

2025年8月 MUS研究会 音楽音響処理 (1)

音楽基盤モデルは 音高情報を螺旋構造に埋め込むか？

八木 颯斗¹，高道 慎之介^{1,2}

(1: 慶大, 2: 東大)



Takamichi Laboratory
慶應義塾大学 高道研究室

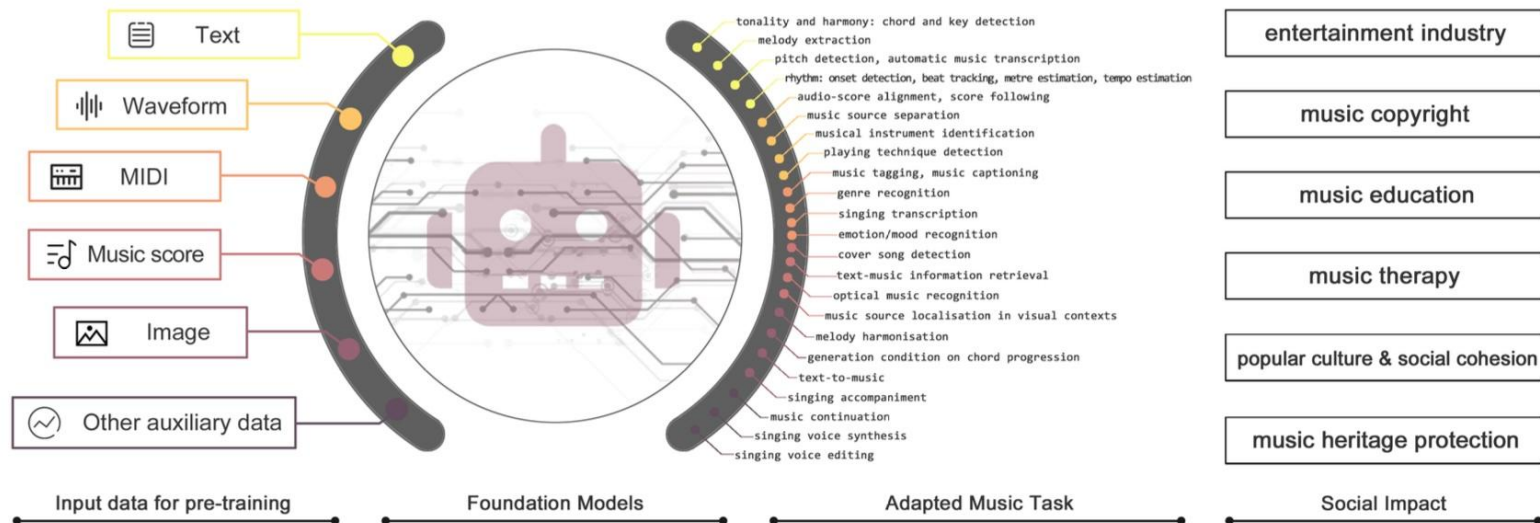
研究背景

基盤モデルとは

音楽基盤モデルは音高情報を螺旋構造に埋め込むか？

特定のタスクだけに特化して訓練されたモデルではなく
大規模な自己教師あり学習によって事前学習されることで**多数の下流タスクに
対応可能な汎用的な機械学習モデル**

例) BERT, GPT, Stable Diffusion, Jukebox, MusicGenなど



音高の表現

音楽基盤モデルは**音高情報**を**螺旋構造**に埋め込むか？

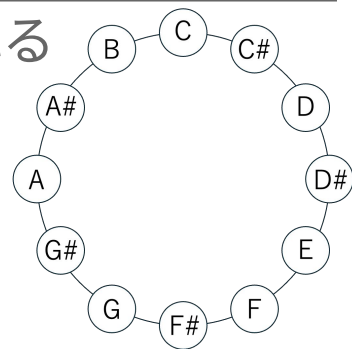
- 音高の知覚的表現は以下の**2つの心理的属性**から構成される

- ピッチクラス (pitch class)**

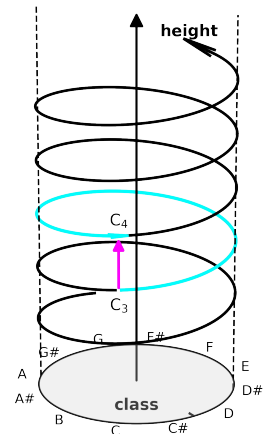
- ・ **トーンクロマ**とも呼ばれ、音高の**循環的な次元**を指す
- ・ 基本的に**12のピッチクラス**として定義

- ピッチハイト (pitch height)**

- ・ 音高の**垂直的な次元**であり、音を「低い」から「高い」へと順序付ける
- ・ 楽譜上では、どのオクターブに属するかによってこの高さが示される
- ・ 例) C4とC5は同じ「ド」の音でも、1オクターブの違いを表す



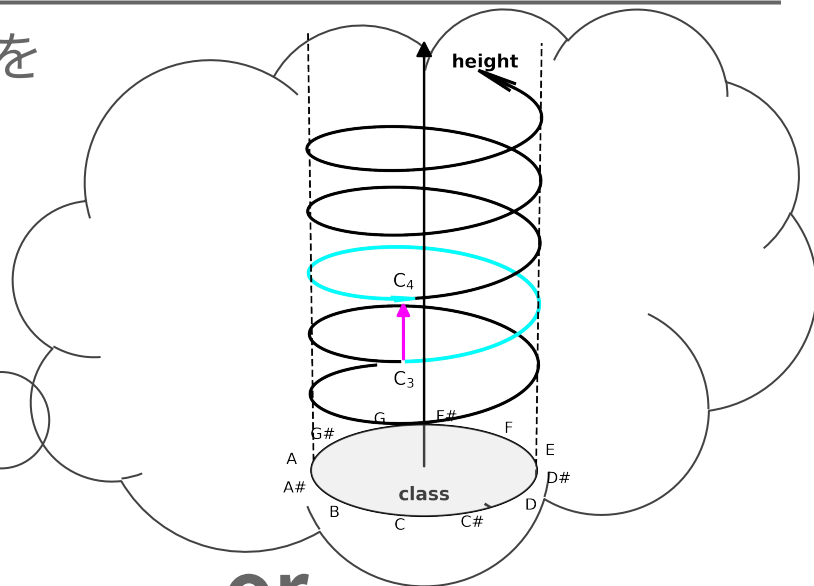
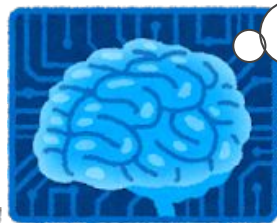
- 音高は**螺旋モデルで説明可能** [Sheard, 1982]



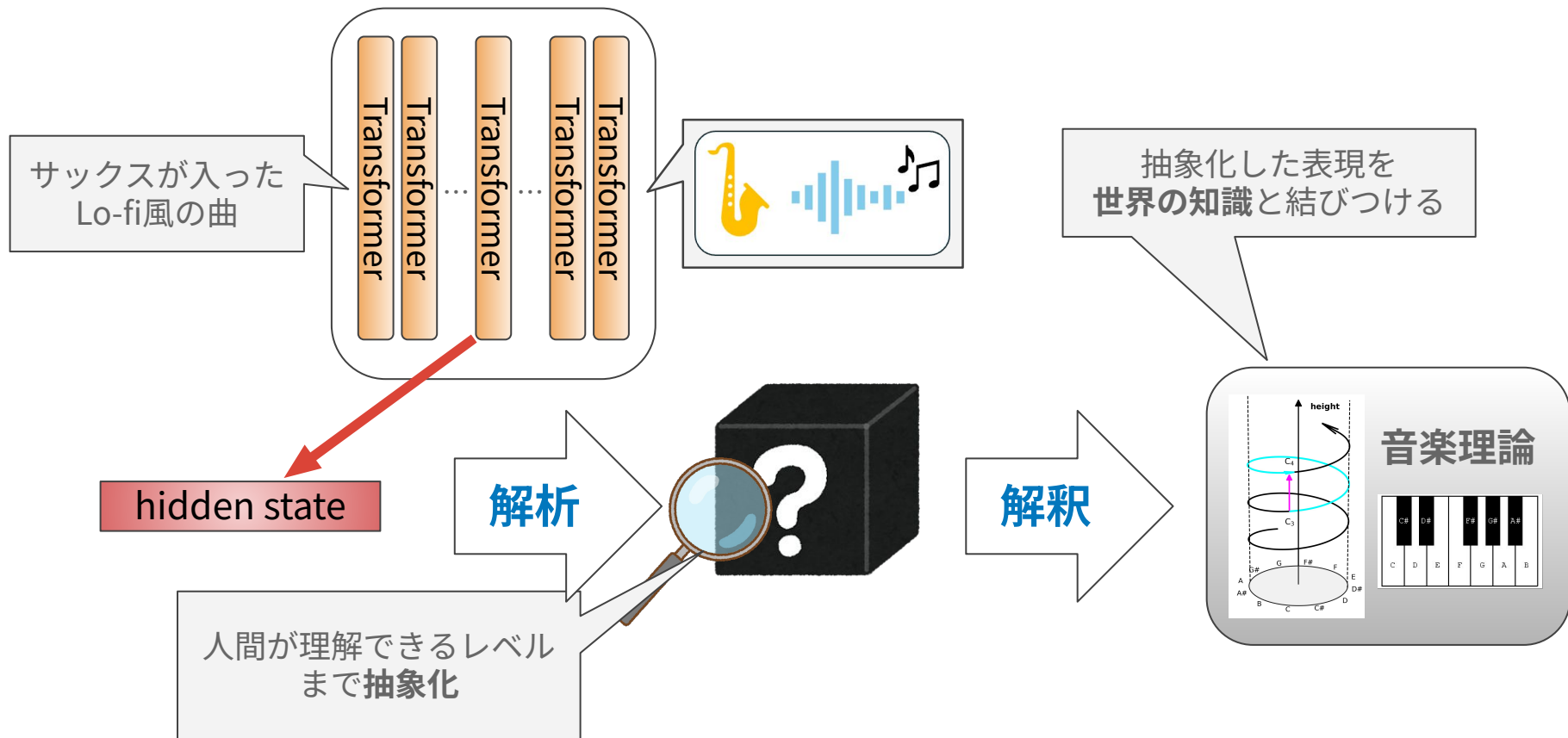
音高の表現

音楽基盤モデルは音高情報を螺旋構造に埋め込むか？

- 音楽基盤モデルはどのように音高情報を認識しているのだろう？
- 螺旋構造 or その他の幾何学的構造



内部解析と解釈



何が嬉しい？

- 透明性の確保
 - モデルが「本当に理解しているか」検証できる
 - 音高に着目 → 人間の知覚に近い構造を獲得しているかどうかを検証
- 音楽の知見を発展
 - 音楽の仮説を検証する手段になる可能性
 - 人間のピッチ知覚の真の構造を知るヒントになる
- 制御性の向上
 - 音楽的要素を特定し、細かく直感的に制御可能
 - 創作活動において欠かせないユーザーの主体性を確保に繋がる

