音楽基盤モデルの表現形成における学習過程の解析手法の検討

佐藤 りん[1,a)]　田中 啓太郎[1,b)]　八木 颯斗[2,c)]　高道 慎之介[2,3,d)]　森島 繁生[1,e)]

概要：音楽基盤モデルは高品質な生成・理解性能を示す一方で，内部表現が学習によってパラメータ空間上のどの領域で，どのように形成されるのかについては，依然として十分に理解されていない．本研究では音高に着目し，モデルの中間表現に内在する音高螺旋構造のパラメータ空間上における顕在化の進行を探索的に調査する．具体的には，未学習モデルと学習済みモデルの間に連続的なパラメータ経路を構成し，経路上の各点における中間表現を解析する枠組みを採用する．基準として線形補間（全体サンプリング／学習済み近傍サンプリング）を用いるとともに，ラベルなし音響データに基づく勾配情報で局所的に補正した経路を導入し，経路の取り方が観測結果に与える影響を比較する．各点の中間表現に対して主成分分析により 3 次元へ低次元化を行い，パラメトリックな音高螺旋モデルをフィッティングすることで，螺旋構造の明瞭さを Helicality スコアにより定量化する．さらに，perplexity を併用して経路の性質を補助的に評価する．MusicGen および Jukebox を対象として，内在音高螺旋がどの程度学習済みモデルに近いパラメータ領域で顕在化するか，ならびに勾配補正がその傾向に与える影響を検討する．

1. はじめに

近年，自然言語処理や画像処理で成功を収めた大規模事前学習モデルの枠組みは，音楽情報処理分野にも急速に波及している．特に，大量の音楽データを用いた自己教師あり学習（self-supervised learning; SSL）により，高品質な音楽生成や高い汎用性を示すモデルが次々と提案されており，これらは音楽基盤モデル（music foundation models）と総称される [1], [2], [3], [4], [5]．代表例として，高精細な音響生成を実現する Jukebox [1]，テキスト条件付き生成を可能にする MusicLM [2]，軽量かつ高性能な生成を実現した MusicGen [3] などが挙げられ，生成のみならず分類や解析といった音楽理解タスクへの応用も報告されている [4]．

1 早稲田大学
Waseda University
2 慶應義塾大学
Keio University
3 東京大学
University of Tokyo
a) rin_sato@akane.waseda.jp
b) keitaro@aoni.waseda.jp
c) hayatobuti523@keio.jp
d) shinnosuke_takamichi@keio.jp
e) shigeo@waseda.jp

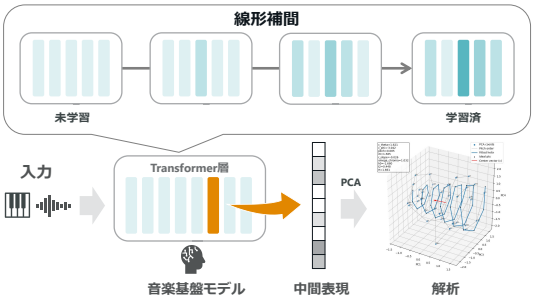

図 1　線形補間による内在音高螺旋の解析の流れ

一方で，音楽基盤モデルは極めて複雑な内部構造を有し，モデルがどのように音楽的知識や理論概念に対応する表現を獲得しているかは十分に解明されていない．このブラックボックス性は，出力の解釈や信頼性評価を困難にするだけでなく，意図した性質を持つモデル設計や制御を行う上での障壁となり得る．とりわけ音楽は，音高・和声・拍節・形式といった理論的枠組みが長年にわたり整備されてきた分野であり，これらの概念とモデル内部表現との対応関係を明らかにすることは，音楽基盤モデルの理解において本質的な課題である．

この問題意識のもと，近年では音楽基盤モデルの中間表現を解析する研究が進められている．我々は

先行研究 [6] において，音楽基盤モデルの中間表現に，人間の音高知覚に対応する音高螺旋構造 [7] が内在的に現れることを示した．具体的には，音高のみが異なる音楽信号を入力とした際に得られる中間表現を低次元空間に射影し，それらに対して螺旋モデルを当てはめることで，螺旋構造の明瞭さを定量化する指標である *Helicality* を導入した．その結果，音高のような抽象的な概念はモデルの深い層においてより顕著に表現されること，ならびにモデルサイズや階層構造の違いが，螺旋構造の明瞭さに影響を与えることが示された．以降では，このように中間表現空間に内在的に形成される音高螺旋構造を**内在音高螺旋**と呼ぶ．この結果は，学習済みモデル内部に音楽理論的構造と整合的な表現が自発的に形成されうることを示唆している．一方で，このような表現がパラメータ空間上のどの領域において，どのように形成されるのかについては，依然として十分に理解されていない．

学習過程を直接解析するためには，スクラッチからの学習を行い，学習中のチェックポイントを高い時間分解能で保存・解析することが理想的である．しかし，大規模音楽基盤モデルの学習には莫大な計算資源と長時間の計算が必要であり，これを再現することは現実的に困難である．さらに，既存の多くの公開モデルでは，学習途中のチェックポイントが公開されていないため，既存の学習済みモデルを用いて学習を新たに実行することなく表現形成過程を間接的に調査する手法の検討が必要である．

