感情分類器の識別性能

感情分類器は，生成モデルが学習した 16 次元の埋め込み空間を直接入力として学習を行った．

1.**学習効率**：生成モデルと埋め込み層の重みを共有しているため，約 1.1 時間の訓練で識別性能が安定した．

2.**収束特性**:訓練ロスの推移から，拡散プロセス中の「ノイズが含まれた潜在変数」に対しても，感情属性を識別可能なロバストな特徴抽出機が構築されたと言える．

3.3 スタイル変換実験の結果

逆拡散過程における目標感情$\boldsymbol{y}$に対する確信度を観察した結果，サンプリングの初期段階から確信度が上昇し，プロセスの大部分において約 90%以上の高い水準を維持し続けた．

1.**制御の有効性**:この結果は，生成モデルが構築した潜在空間において，分類器からの勾配ガイダンスが支配的に作用し，意図した感情領域(Q1~Q4)へと潜在変数を確実に引き込めていることを示している．

2.**変化の傾向**：例えば Q3(Sad)を目標とした場合，確信度の向上に伴いベロシティ値の統計的な下降が確認され，聴感上弱い打鍵へと変容する様子が確認された．

4. 考察

4.1 構造保持と感情表現の排他性の解消

スタイル変換における最大の課題は，制御対象(表現)と維持対象(構造)の干渉であった．逆拡散中の Infilling による「動的な固定」と，デコード時のカテゴリ制約による「静的な保護」という二重の防御メカニズムを構築した.カテゴリ制約を導入することで，属性を跨ぐトークンの「化け」を排除できた点は，離散的な音楽データにおいて有効なアプローチであるといえる．

4.2 感情誘導における「過剰制御」と音の欠落

感情分類器による強力なガイダンスは，潜在変数を目標感情の方向へ極端に押し込む．この際，変数が「音楽的に意味を持つ埋め込み空間」の境界を越えてしまうと，デコード時に最近傍のトークン(パディングや無音)へと収束し，「音の欠落」を引き起こすと考えられる．この現象は，芸術性の観点から，ガイダンス強度の動的な調整や，より広範なデータセットによる埋め込み空間の密度向上が必要である．

4.3 潜在空間の安定化における設計思想

本研究で導入した L2 正規化と「緩和フェーズ」は，非自己回帰型モデルの不安定性を補完する上で重要な役割を果たした。潜在変数を常に半径 4.0 の超球内に拘束することで，勾配適用に伴う数値的な発散を防止できた．これは，高次元の埋め込み空間において生成の軌道安定させるための必要要件である.最終 100 ステップで誘導を停止し,生成モデル本来の復元力に委ねる戦略は，分類器が捉える「瞬間的な表情」と生成モデルが持つ「音楽的文脈」を最適に融合させる要因になったと考えられる．

5. まとめ

本研究では，Diffusion-LM を用いた独自のスタイル変換手法により，学習者の個性を尊重した個別演奏例の生成を実現した．今後の展望として，楽曲全体の起承転結を考慮した長距離的な文脈制御や，感情ラベル付与における,楽曲の連続性を担保したセグメンテーション手法の検討，実際の学習者を対象とした評価実験を通じた教育的効果の検証が挙げられる．

参考文献

(1) 川邉 もゆ，小林 一郎: “拡散モデルを用いた感情と音楽属性に着目した音楽生成への取り組み”，情報処理学会研究報告 2025

(2) 和田 陽一郎，長名 優子:“Music Transformer を用いたポピュラー音楽における伴奏を元にしたメロディ生成”，情報処理学会研究報告 2025

(3) James A. Russell, "A circumplex model of affect," Journal of Personality and Social Psychology, Vol. 39, No. 6, pp. 1161--1178, 1980.

(4) Curtis Hawthorne, et al., "MAESTRO: A Dataset and Benchmark for Estimating Algorithmic Music Performance," International Conference on Learning Representations (ICLR), 2019.

(5) Hao Sun, et al., "Symbolic Music Generation with Diffusion-LM," arXiv preprint arXiv:2210.15557, 2022.

(6) Hao Sun, Liwen Ouyang.:“Diffusion-LM on Symbolic Music Generation with Controllability”, cs230.stanford.edu/projects_fall_2022/reports/16.pdf(2026 年 2 月 5 日確認)