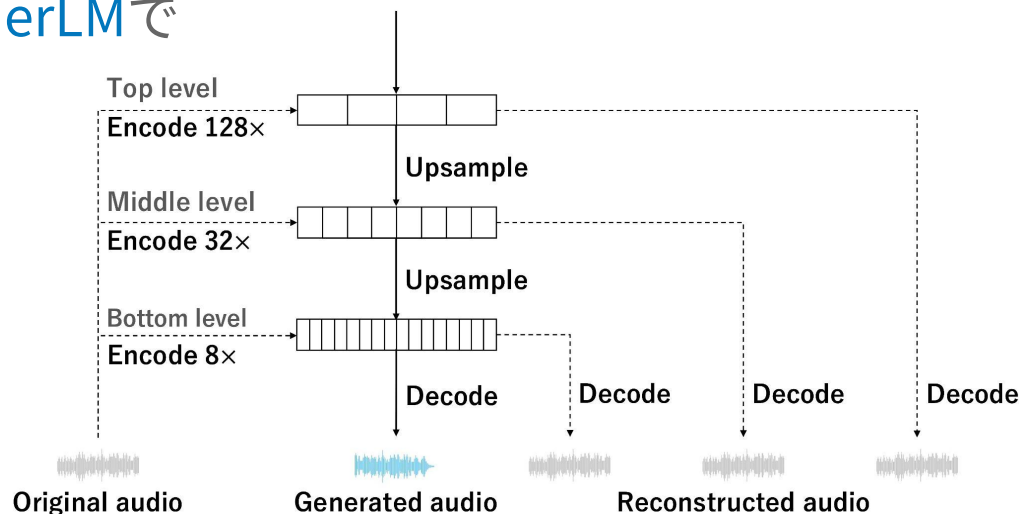
関連研究

本実験で使用する音楽基盤モデル

- Jukebox
- MusicGen

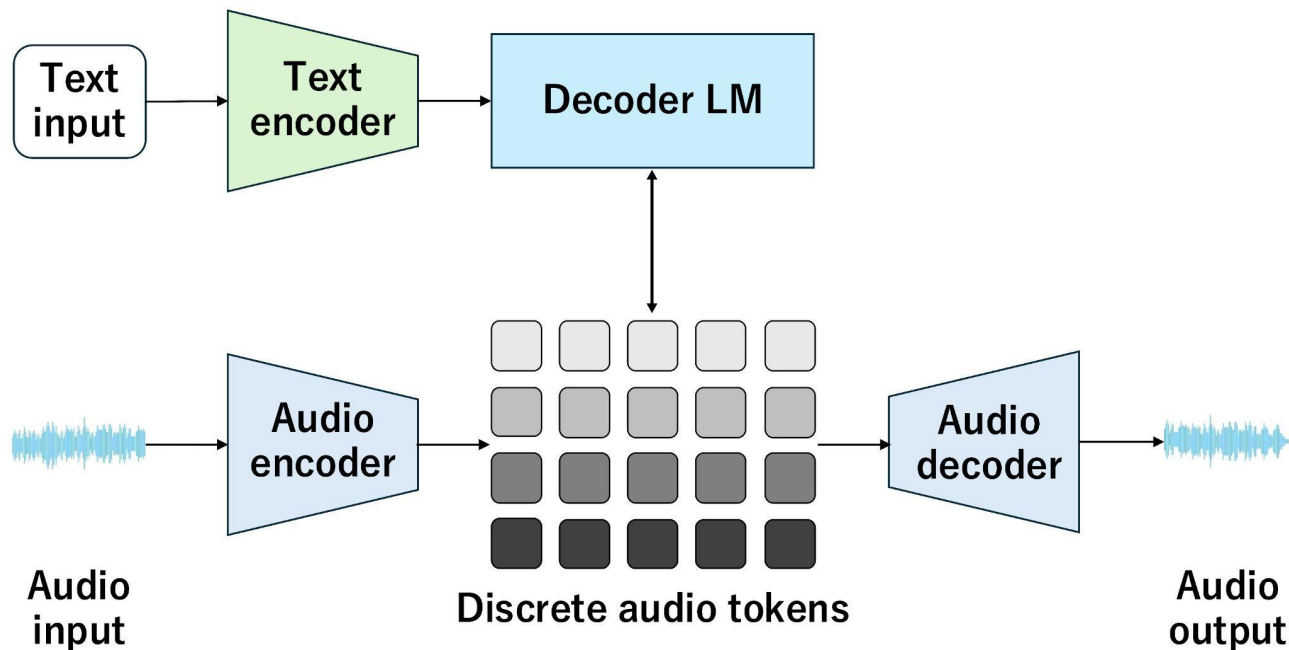
Jukebox：階層的な音楽生成モデル [Dhariwal, 2020]

- 3つのレベルのVQ-VAE [Van Den Oord, 2017] によって音楽信号を離散トークン化
 - Top：最も圧縮率が高い．大域な情報を捉える．
 - Middle：中間の圧縮率
 - Bottom：最も圧縮率が低い．局所的な情報を捉える．
- TransformerベースのDecoderLMで階層的にトークン生成



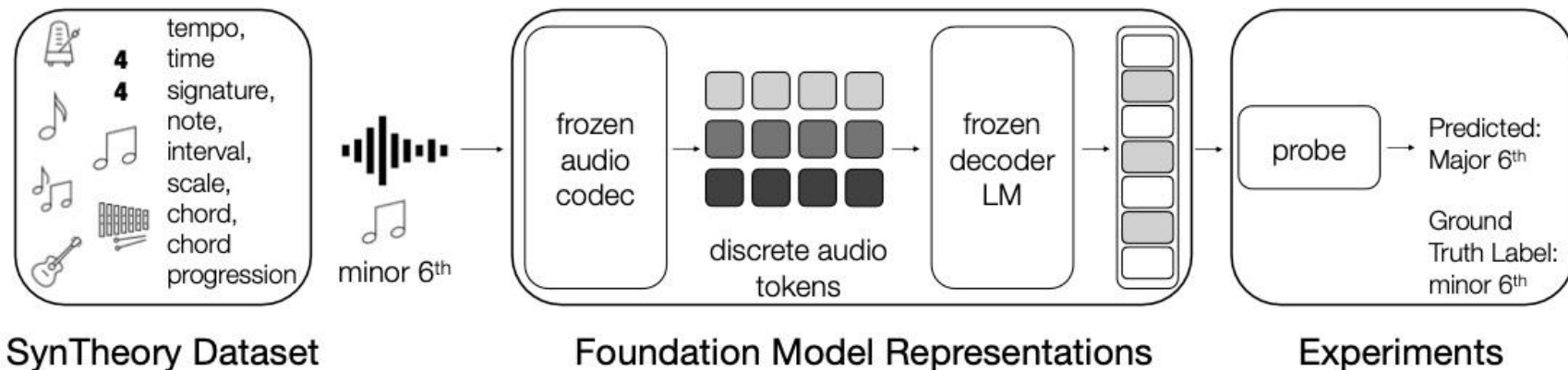
MusicGen：非階層的な音楽生成モデル [Copet, 2023]

- **EnCodec** [Défossez, 2022]によって音楽信号を離散トークン化
- **DecoderLM**でトークン列生成



音楽の基本概念に対しての解析 [Wei, 2024]

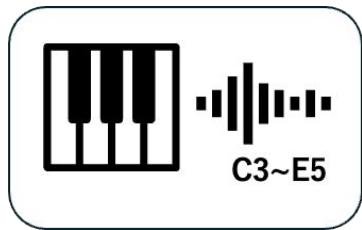
- テンポや音高、コードなど音楽の基本概念を対象
- モデルから抽出した中間表現を**小規模な識別器**に入力して学習
- 識別器の精度を確かめ、モデルが「どれだけ基本概念情報を保持しているか」を検証
- しかし、「どのような構造で表現されているか」は未知



提案手法

全体像

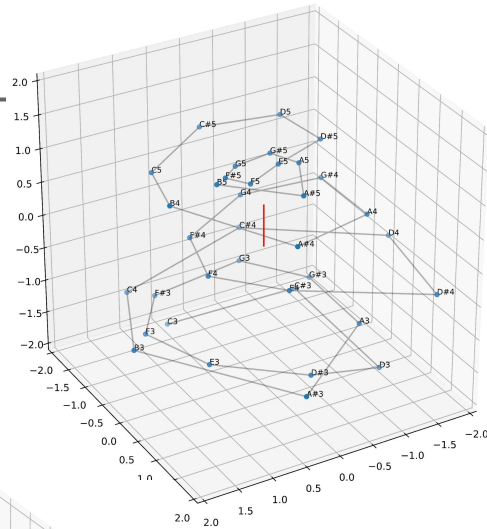
- **音高に関する中間表現の低次元化**



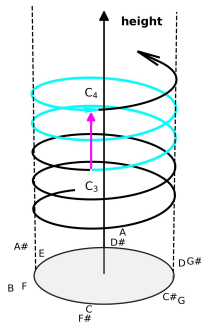
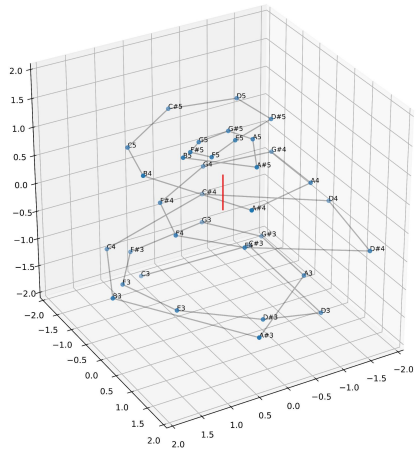
Foundation Model

中間表現
抽出

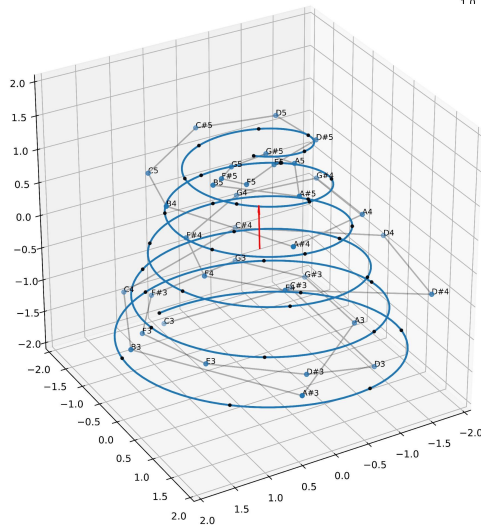
PCA



- 螺旋構造らしさの定量化



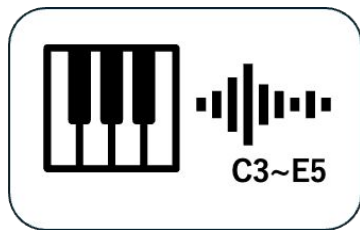
フィッティング



H = 2.60

全体像

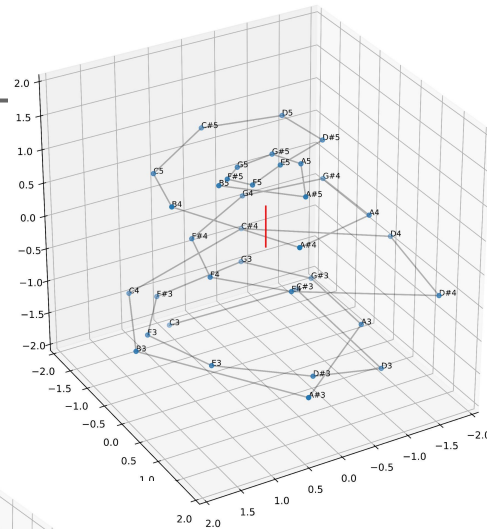
- **音高に関する中間表現の低次元化**



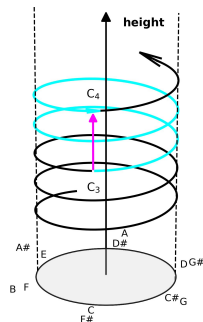
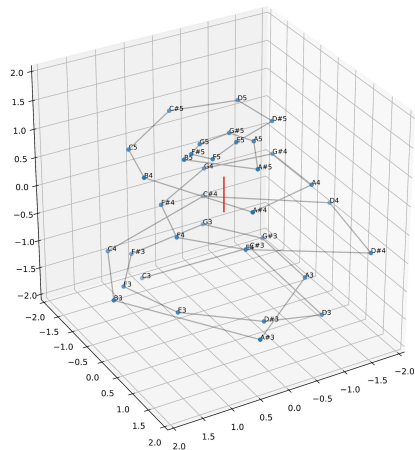
Foundation Model