以上を踏まえ，本研究では未学習モデルと学習済みモデルのパラメータの間を補間することで，両者を連続的につなぐ仮想的なパラメータ遷移を構成し，その経路上における中間表現を解析する（図 1）．ここでいうパラメータ経路とは，モデルの重み空間上において，初期状態から学習済み状態へと連続的に変化する一連のパラメータ集合を指す．

さらに，単純な線形補間に基づく基準的な経路に加え，ラベルなし音響データに基づく勾配情報を用いて局所的に補正した経路も構成し，経路の設計が観測される内部表現に与える影響を比較する．本研究は，音楽基盤モデルにおける表現形成の過程をパラメータ空間上で捉え，学習を再実行することなく探索的に理解するための初期的検討として位置づけられ，モデル理解および将来的な制御性向上に資する基礎的知見の獲得を目的とする．

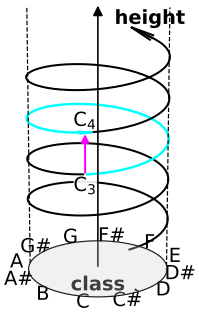

図 2　Shepard の音高螺旋構造 [7]

2. 関連研究

2.1 基盤モデルの内部解析

基盤モデルの内部表現を理解する試みは，自然言語処理を中心に活発化しており，中間表現に現れる構造を可視化・抽象化し，概念や推論との対応を検討する研究が報告されている [8], [9], [10]．例えば，言語基盤モデルが周期的概念を幾何学的構造として保持することが示されており [8]，内部表現の構造分析がモデル理解に有効であることが示唆される．

音楽領域においても，中間表現に含まれる音楽的属性を検証する研究が進められており，特にプロービング [11] に基づく分析が多い．これは，中間表現から線形分類器等で音高・和音・テンポなどの属性がどの程度予測可能かを調べることで，当該情報の符号化を間接的に評価する手法である．実際に，Jukebox や MusicGen 等が音楽理論的概念をどの程度捉えるかを検証する研究 [12] や，層が深くなるにつれて音高や和音のルート音がより識別的に表現されることを報告する研究 [13]，さらに多様な音楽的特徴が中間表現に含まれることを示す研究 [5] が存在する．

一方で，プロービングは表現に含まれる識別的情報の有無を評価する枠組みであり，中間表現が形成する幾何学的構造そのものに踏み込む研究は限定的である．我々の先行研究 [6] はこの点に着目し，音楽基盤モデルの中間表現が音高螺旋構造を示し得ることを報告したが，その形成過程については未解明である．本研究は，この形成過程の理解に向けた初期的検討として位置づけられる．

2.2 内在音高螺旋

音高は，高さ方向の連続性に対応する pitch height と，オクターブ周期性に基づく循環的性質を表す pitch class から構成される二次元的な知覚構造を有することが知られている．これら二つの側面を統合的に表現する枠組みとして，音高螺旋（pitch helix）が提案されている [7]（図 2）．近年の研究では，こ

のような音高の幾何学的構造が，明示的なラベルを用いず，音響信号のみから教師なしに抽出可能であることも報告されている [14], [15].

これに対して，我々の先行研究 [6] では，音楽信号そのものではなく，音楽基盤モデルが内部で形成する中間表現に着目し，モデル内部に音高螺旋構造が内在的に表現されているかを検討した．具体的には，学習済み音楽基盤モデルに対して音高のみが異なる N_{key} 種類の単音を入力として与え，各 Transformer 層から抽出された中間表現を主成分分析（PCA）により低次元空間へ射影した．その後，得られた 3 次元表現にパラメトリックな音高螺旋モデルを当てはめることで，音高螺旋構造の明瞭さを評価した．

音高インデックス t に対する螺旋モデル $\boldsymbol{y}(t) \in \mathbb{R}^3$ は，高さ変化係数 h_{pitch}，初期高さ h_0，初期半径 r_0，半径変化係数 r_{slope}，角周波数 ω_{chroma}，および位相 t_0 を用いて次式のように定式化される．

$$\boldsymbol{y}(t) = h(t) \cdot \boldsymbol{c} + r(t)\left(\cos\theta(t) \cdot \boldsymbol{u} + \sin\theta(t) \cdot \boldsymbol{v}\right)$$
$$\text{where } \begin{cases} h(t) = h_{\mathrm{pitch}} \cdot t + h_0 \\ r(t) = r_{\mathrm{slope}} \cdot t + r_0 \\ \theta(t) = \omega_{\mathrm{chroma}} \cdot (t - t_0) \end{cases} \tag{1}$$

ここで $\boldsymbol{c}, \boldsymbol{u}, \boldsymbol{v}$ は $\mathbb{R}^3$ の正規直交基底であり，$\boldsymbol{c}$ は螺旋の中心軸，$\boldsymbol{u}, \boldsymbol{v}$ は回転平面を定義する．なお，r_{slope} を導入することで，半径が音高に応じて変化し得る円錐状の螺旋も表現できる．

さらに，PCA により得られた特徴量群と螺旋モデルの適合度を定量的に評価するため，先行研究 [15] で提案された Helicality スコア を用い，3 次元特徴量 $\{\boldsymbol{x}_t\}_{t=1}^{N_{\mathrm{key}}}$ に対して以下の指標を定義する．

$$\text{Helicality} = \left(\frac{1}{N_{\mathrm{key}}} \sum_{t=1}^{N_{\mathrm{key}}} |\boldsymbol{x}_t - \boldsymbol{y}(t)|^2 \right)^{-1} \tag{2}$$

この指標を用いることで，層ごとおよびモデルごとに内在音高螺旋の明瞭さを比較可能とした．

その結果，Jukebox [1] や MusicGen [3] といった音楽基盤モデルにおいて，音高の螺旋構造が主として深い層で明瞭に出現することが確認された．また，モデルサイズや階層的アーキテクチャの違いによって，螺旋構造の局在性や明瞭さが変化することが示され，音高のような抽象的音楽概念がモデル内部で段階的に形成される可能性が示唆された．

一方で，先行研究 [6] は学習済みモデルを対象とした静的解析に基づくものであり，未学習状態から学習済み状態へと連続的に変化するパラメータ空間上のどの領域において，内在音高螺旋が顕在化するのかについては未解明である．本研究ではこの点に着目し，未学習モデルと学習済みモデルの間にパラメータ経路を構成し，Helicality の変化を解析することで，表現形成の段階性を検討する．

2.3 モデルマージとパラメータ空間の幾何

近年，複数の学習済みモデルを再学習なしで統合する**モデルマージ**（*model merging*）の研究が盛んに行われている [16]．これらの手法は，元データが非公開である場合や，大規模モデルの再学習・追加学習に要する計算資源が制約される場合など，学習過程を再実行することが困難な状況において，複数のモデルが獲得した知識をパラメータ空間上の操作として統合する枠組みを提供するものである．

モデルマージ研究の多くは，同一アーキテクチャかつ共通の事前学習初期値をもつ**学習済みモデル同士**を対象とし，重み平均（model soup） [17]，タスクベクトルに基づく加減算（task arithmetic） [18]，あるいはパラメータ干渉を抑制するための剪定・再重み付け手法などを提案してきた [19], [20]．これらの研究は，複数タスクやドメインの能力を単一モデルに統合する実用的手法として高い有効性を示している一方で，統合後の性能を主たる評価対象としており，学習過程における内部表現の形成そのものを分析対象とするものではない．

一方で，損失関数上で性能劣化を伴わずにモデル間を接続する経路の存在を調べる *mode connectivity* の研究 [21], [22] は，ニューラルネットワークのパラメータ空間を幾何学的対象として捉える視点を提供している．特に，線形補間や曲線補間を用いて異なる学習結果が低損失経路で接続可能であることが示されており，この知見はモデルマージ手法の理論的基盤とも密接に関連している．ただし，mode connectivity 研究においても，主な関心は最終的な性能保存や接続可能性にあり，表現がどの段階でどのように顕在化するかという学習過程の解釈には十分踏み込まれていない．

これらに対し，本研究は未学習モデル（初期化状態）と学習済みモデルの間にパラメータ空間上の経路を構成し，その経路に沿って内部表現がどのように形成されるかを解析対象とする点に焦点を当てる．これは，学習済みモデル同士の統合を目的とする従来のモデルマージ研究とは異なり，ニューラルネットワークが表現を獲得していく過程を幾何学的に可視化・定量化する試みであり，モデル理解や制御性の向上に資する新たな視点を提供するものである．

3. 分析手法

本研究では，未学習モデルと学習済みモデルの間に連続的なパラメータ経路を構成し，経路上の各点における中間表現を先行研究 [6] の枠組みに基づいて解析することで，内在音高螺旋が顕在化する段階を調査する．

さらに，単純な線形補間に基づく経路に加え，ラベルなし音響データに基づく勾配情報を用いて補正した経路を構成し，経路の取り方が観測結果に与える影響を比較する．

3.1 未学習モデルと学習済みモデルの線形補間

未学習モデルのパラメータを $\boldsymbol{\theta}_{\text{init}}$，学習済みモデルのパラメータを $\boldsymbol{\theta}_{\text{trained}}$ とする．線形補間により，補間係数 $\alpha \in [0,1]$ に対する中間パラメータ $\boldsymbol{\theta}(\alpha)$ を次式で定義する．

$$\boldsymbol{\theta}(\alpha) = (1-\alpha)\boldsymbol{\theta}_{\text{init}} + \alpha\boldsymbol{\theta}_{\text{trained}} \qquad (3)$$

ここで $\alpha = 0$ は未学習モデル，$\alpha = 1$ は学習済みモデルに対応する．本研究では，未学習モデルと学習済みモデルを結ぶ最も単純なパラメータ空間上の接続として，この線形補間系列を用いて各 α における中間表現の変化を解析する．

このような線形補間は真の学習軌跡を正確に再現するものではないが，ニューラルネットワークのパラメータ空間が高次元かつ冗長であること，および学習済み解同士が低損失経路で接続可能であることが先行研究により示されている [21], [23], [24]．これらの知見を踏まえ，本研究では追加の仮定を導入しない解析の初期検討として線形補間を採用する．