中間表現
抽出

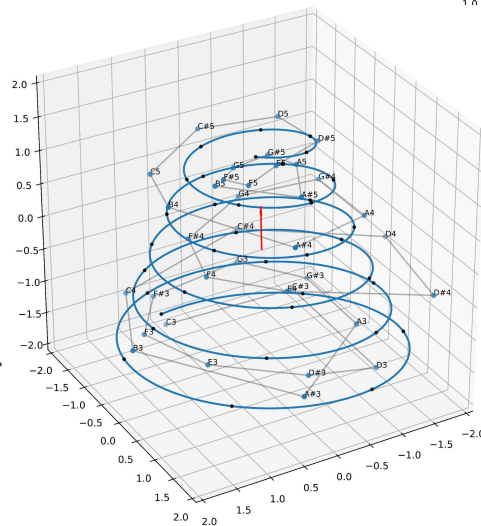
PCA



- 螺旋構造らしさの定量化



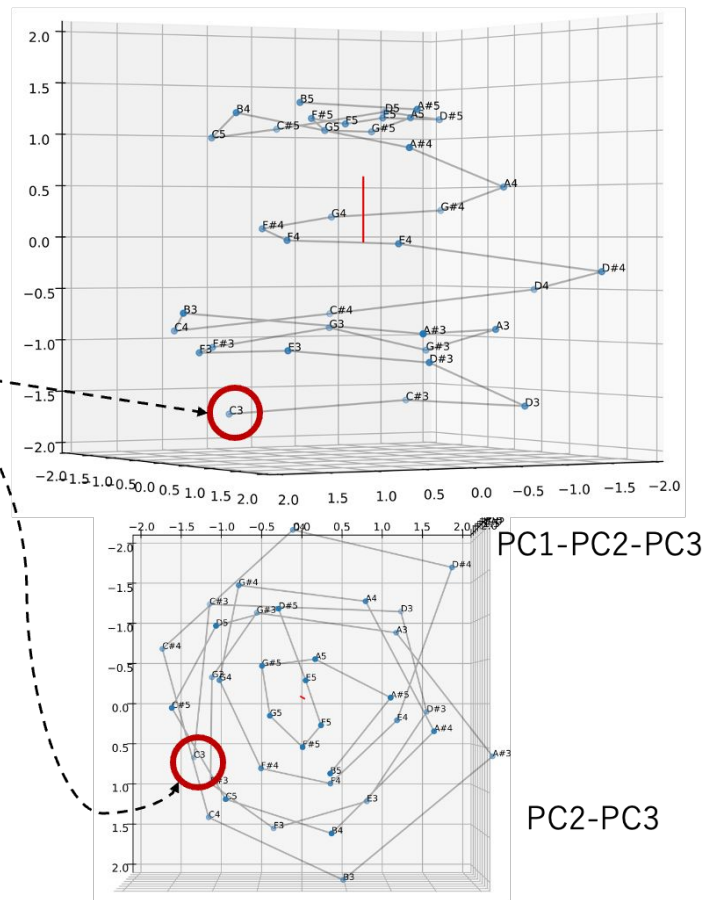
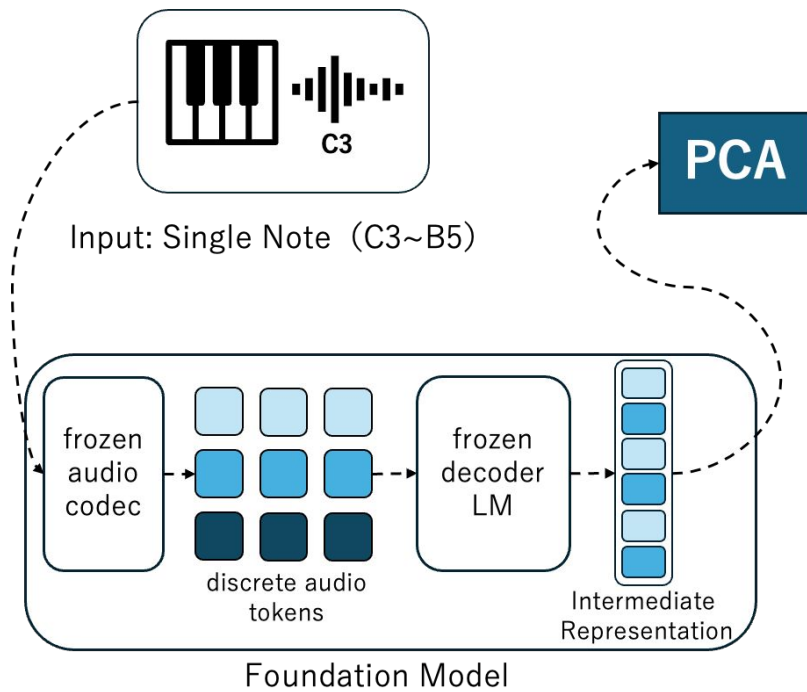
フィッティング



H = 2.60

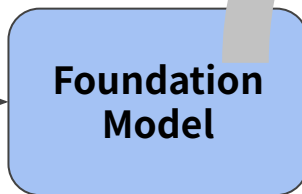
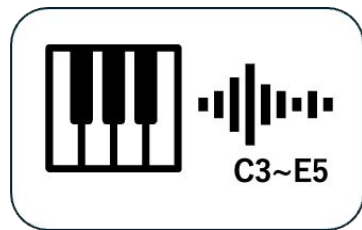
中間表現の獲得→次元削減 [Wei, 2024]

- 音高のみ異なる音楽信号を入力
- 音楽基盤モデルの**中間表現**を抽出
- **PCA**によって低次元化



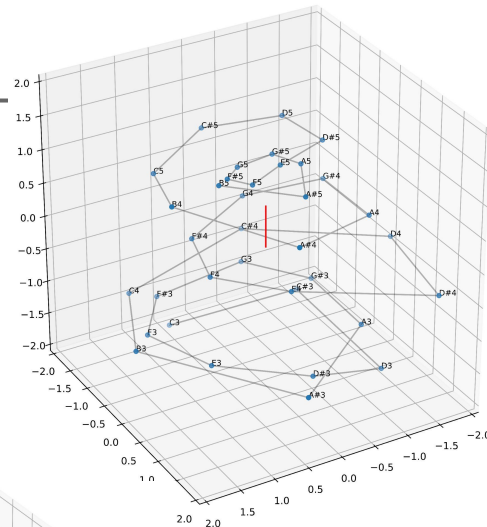
全体像

- 音高に関する中間表現の低次元化

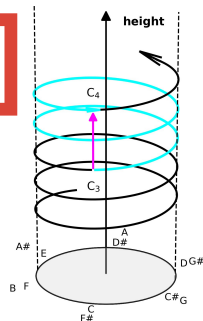
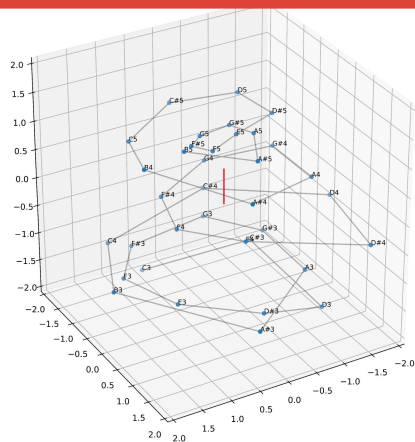


中間表現
抽出

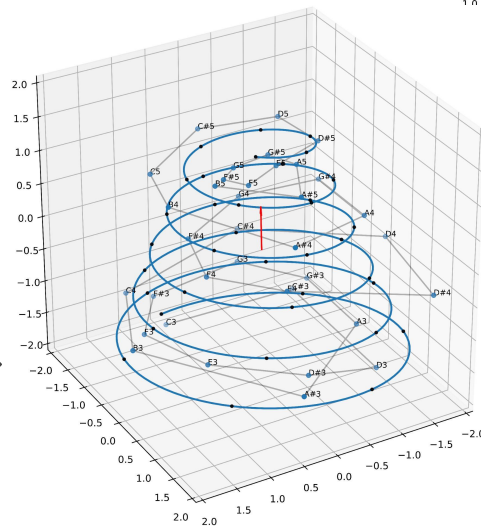
PCA



- 螺旋構造らしさの定量化

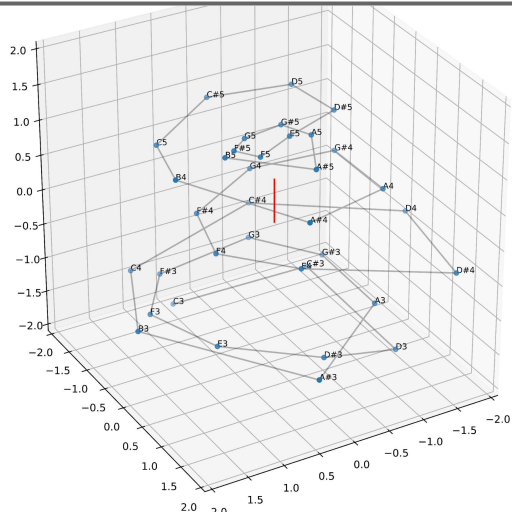


フィッティング

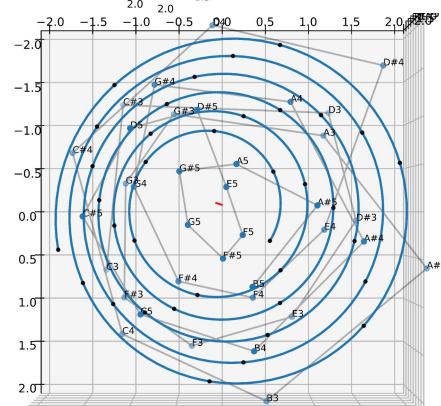
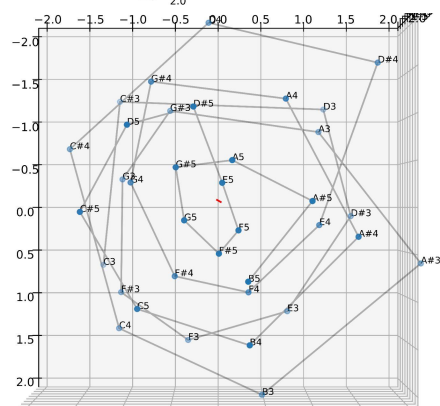
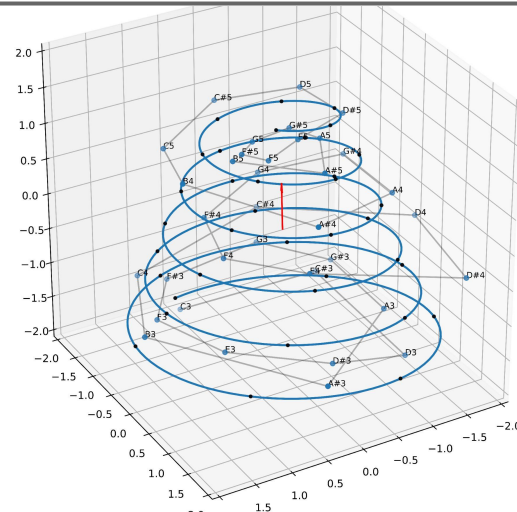


H = 2.60

螺旋フィッティング例



フィッティング



螺旋フィッティング関数

高さ方向

回転方向

$$\mathbf{y}(t) = \underbrace{h(t)}_{\text{高さ}} \cdot \underbrace{\mathbf{c}}_{\text{半径}} + r(t) \{ \underbrace{\cos \theta(t)}_{\text{位相}} \cdot \mathbf{u} + \sin \theta(t) \cdot \mathbf{v} \}$$