3.2 勾配情報に基づく補正経路

線形補間によって得られた各 $\boldsymbol{\theta}(\alpha)$ に対し，ラベルを付与しない音響データ集合 $\mathcal{D}$ を用いて，モデル実装において定義されている生成損失 $\mathcal{L}(\boldsymbol{\theta};\mathcal{D})$ を評価し，対応する勾配 $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(\alpha);\mathcal{D})$ を計算する．ここで $\mathcal{L}$ は，離散化された音響表現に対する次トークン予測に基づくクロスエントロピー損失であり，各モデルが学習時に用いる尤度最大化目的関数に対応する．

本研究では，勾配降下法における 1 ステップ分の更新を近似的に導入し，線形補間点を学習方向へ局所的に補正したパラメータ $\tilde{\boldsymbol{\theta}}(\alpha)$ を次式で定義する．

$$\tilde{\boldsymbol{\theta}}(\alpha) = \boldsymbol{\theta}(\alpha) - \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(\alpha);\mathcal{D}) \qquad (4)$$

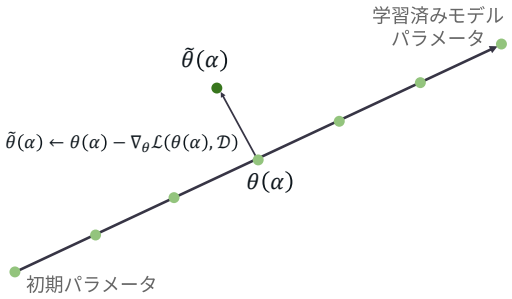

図 3 補正の概略図

この操作は，線形補間によって定義されたパラメータ経路に対し，局所的に学習時の更新方向の情報を付加するものであり，図 3 に示すように，各補間点を学習方向へ微小に変位させることに相当する．これにより，純粋な線形補間と比較して，実際の学習に伴う表現変化に近い挙動が観測されることが期待される．

本研究では，この局所補正が線形補間に基づく観測結果を質的に変化させるかを検討するため比較条件の一つとして導入する．以降では，式 (3) に基づく系列を線形補間，式 (4) に基づく系列を勾配補正補間と呼ぶ．

3.3 内在音高螺旋の評価指標

各補間点（線形補間/勾配補正補間）について，先行研究 [6] に従い，中間表現に対して主成分分析（PCA）により 3 次元へ低次元化を行い，パラメトリックな音高螺旋モデル $\boldsymbol{y}(t)$（式 (1)）をフィッティングする．得られた 3 次元特徴量 $\{\boldsymbol{x}_t\}_{t=1}^{N_{\text{key}}}$ と螺旋モデルの適合度は，Helicality スコア（式 (2)）により定量化する．本研究では，Helicality の α 依存性を観測することで，内在音高螺旋がどの α 近傍で顕在化するかを評価する．

3.4 パープレキシティに基づく補助的評価

経路の妥当性を補助的に検討するため，線形補間と勾配補正補間の各補間点における生成モデルの損失に基づくパープレキシティ（perplexity; PPL）を算出する．具体的には，評価用データ $\mathcal{D}_{\text{eval}}$ に対する平均損失 $\mathbb{E}[\mathcal{L}]$ を用い，

$$\text{PPL} = \exp(\mathbb{E}[\mathcal{L}(\boldsymbol{\theta};\mathcal{D}_{\text{eval}})]) \qquad (5)$$

により定義する．

PPL の比較により，勾配補正が予測性能（損失）の改善方向に寄与しているか，ならびに線形補間経路との性質の違いを確認する．

4. 実験的評価

4.1 実験条件

本研究では，先行研究 [6] で用いられた解析手順を基礎として，音楽基盤モデルのパラメータ空間においてモデルが学習済み状態へと近づく過程のどの段階で音高に関する螺旋構造が顕在化するのかを実験的に検証した．本節では，本研究で採用した実験条件について詳細を述べる．

4.1.1 モデル条件

解析対象となる音楽基盤モデルとして，MusicGen [3] のモデルサイズが異なる small（300M）と large（3.3B），Jukebox [1] の bottom-level decoder（1B）の計 3 種類を用いた．モデルの入手方法は先行研究 [6] と同様である．

未学習モデルについては，各モデルについて公開されている実装に基づいて初期化したモデルを用いる．

MusicGen については，Hugging Face Transformers [25] ライブラリを用いて事前学習済みモデルの設定（decoder configuration）のみを取得し，decoder-only 構成のモデルを新たに構築した．その後，公式実装で用いられている初期化規則に従い，線形層および埋め込み層の重みを正規分布により再初期化することで，学習前状態のモデルを生成した．

Jukebox については OpenAI により公開されている GitHub リポジトリ[*1]に基づいてパラメータを初期化したモデルを使用した．この手順により，学習済みモデルと同一のアーキテクチャを保ちつつ，パラメータが学習に依存しない初期状態から解析を行っている．

補助的解析として perplexity および勾配補正の影響については，MusicGen (Small) を代表モデルとして実施した．これは，先行研究 [6] において MusicGen の各モデルの中で最も高い Helicality が報告されていることに加え，MusicGen が学習過程においてラベルを部分的にドロップする訓練を含み，入力音響のみに基づく表現学習が行われている点を踏まえたものである．