高さ

半径

位相

$\{\mathbf{c}, \mathbf{u}, \mathbf{v}\}$: 正規直交基底

t : 音高インデックス

$$h(t) = h_{\text{pitch}} \cdot t + h_0$$

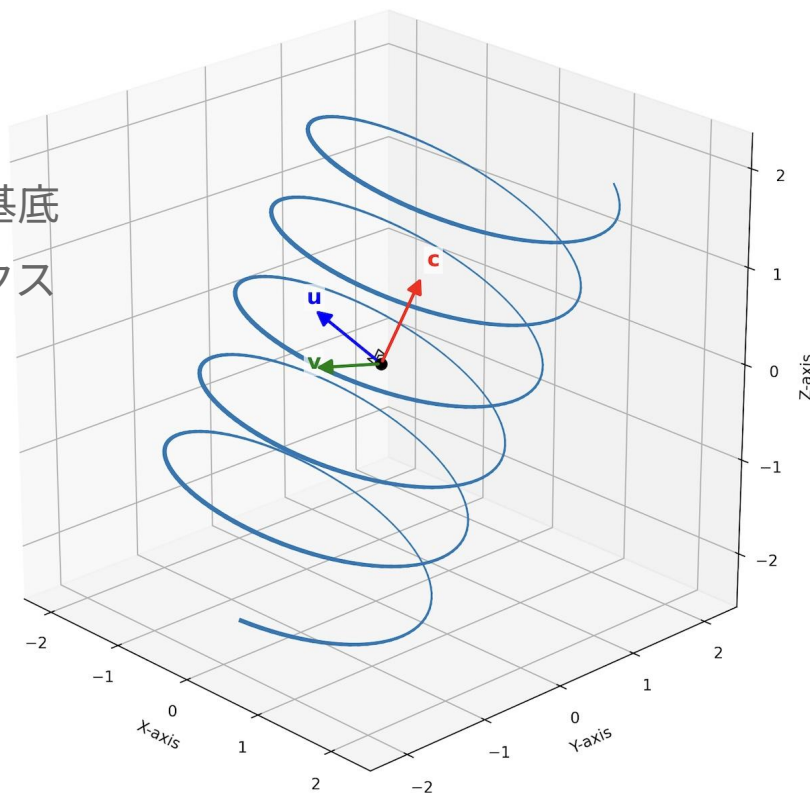
高さ変化係数 初期高さ

$$r(t) = r_{\text{slope}} \cdot t + r_0$$

半径変化係数 初期半径

$$\theta(t) = \omega_{\text{chroma}}(t - t_0)$$

角周波数 回転の初期位置



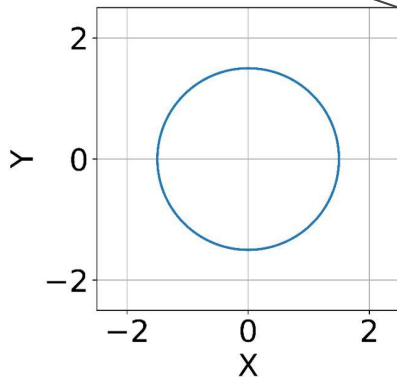
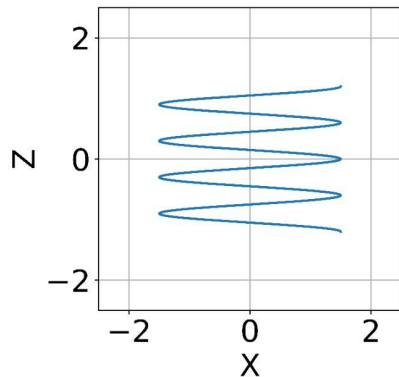
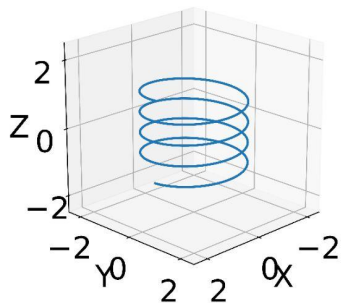
パラメータ比較

$$y(t) = h(t) \cdot c + r(t) \{ \cos \theta(t) \cdot u + \sin \theta(t) \cdot v \}$$

高さ

基準値
 $h_pitch = 0.15$,
 $\omega_chroma = \pi/3$,
 $c : z$ 軸,
 $r_0 = 1.5$, $r_slope = 0$
 $h_0 = 0$, $t_0 = 0$

$h_{pitch} = 0.1$



$$h(t) = h_{pitch} \cdot t + h_0$$

高さ変化係数 初期高さ

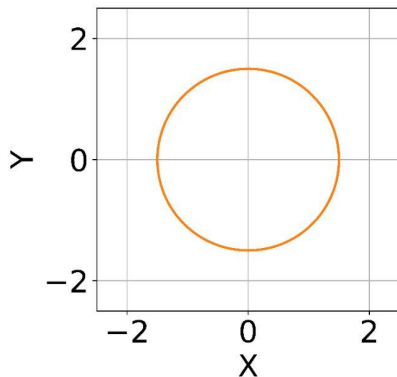
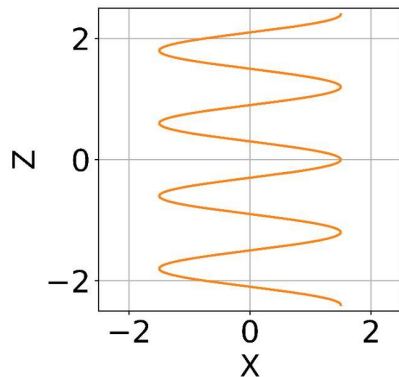
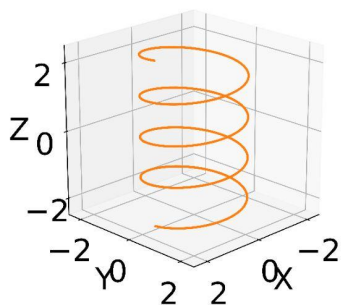
$$r(t) = r_{slope} \cdot t + r_0$$

半径変化係数 初期半径

$$\theta(t) = \omega_{chroma} (t - t_0)$$

角周波数 回転の初期位置

$h_{pitch} = 0.2$



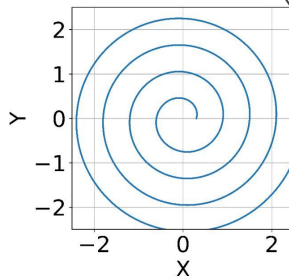
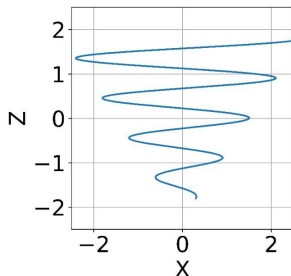
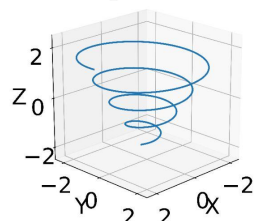
パラメータ比較

$$y(t) = h(t) \cdot c + r(t) \{ \cos \theta(t) \cdot u + \sin \theta(t) \cdot v \}$$

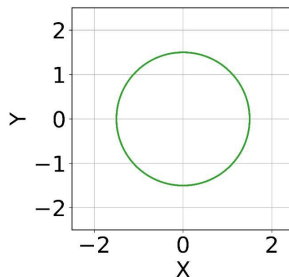
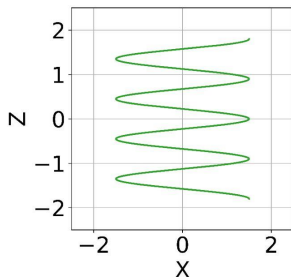
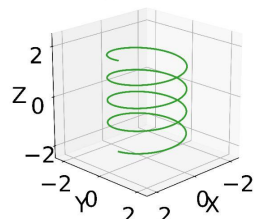
半径

基準値
 $h_pitch = 0.15$,
 $\omega_chroma = \pi/3$,
 $c : z$ 軸,
 $r0 = 1.5$, $r_slope = 0$
 $h0 = 0$, $t0 = 0$

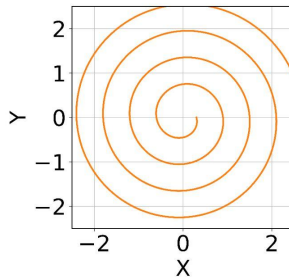
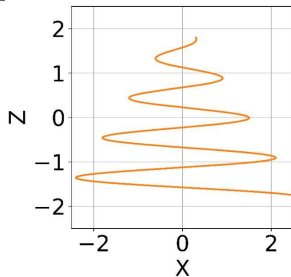
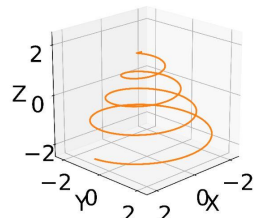
$r_{slope} = 0.1$



$r_{slope} = 0.0$



$r_{slope} = -0.1$



$$h(t) = h_{pitch} \cdot t + h_0$$

高さ変化係数 初期高さ

$$r(t) = r_{slope} \cdot t + r_0$$

半径変化係数 初期半径

$$\theta(t) = \omega_{chroma}(t - t_0)$$

角周波数 回転の初期位置

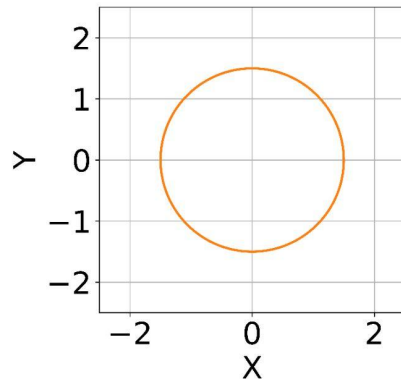
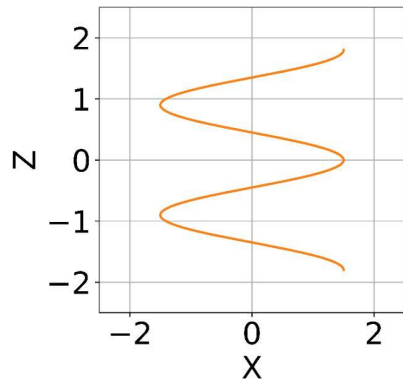
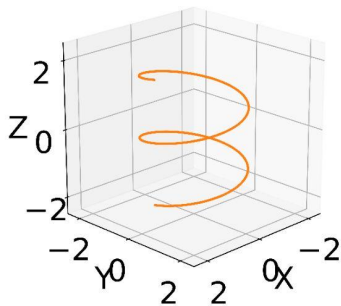
パラメータ比較

$$y(t) = h(t) \cdot c + r(t) \{ \cos \theta(t) \cdot u + \sin \theta(t) \cdot v \}$$

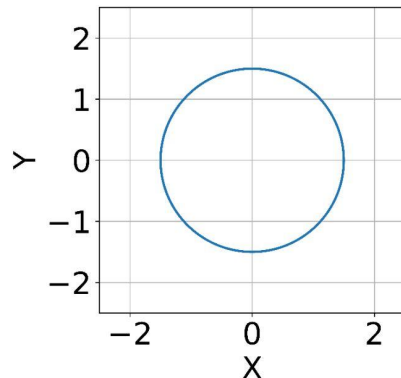
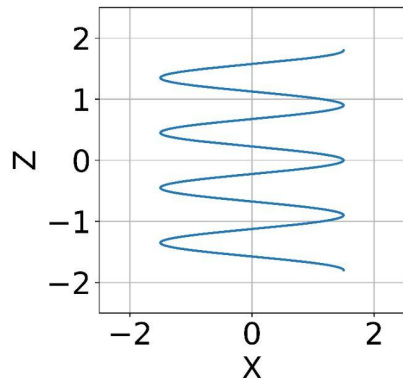
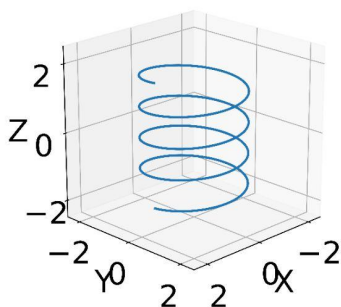
位相

基準値
 $h_pitch = 0.15$,
 $\omega_chroma = \pi/3$,
 $c : z$ 軸,
 $r_0 = 1.5$, $r_slope = 0$
 $h_0 = 0$, $t_0 = 0$

$$\omega_{chroma} = \pi/6$$



$$\omega_{chroma} = \pi/3$$



$$h(t) = h_{pitch} \cdot t + h_0$$

高さ変化係数 初期高さ

$$r(t) = r_{slope} \cdot t + r_0$$

半径変化係数 初期半径

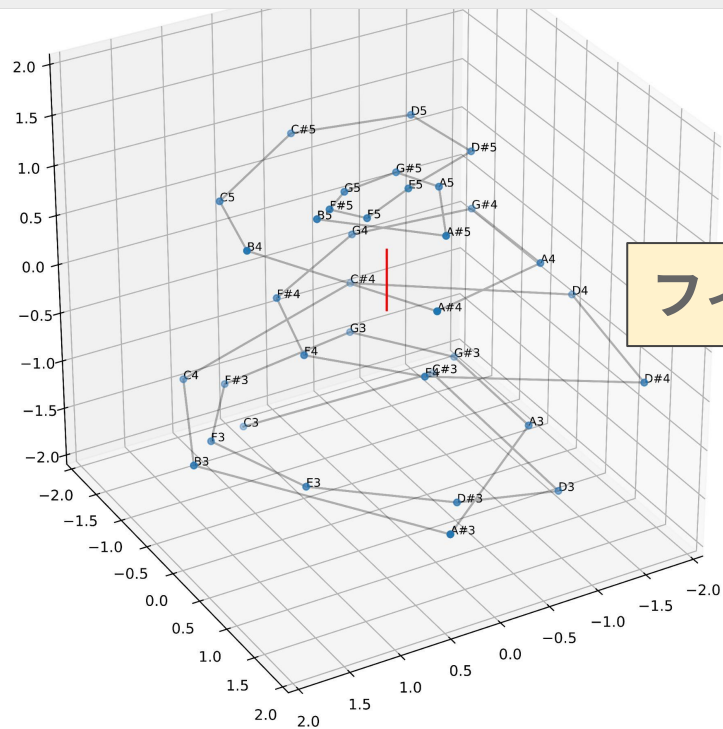
$$\theta(t) = \omega_{chroma}(t - t_0)$$

角周波数 回転の初期位置

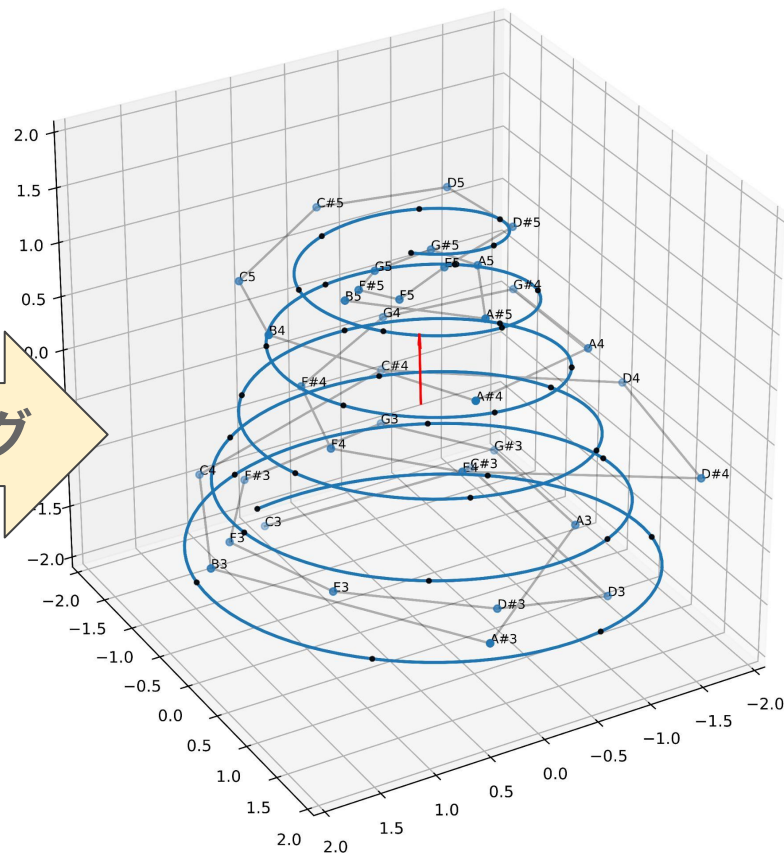
Helicityスコア

スコアが高い＝綺麗な螺旋

$$\text{Helicity} = \left(\frac{1}{N_{\text{key}}} \sum_{t=1}^{N_{\text{key}}} |x_t - y(t)|^2 \right)^{-1}$$



フィッティング

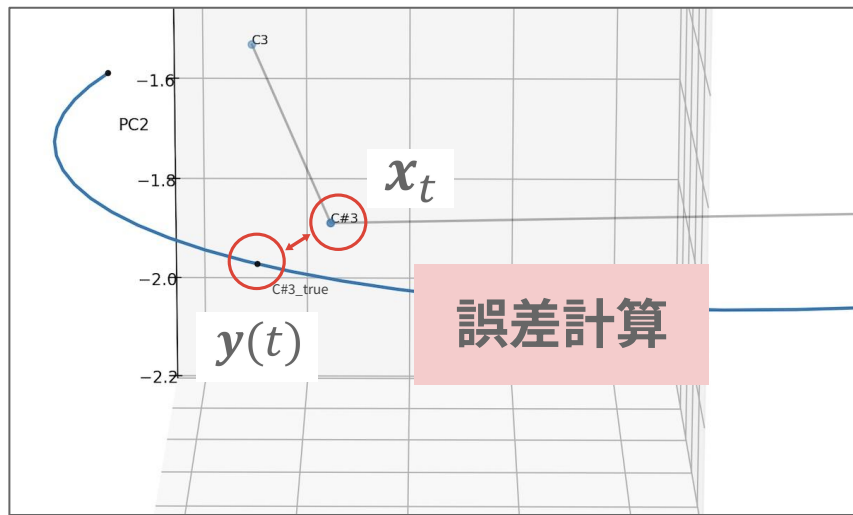


Helicityスコア

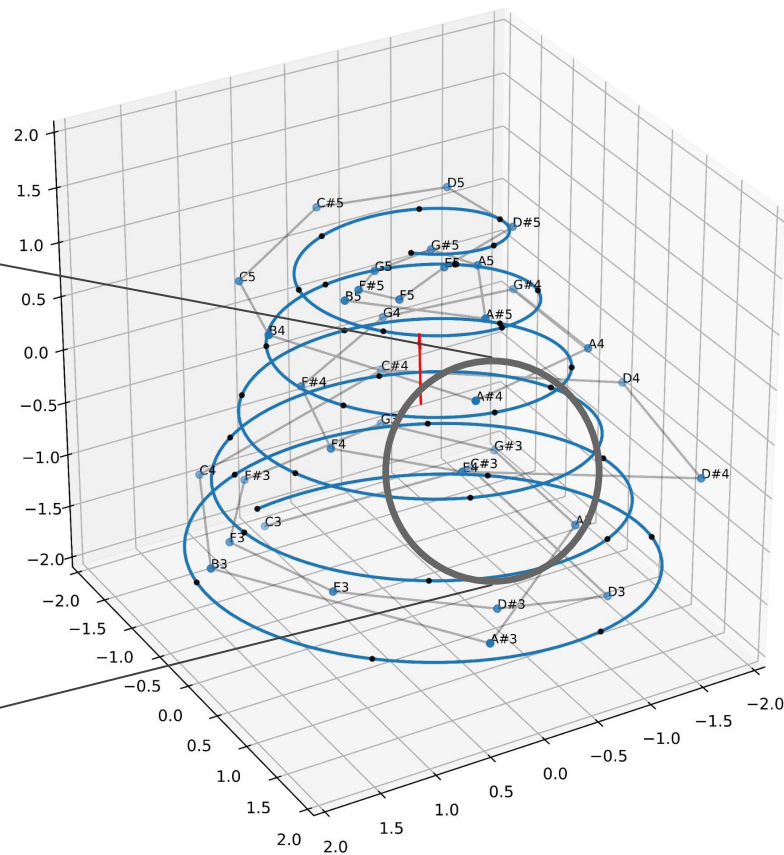
スコアが高い＝綺麗な螺旋

$$\text{Helicity} = \left(\frac{1}{N_{\text{key}}} \sum_{t=1}^{N_{\text{key}}} \underline{|\mathbf{x}_t - \mathbf{y}(t)|^2} \right)^{-1}$$

例) C#3の場合



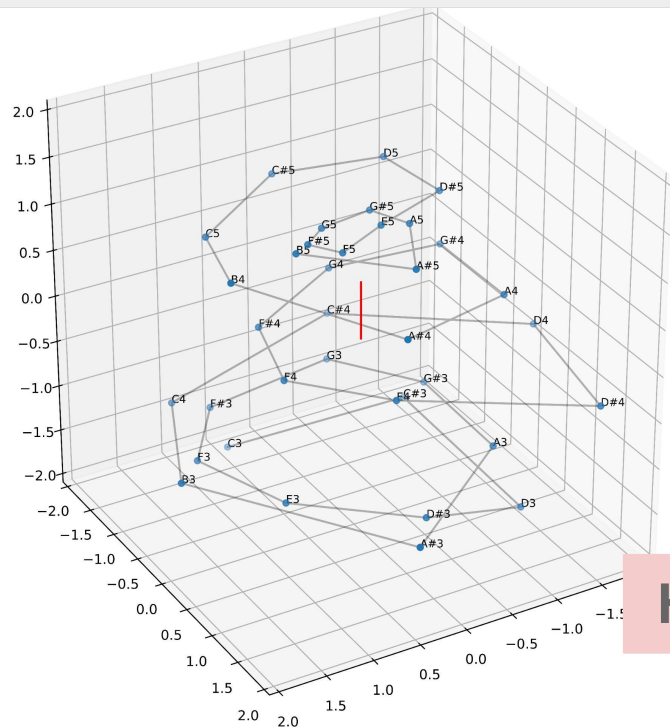
$$\mathbf{y}(t) = h(t) \cdot \mathbf{c} + r(t) \{ \cos \theta(t) \cdot \mathbf{u} + \sin \theta(t) \cdot \mathbf{v} \}$$



Helicityスコア

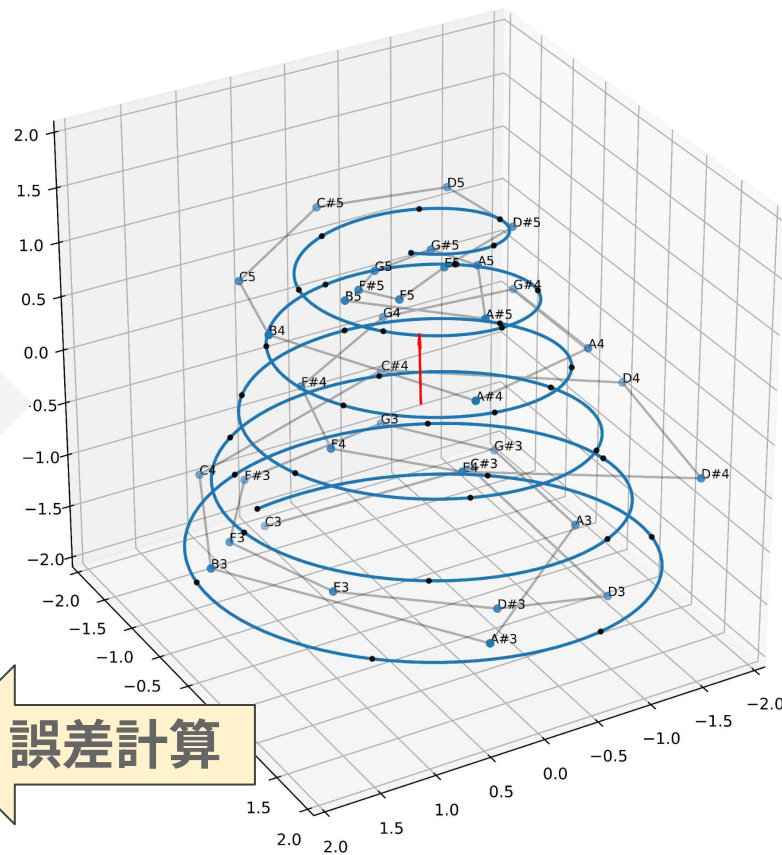
スコアが高い＝綺麗な螺旋

$$\text{Helicity} = \left(\frac{1}{N_{\text{key}}} \sum_{t=1}^{N_{\text{key}}} |x_t - y(t)|^2 \right)^{-1}$$