4.1.2 データ条件

音高螺旋構造の解析には，先行研究 [12] で提案された合成音楽理論データセット SynTheory を用いた．このデータセットは，先行研究 [6] と同様に，本研究の解析目的に適合するよう一部調整を施した上で使用している．

Jukebox や MusicGen といった音楽基盤モデルの学習データは非公開であり，学習時に用いられた音源分布を直接再現することは困難である．そのため本研究では，公開データセットの中から，多様な楽器音・音高・音色を含み，音楽的性質がこれらのモデルの学習データと近いと考えられる Music4All [26] を代替データとして採用した．

本実験では，勾配計算においてラベル情報は使用せず，音響信号のみを入力として用いることで，特定タスクへの適合ではなく，モデル内部の表現構造およびパラメータ空間上の挙動を解析対象としている．

4.1.3 フィッティング条件

式 (2) に基づく Helicality スコアのパラメータ最適化手法は，先行研究 [6] と同一の設定を採用した．各条件において試行回数 1000 回の最適化試行を異なる乱数シードを用いて 3 回行い，得られた Helicality スコアのうち最も高い値を，当該条件を代表するスコアとして記録した．

4.2 比較設定

本研究では，補間係数 α のサンプリング方法および補間経路の定義に応じて，以下の 4 条件を設定し，Helicality およびパープレキシティ（PPL）の観点から比較を行う．

4.2.1 線形補間（全体サンプリング）

$\alpha \in [0, 1]$ を等間隔に分割した $\alpha = \{0, 0.25, 0.5, 0.75, 1.0\}$ を補間点として用いる．この条件は，未学習モデルから学習済みモデルへの遷移をパラメータ空間全体の観点から俯瞰するための基準的な設定である．

4.2.2 線形補間（学習済み近傍サンプリング）

$\alpha \in [0.9, 1.0]$ を高解像度に分割した $\alpha = \{0.90, 0.91, \ldots, 0.99, 1.00\}$ を補間点として用いる．学習済み近傍において表現変化が集中する可能性を考慮し，$\alpha \approx 1$ 付近での内在音高螺旋構造の立ち上がり過程を詳細に観測することを目的とする．

4.2.3 勾配補正補間（全体サンプリング）

線形補間（全体サンプリング）と同一の α 点に対し，式 (4) に基づく勾配補正を施したパラメータ $\tilde{\boldsymbol{\theta}}(\alpha)$ を用いる．これにより，線形補間経路の局所近傍に学習方向の一次情報を付与した場合に，内在音高螺旋の顕在化段階や明瞭さがどのように変化するかを検証する．

4.2.4 パープレキシティによる補助的解析

補間経路の性質を補助的に評価するため，線形補間（全体サンプリング）および勾配補正補間（全体サンプリング）の各補間点に対してパープレキシティ（PPL）を算出する．PPL は予測損失に基づく指標

[*1] https://github.com/openai/jukebox

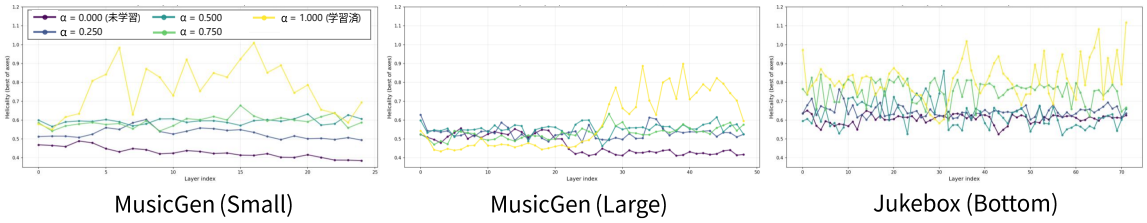

図 4 線形補間における Helicality の α 依存性

であり，α に伴う性能変化の大きさを表す．これにより，Helicality の変化が予測性能の改善と同時に生じているのか，あるいは性能とは独立に生じているのかを切り分ける．

4.3 評価指標

本研究では，内在音高螺旋の顕在化段階とその明瞭さを評価するため，Helicality スコアおよびパープレキシティ（PPL）の観点から評価を行う．以下では，各指標の定義と集計方法を述べる．

4.3.1 Helicality スコア

各補間点（線形補間/勾配補正補間）および各層において，中間表現を PCA により 3 次元に低次元化し，音高螺旋モデル（式 (1)）をフィッティングする．得られた 3 次元特徴量 $\{\boldsymbol{x}_t\}_{t=1}^{N_{\text{key}}}$ とモデル $\boldsymbol{y}(t)$ の適合度は，Helicality スコア（式 (2)）により定量化する．本研究では，各条件につき複数回の最適化試行を行い，得られた Helicality スコアの最大値を当該条件の代表値として用いる（§4.1.3）．

4.3.2 パープレキシティ（PPL）

経路の性質を補助的に評価するため，線形補間（全体サンプリング）および勾配補正補間（全体サンプリング）の各補間点に対して PPL を算出する．PPL は式 (5) に基づき，評価用データ $\mathcal{D}_{\text{eval}}$ に対する平均損失 $\mathbb{E}[\mathcal{L}]$ から算出する．これにより，勾配補正が損失の観点で改善方向に働くか，および Helicality の変化と損失の変化の関係性を確認する．