フィッティング

H = 2.60



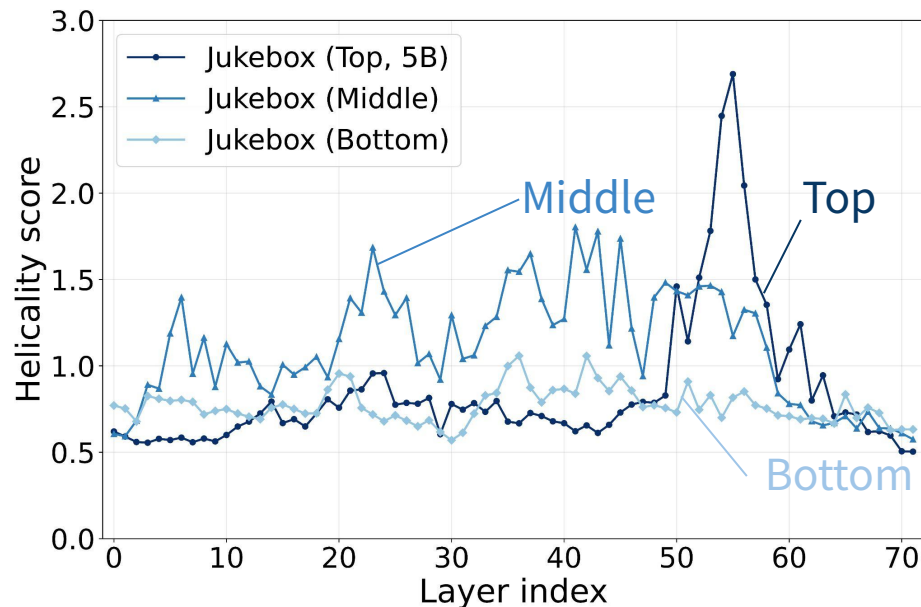
誤差計算

実験の評価

実験条件

モデル	Jukebox	Top(5B)	72層, 4800次元, 1トークン = 2.9ms
		Top(1B)	72層, 4800次元, 1トークン = 2.9ms
		Middle(1B)	72層, 1920次元, 1トークン = 0.73ms
		Bottom(1B)	72層, 1920次元, 1トークン = 0.18ms
	MusicGen	Small(300M)	24層, 1024次元, 1トークン = 20ms
		Medium(1.5B)	48層, 1536次元, 1トークン = 20ms
		Large(3.3B)	48層, 2048次元, 1トークン = 20ms
データ	SynTheory [Wei, 2024]	Notes	C3~B5 までの3 オクターブ, ピアノの合成音源
フィッティング	Optuna [Akiba, 2019]	試行回数1000 回, 繰り返し3回 → 最も良いHelicalityスコアを採用	

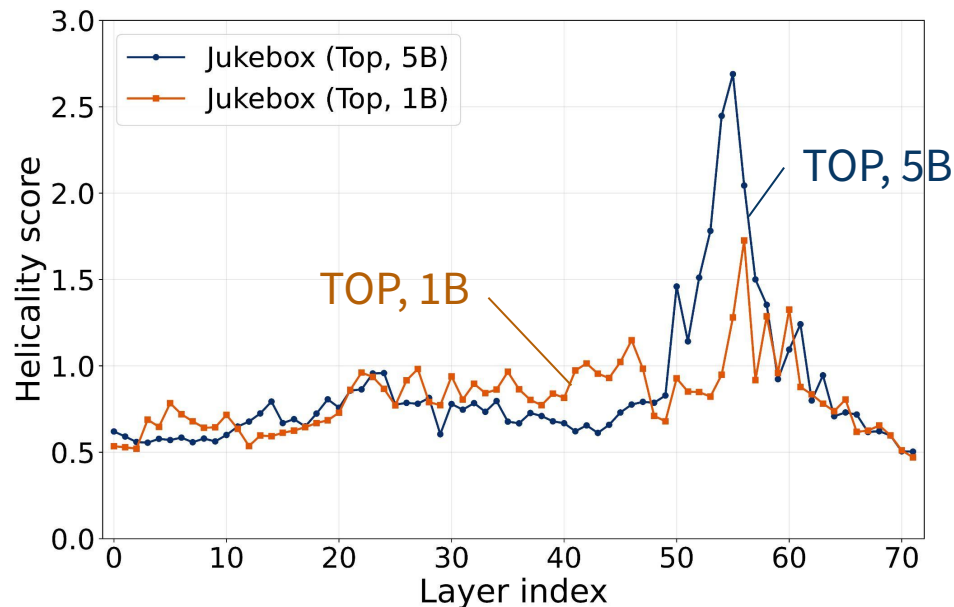
Jukeboxレベル別の比較



Top: 1トークン = 2.9ms
Middle: 1トークン = 0.73ms
Bottom: 1トークン = 0.18ms

- 最大値は**Top**が持つ
 - **Max:** 2.6891 (Layer 55)
 - **Mean:** 0.8510
 - 平均値は**Middle**が最も高い
 - **Max:** 1.7138 (Layer 42)
 - **Mean:** 1.0378
 - **Bottom**は全体的にスコアが低い
 - **Max:** 1.0584 (Layer 36)
 - **Mean:** 0.7714
- Topは**受容野が広く**，信号を粗く圧縮
→ 音高の**階層構造**を獲得
 - Bottomは**受容野が狭く**，音楽波形の再現を担う
→ 抽象的な**階層構造**を獲得しない

モデルサイズ別の比較 (Jukebox)



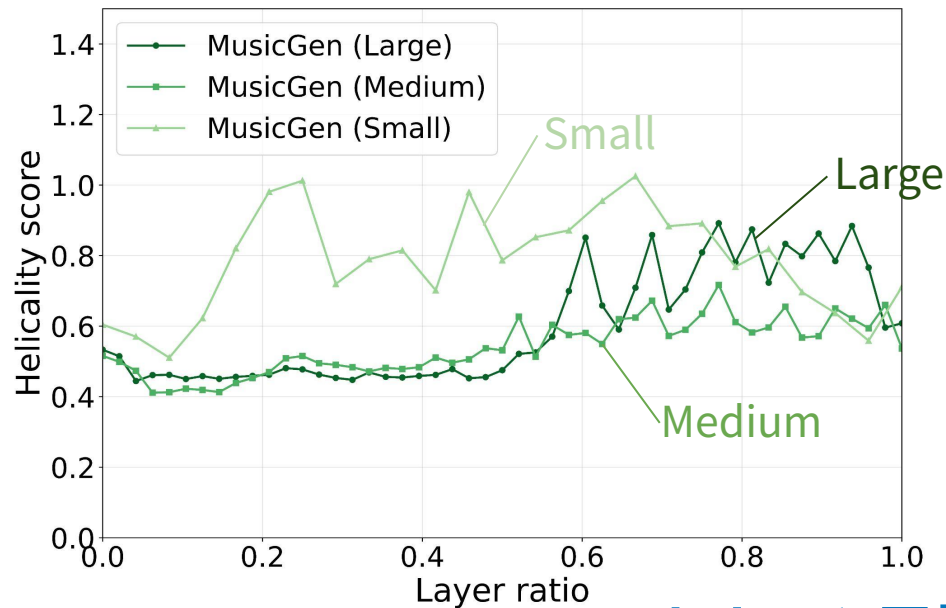
- 5Bモデル
 - 層の後半ほどスコアが上昇
 - 第55層付近で非常に高いスコア
- 1Bモデル
 - 分布は5Bと似ているが、ピークのスコアは控えめ

5B: Mean: 0.8510

1B: Mean: 0.8129

- モデルサイズが大きいほど、音高幾何に特化する層が生まれる可能性
- Top層の後半で音高情報を埋め込んでいる

モデルサイズ別の比較 (MusicGen)



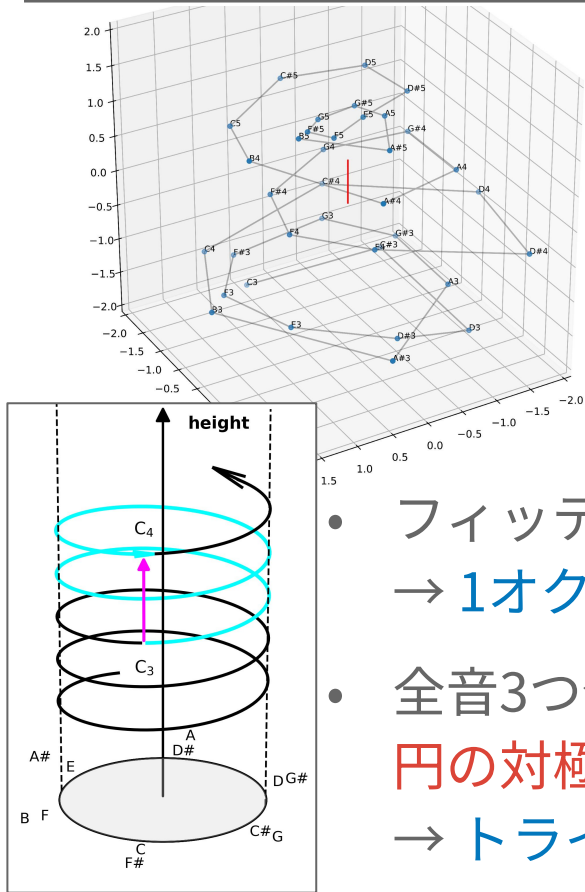
Large: 49層, 300M
Medium: 49層, 1.5B
Small: 25層, 3.3B

- Smallが**全体的に高スコア**を示す
- L/Mは**全体的にスコアが低い**
- モデルに関わらず**中～後半層でピークが出現**する傾向
- モデルサイズが大きくなるほど、**ピークが明確に出現**
- **Jukeboxと同様の傾向**を持つ
 - 「モデルサイズが大きくなるほど、音高幾何に特化した層が生まれる」という仮説を支持
- Jukeboxとは対照的に、**Small**が最も高いピーク

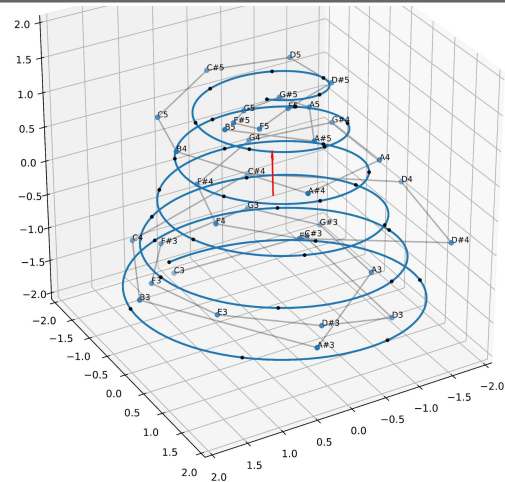
パラメータ分析：トライトーン

$$\theta(t) = \omega_{\text{chroma}}(t - t_0)$$

角周波数 回転の初期位置



フィッティング



- フィッティングパラメータ: $\omega_{\text{chroma}} = 1.008 \div \pi/3$ (1.047)
→ **1オクターブで2回転**する螺旋のフィッティング
- 全音3つ分の関係にある音（トライトーン）は螺旋上で
円の対極に位置するのではなく、**円の同じ位置**となる
→ **トライトーン代理**などの役割を捉えている？