4.4 結果

本節では，§4.2 で定義した 4 条件（線形補間（全体サンプリング），線形補間（学習済み近傍サンプリング），勾配補正補間（全体サンプリング），PPL による補助的解析）に関する結果を示す．結果の可視化は図として提示し，本文では観測された傾向と比較の要点を述べる．

4.4.1 Helicality の α 依存性（全体サンプリング）

図 4 に，線形補間（全体サンプリング）における Helicality の α 依存性を示す．

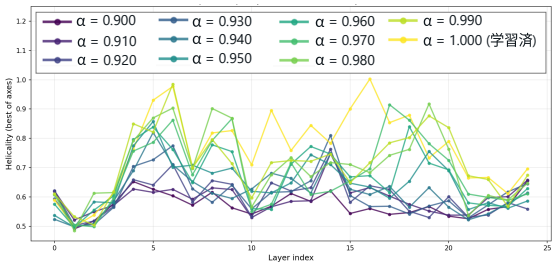

図 5 MusicGen (Small) について学習済み近傍サンプリングにおける Helicality の α 依存性（$\alpha \in [0.9, 1.0]$）

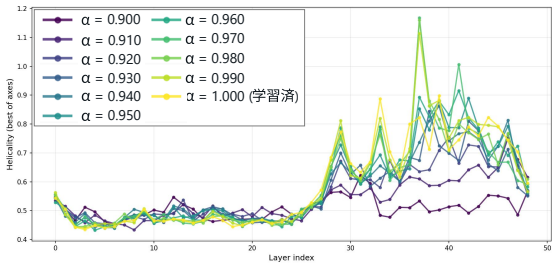

図 6 MusicGen (Large) について学習済み近傍サンプリングにおける Helicality の α 依存性（$\alpha \in [0.9, 1.0]$）

多くの層において，$\alpha < 1$ の範囲では Helicality は低い値に留まり，学習済み点 $\alpha = 1$ において急激な上昇が観測された．この結果は，内在音高螺旋構造が学習過程の後半で集中的に形成される可能性を示唆している．

なお，未学習点（$\alpha = 0$）においても Helicality は 0.4–0.5 程度を示し，ランダム変数に対する基準値 0.371 ± 0.02（[27]）をわずかに上回った．この差は小さいものの，完全に無構造な表現ではなく，Transformer による多層写像（自己注意・残差接続・正規化など）そのものが音高変化に対して弱い幾何学的整合性を誘起している可能性を示唆する．ただし，本結果は初期化規則や PCA の不変性の影響も受け得るため，初期化分散や入力集合を変えた統制実験により検証する必要がある．

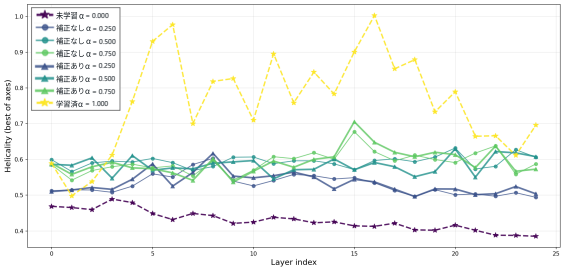

図 7 MusicGen (Small) における線形補間と勾配補正補間の Helicality の α 依存性

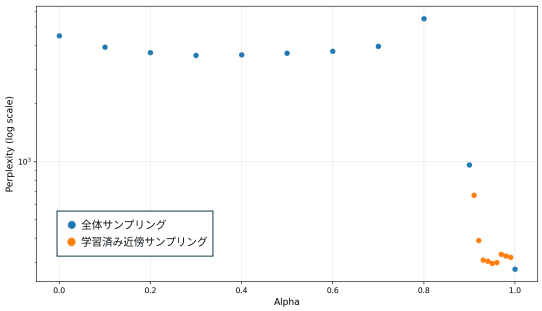

図 8 MusicGen (Small) における Perplexity の α 依存性

4.4.2 層方向の傾向と学習による反転

未学習モデルでは，層が深くなるにつれて Helicality が低下する傾向が観測された一方，学習済みモデルでは深い層ほど Helicality が高くなるという逆の傾向が確認された．この反転は，初期状態では深層ほど表現が入力差（音高差）を保持しにくいのに対し，学習を通して深層が音高のような抽象的概念を安定に符号化する役割を獲得していく可能性を示唆する．すなわち，深層ほど Helicality が上がるという事実は，音高螺旋が入力に近い層で自然に現れる構造ではなく，学習を通して深層側で選択的に強調・整列される構造である可能性を示す．

4.4.3 学習済み近傍の詳細解析（学習済み近傍サンプリング）

図 5 および 6 に，MusicGen の Small および Large モデルでの $\alpha \in [0.9, 1.0]$ における学習済みモデルの近傍における線形補間の結果を示す．全体をサンプリングした際には観測されなかった Helicality の漸増的な変化が，学習済み近傍では複数の層において明瞭に確認された．

特に，学習済みモデルで高い Helicality を示す層では，α の増加に伴って Helicality が段階的に増大し，学習済みモデルの値へと連続的に近づく傾向が見られる．このことは，音高螺旋構造の顕在化が特定の一点で不連続に生じるのではなく，$\alpha \approx 1$ 近傍において層依存的かつ連続的に進行している可能性を示唆する．

4.4.4 勾配補正の影響

図 7 に，勾配補正補間の結果を示す．本設定では，Helicality の定性的な挙動は線形補間の場合と大きく変化せず，螺旋構造の明瞭さに顕著な差は観測されなかった．これは，各補間点において勾配補正を 1 ステップのみ適用しているため，パラメータ更新量が限定的であることに起因すると考えられる．

MusicGen の Large モデルでは，後半の一部の層において，学習済み点（$\alpha = 1$）に到達する前の補間点で Helicality が学習済み点を上回る区間が観測された．この現象は，Helicality が生成性能そのものではなく音高に関する幾何学的整合性の明瞭さを測る指標であり，学習目的（次トークン予測）と必ずしも単調に対応しないことを示唆する．

すなわち，大規模モデルでは学習の進行に伴い音高表現が一貫して強化されるだけでなく，他の音楽属性（音色・和声・文脈情報など）との共表現化により，音高に特化した幾何が相対的に見えにくくなる段階が存在する可能性がある．今後は，層ごとの PPL や他属性の指標と併せて解析することで，音高螺旋が最も明瞭な点がどの要因で移動するのかを検証する必要がある．

4.4.5 Perplexity の α 依存性

図 8 に，線形補間 (全体サンプリング) における PPL の α 依存性を示す．

PPL は，α が学習済み点に近づくにつれて急激に低下する傾向を示し，Helicality の立ち上がりと時期的に対応していることが確認された．この結果は，内在音高螺旋の顕在化が，モデルの予測性能の向上と同時期に生じている可能性を示唆している．

さらに図 8 では，PPL の低下が $\alpha \approx 0.8$ 付近から急激になる傾向が見られる．一般に学習過程では，対数軸上で損失（あるいは PPL）が概ね滑らかに低下することが多いことを踏まえると，$\alpha \in [0, 0.8]$ の領域は学習の時間進行を素直に反映した区間というより，初期解から学習済み解への幾何学的接続（線形補間）に特有の経路である可能性が高い．この観点から，本研究の α は学習ステップの代替変数ではなく，あくまでパラメータ空間上の連続経路のパラメータとして解釈すべきであり，PPL の変化が急になる領域を, 学習終盤に対応する候補領域として扱うのが妥当である．

5. まとめ

本研究では，音楽基盤モデルにおける内在音高螺旋が，未学習状態から学習済み状態へ至るパラメー

タ空間上のどの領域で顕在化するのかを調べるため，未学習モデルと学習済みモデルを連続的に結ぶパラメータ経路を構成し，経路上の中間表現を解析した．

線形補間に基づく全体サンプリングおよび学習済み近傍サンプリングの結果から，内在音高螺旋は補間全体に一様に現れるのではなく，学習済みモデル近傍において層依存的かつ連続的に顕在化することが確認された．特に，学習済み近傍を高解像度に観測することで，全体サンプリングでは不連続に見えていた Helicality の立ち上がりが，実際には緩やかな変化として進行している可能性が示された．

さらに，ラベルなし音響データに基づく勾配情報を用いて，線形補間点を局所的に補正した経路を導入し，経路設計が観測結果に与える影響を検討した．その結果，本設定における 1 ステップの勾配補正は，Helicality の定性的な傾向を大きく変化させないことが確認された．このことは，単純な線形補間に対する局所的一次補正のみでは，実際の学習過程に伴う表現形成を十分に捉えきれない可能性を示唆している．

以上より，本研究は学習を再実行することなく音楽基盤モデルの表現形成過程をパラメータ空間上の連続経路として捉える枠組みの有効性を示すとともに，同時に，線形補間を基準とした単純な経路設計の限界を明らかにした点に意義がある．今後は，より学習過程に整合的なパラメータ経路の構成法や，勾配情報をより本質的に取り込む手法の検討を通して，内在音高螺旋を含む音楽的表現がどのように形成されるのかについてより明確に記述できる枠組みの構築を目指す．

謝辞： 本研究は，JST 創発的研究支援事業 JPMJFR226V の支援を受けて実施した．

参考文献

[1] Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A. and Sutskever, I.: Jukebox: A generative model for music, *arXiv preprint arXiv:2005.00341* (2020).

[2] Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N. and Frank, C.: MusicLM: Generating Music From Text, *arXiv arXiv:2301.11325* (2023).

[3] Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y. and Défossez, A.: Simple and Controllable Music Generation, *Advances in Neural Information Processing Systems* (2023).

[4] Li, Y., Yuan, R., Zhang, G., Ma, Y., Chen, X., Yin, H., Xiao, C., Lin, C., Ragni, A., Benetos, E. et al.: MERT: Acoustic music understanding model with large-scale self-supervised training, *arXiv preprint arXiv:2306.00107* (2023).