中心軸cとpitch heightの関係

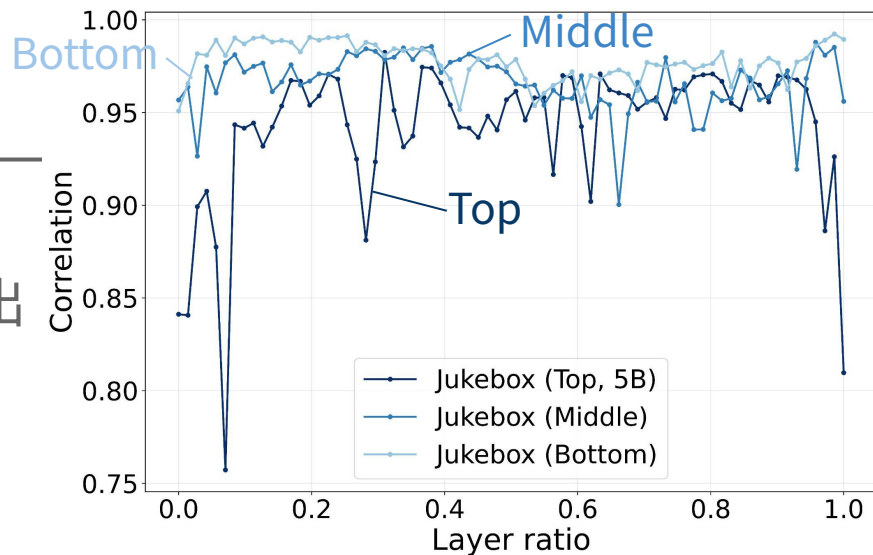
- 中心軸c へ射影した値と実際の音高の順序との間の相関係数を算出
- Jukeboxは全ての階層において高く安定した相関を示す
- Bottom階層ほど相関が高い
→ 周波数を捉える能力は、
受容野に依存する可能性

受容野

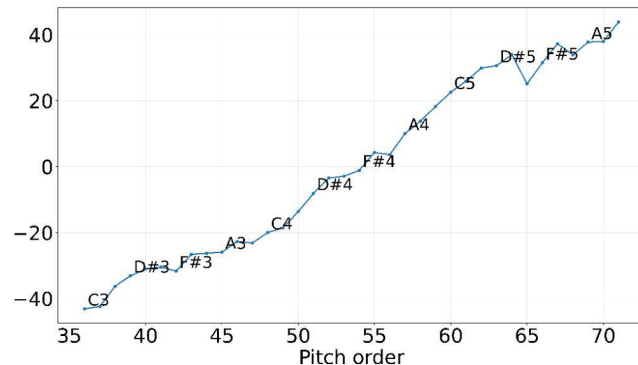
Top: 1トークン = 2.9ms
Middle: 1トークン = 0.73ms
Bottom: 1トークン = 0.18ms

各層の相関係数の平均値

Top: Mean: 0.942
Middle: Mean: 0.965
Bottom: Mean: 0.977



Jukebox bottomの第72層目
中心軸c への射影と音高の関係図 (r = 0.992)



中心軸cとpitch heightの関係

- MusicGenはモデルサイズによって振る舞いが異なる

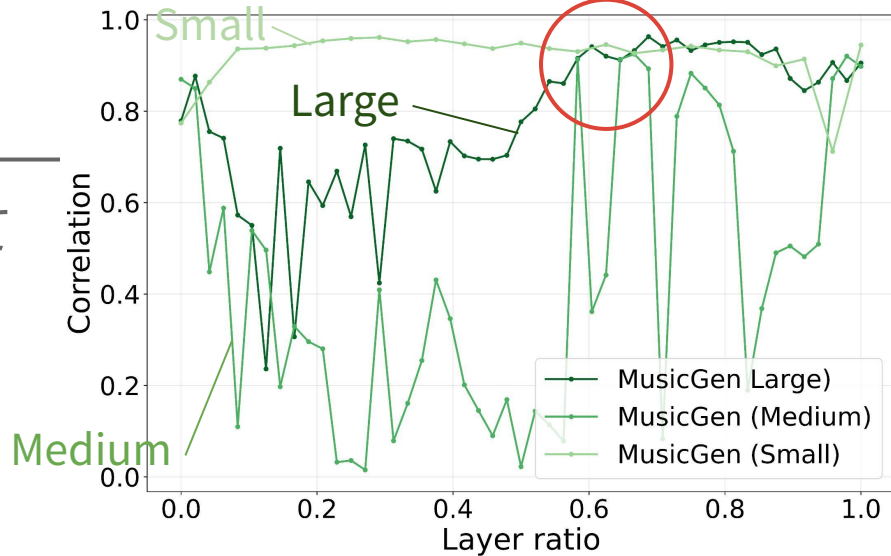
- **Small**はJukeboxに匹敵するほど高く安定した相関を示す

→ 限られたモデル容量の中で、

音楽の最も基本構造であるpitch heightを優先的に学習した可能性

- Large/Mediumは平均相関が大幅に低い (L: 0.762, M: 0.440)
- 相関の最大値はSmallに匹敵する (S: 1.000, M: 0.996, L: 1.000)

→ 「モデルサイズが大きくなるほど、音高幾何に特化した層が生まれる」
ということに対応する



まとめ

- 目的
 - 音楽基盤モデルが音高情報を螺旋構造に埋め込むかを調査する。
- 提案手法
 - 音高のみ異なる音楽データを入力とした時の中間表現をPCAで低次元化。
 - 螺旋モデルをフィッティングし，螺旋構造らしさを定量評価。
- 評価結果
 - 音高のような抽象的な概念はモデルの**深い層**でより明確に表現される。
 - モデルサイズが大きくなるほど，音高幾何に特化した層が生まれる。
- 結論
 - 音楽基盤モデルは**人間の音高知覚と類似した螺旋構造を内部に獲得し**，モデルの**階層構造**や**サイズ**が螺旋表現の明瞭さと局所性に影響する。

補足

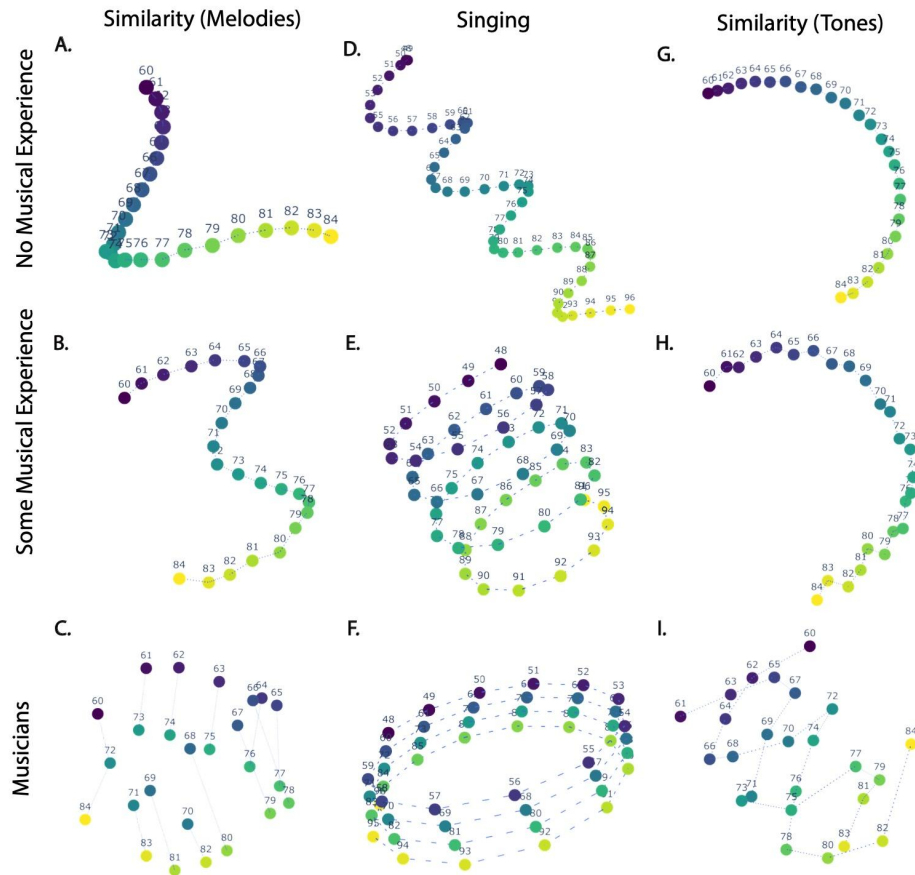
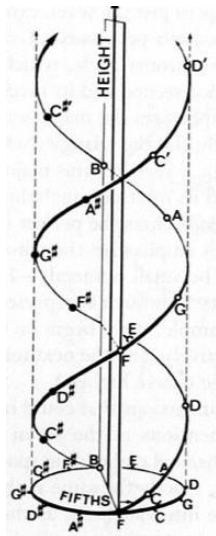
各モデルの最高スコア層におけるパラメーター一覧

モデル	層	H	r0	h_pitch	r_slope	ω_{chroma}	h0	t0
Jukebox(Top)	55/72	2.69	1.77	0.095	-0.025	1.008	-1.588	0.238
Jukebox(Middle)	41/72	1.80	2.03	0.094	-0.050	-1.023	-1.545	-0.103
Jukebox(Bottom)	36/72	1.06	1.15	0.102	-0.006	1.058	-1.866	-4.313
MusicGen(Large)	37/48	0.89	1.40	0.088	-0.039	-0.567	-1.557	5.430
MusicGen(Medium)	37/48	0.72	1.42	0.061	-0.036	0.525	-1.102	-0.367
MusicGen(Small)	16/24	1.03	1.53	0.088	-0.026	-1.041	-1.482	3.998

人間の音高知覚の研究

[Raja, 2023] [Probing the Structure of Musical Pitch Across Tasks and Experience](#)