[5] Liao, W.-H., Takida, Y., Ikemiya, Y., Zhong, Z., Lai, C.-H., Fabbro, G., Shimada, K., Toyama, K., Cheuk, K. W., Martínez-Ramírez, M. A., Takahashi, S., Uhlich, S., Akama, T., Choi, W., Koyama, Y. and Mitsufuji, Y.: Music Foundation Model as Generic Booster for Music Downstream Tasks, *TMLR*, (online), available from ⟨https://openreview.net/forum?id=kHl4JzyNzF⟩ (2025).

[6] 八木 颯斗，高道 慎之介：音楽基盤モデルは音高情報を螺旋構造に埋め込むか？，情報処理学会 音楽情報科学研究会 (2025).

[7] Shepard, R. N.: Geometrical Approximations to the Structure of Musical Pitch, *Psychological Review*, Vol. 89, No. 4, pp. 305–333 (1982).

[8] Engels, J., Michaud, E. J., Liao, I., Gurnee, W. and Tegmark, M.: Not all language model features are one-dimensionally linear, *ICLR* (2025).

[9] Liu, Z., Kitouni, O., Nolte, N., Michaud, E. J., Tegmark, M. and Williams, M.: Towards Understanding Grokking: An Effective Theory of Representation Learning, *NeurIPS* (Oh, A. H., Agarwal, A., Belgrave, D. and Cho, K., eds.), (online), available from ⟨https://openreview.net/forum?id=6at6rB3IZm⟩ (2022).

[10] Heinzerling, B. and Inui, K.: Monotonic Representation of Numeric Attributes in Language Models, *Proceedings of the 62nd ACL (Volume 2: Short Papers)* (Ku, L.-W., Martins, A. and Srikumar, V., eds.), Bangkok, Thailand, Association for Computational Linguistics, pp. 175–195 (online), DOI: 10.18653/v1/2024.acl-short.18 (2024).

[11] Alain, G. and Bengio, Y.: Understanding intermediate layers using linear classifier probes, *ICLR* (2017).

[12] Wei, M., Freeman, M., Donahue, C. and Sun, C.: Do Music Generation Models Encode Music Theory?, *ISMIR* (2024).

[13] Ma, W., Li, X. and Xia, G.: Do music LLMs learn symbolic concepts? A pilot study using probing and intervention, *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, (online), available from ⟨https://openreview.net/forum?id=uvzw0gS0Nn⟩ (2024).

[14] Lostanlen, V., Sridhar, S., McFee, B., Farnsworth, A. and Bello, J. P.: Learning the Helix Topology of Musical Pitch, *IEEE ICASSP*, pp. 11–15 (online), DOI: 10.1109/ICASSP40776.2020.9053644 (2020).

[15] Sridhar, S. and Lostanlen, V.: Helicality: An Isomap-based Measure of Octave Equivalence in Audio Data, *ISMIR* (2020).

[16] Ruan, W., Yang, T., Zhou, Y., Liu, T. and Lu, J.: From Task-Specific Models to Unified Systems: A Review of Model Merging Approaches, *arXiv preprint arXiv:2503.08998* (2025).

[17] Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S. et al.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, *International conference on machine learning*, PMLR, pp. 23965–23998

(2022).
[18] Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H. and Farhadi, A.: Editing models with task arithmetic, *arXiv preprint arXiv:2212.04089* (2022).
[19] Yadav, P., Tam, D., Choshen, L., Raffel, C. A. and Bansal, M.: Ties-merging: Resolving interference when merging models, *Advances in Neural Information Processing Systems*, Vol. 36, pp. 7093–7115 (2023).
[20] Yu, L., Yu, B., Yu, H., Huang, F. and Li, Y.: Language models are super mario: Absorbing abilities from homologous models as a free lunch, *Forty-first International Conference on Machine Learning* (2024).
[21] Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P. and Wilson, A. G.: Loss surfaces, mode connectivity, and fast ensembling of dnns, *Advances in neural information processing systems*, Vol. 31 (2018).
[22] Adilova, L., Andriushchenko, M., Kamp, M., Fischer, A. and Jaggi, M.: Layer-wise linear mode connectivity, *ICLR* (2024).
[23] Entezari, R., Sedghi, H., Saukh, O. and Neyshabur, B.: The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks, *International Conference on Learning Representations* (2022).
[24] Zhou, Z., Chen, Z., Chen, Y., Zhang, B. and Yan, J.: On the Emergence of Cross-Task Linearity in Pretraining-Finetuning Paradigm, *ICML* (2024).
[25] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. and Rush, A. M.: Transformers: State-of-the-Art Natural Language Processing, *EMNLP*, Association for Computational Linguistics, pp. 38–45 (online), available from ⟨https://www.aclweb.org/anthology/2020.emnlp-demos.6⟩ (2020).
[26] Santana, I. A. P., Pinhelli, F., Donini, J., Catharin, L., Mangolin, R. B., Feltrim, V. D., Domingues, M. A. et al.: Music4all: A new music database and its applications, *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, IEEE, pp. 399–404 (2020).
[27] 八木 颯斗，高道 慎之介，佐藤 りん，田中 啓太郎，森島 繁生：音楽基盤モデルにおける音響特徴と内在音高螺旋の関係，情報処理学会 音楽情報科学研究会 (2026).