- 音楽経験によって現れる構造が異なる
- 熟練者ほど音楽的性質を反映した形
- 音楽家の螺旋構造の高さ方向がほぼない
→ オクターブ識別力が高い

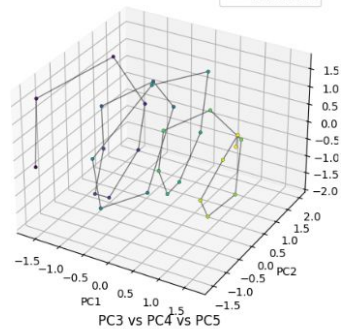


例1

Acoustic_Grand_Piano, octave[3, 4, 5], Layer 55

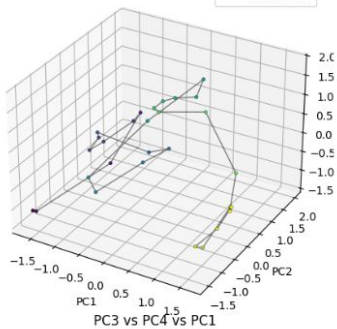
PC1 vs PC2 vs PC3

Pitch Order



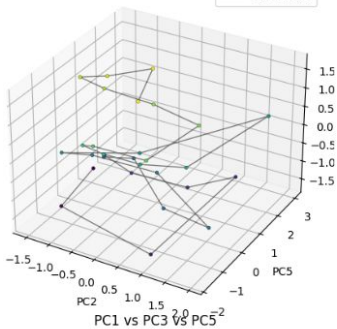
PC1 vs PC2 vs PC4

Pitch Order



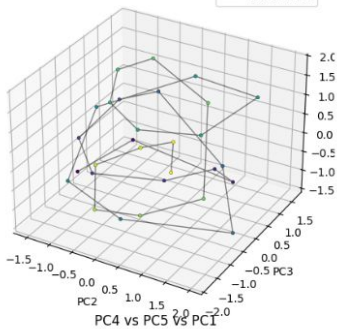
PC2 vs PC5 vs PC1

Pitch Order



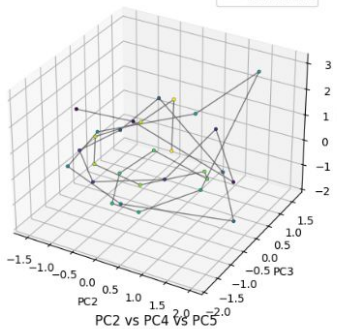
PC2 vs PC3 vs PC4

Pitch Order



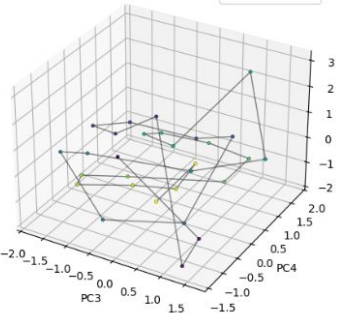
PC2 vs PC3 vs PC5

Pitch Order



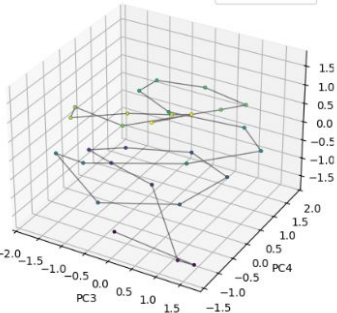
PC3 vs PC4 vs PC5

Pitch Order



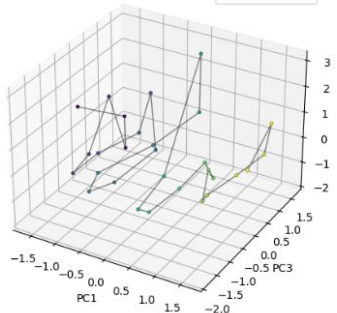
PC3 vs PC4 vs PC1

Pitch Order



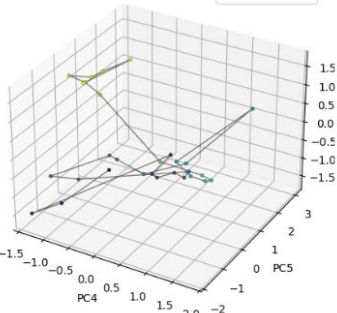
PC1 vs PC3 vs PC5

Pitch Order



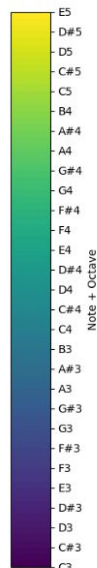
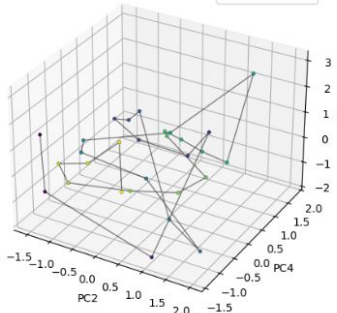
PC4 vs PC5 vs PC1

Pitch Order



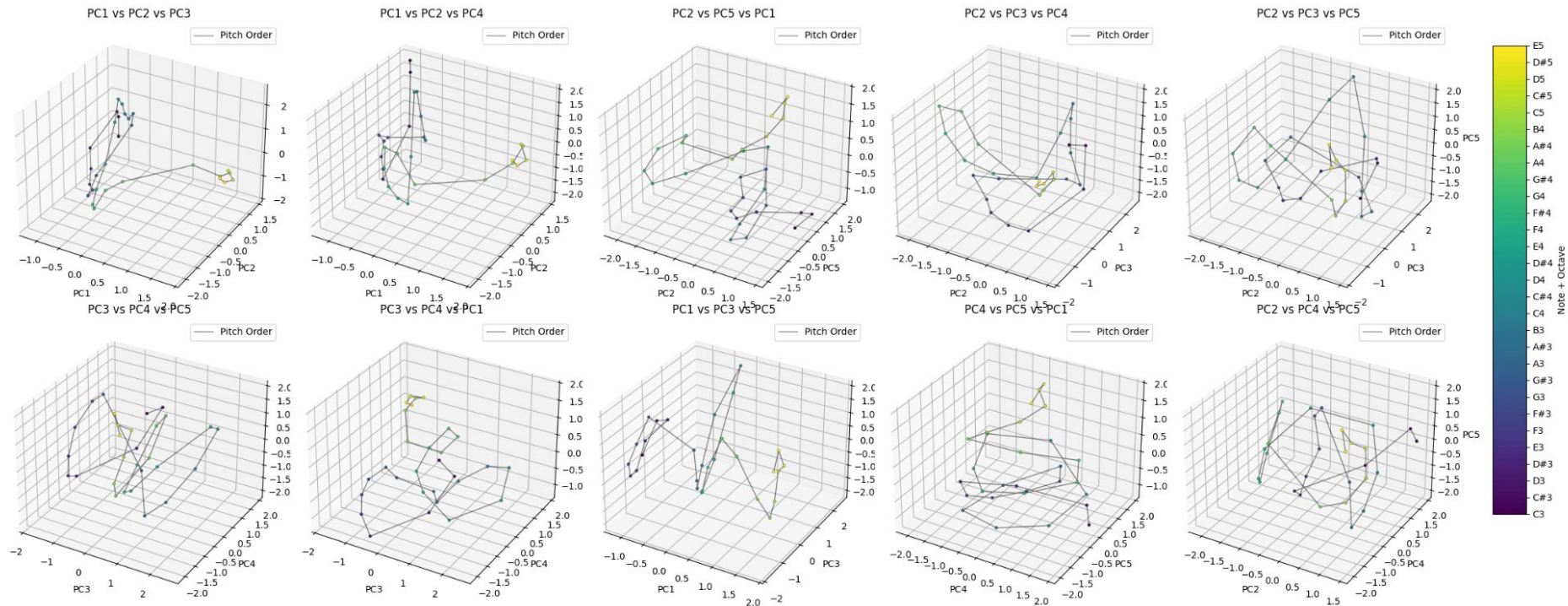
PC2 vs PC4 vs PC5

Pitch Order



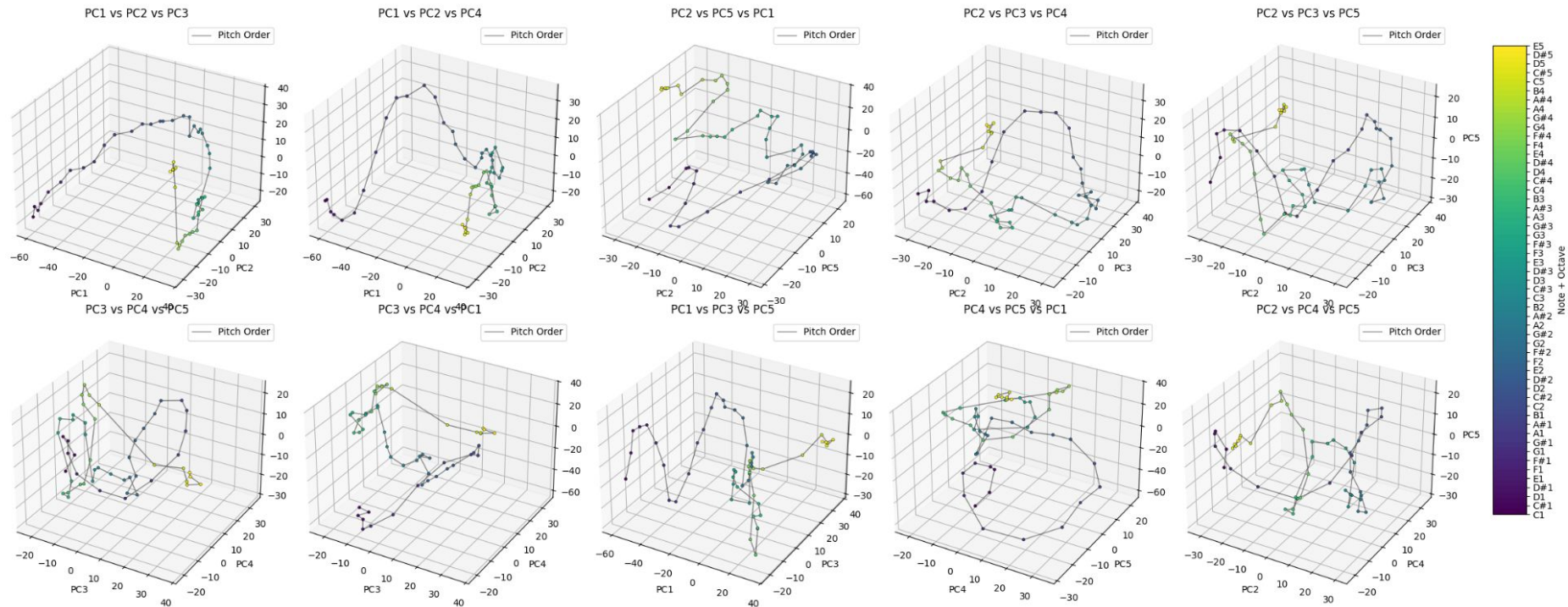
例2

Acoustic_Grand_Piano, octave[3, 4, 5], Layer 48



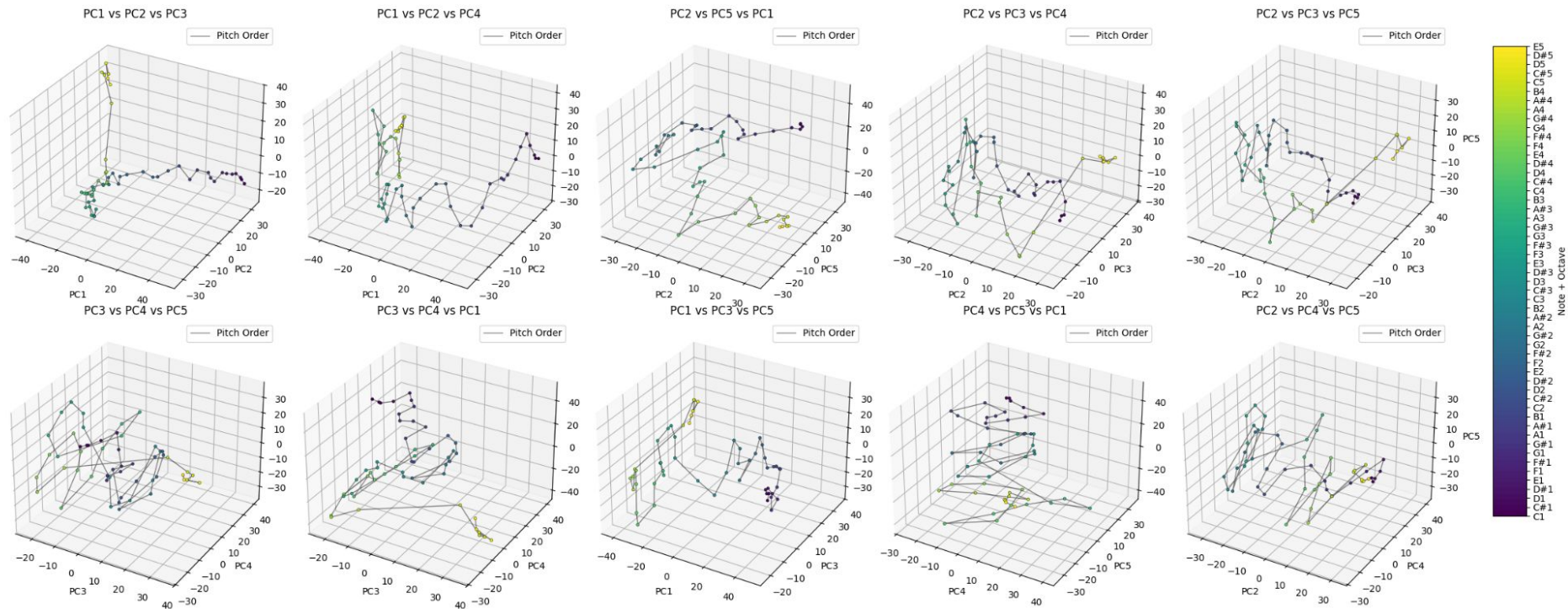
例3

Acoustic_Grand_Piano, octave[1, 2, 3, 4, 5], Layer 45

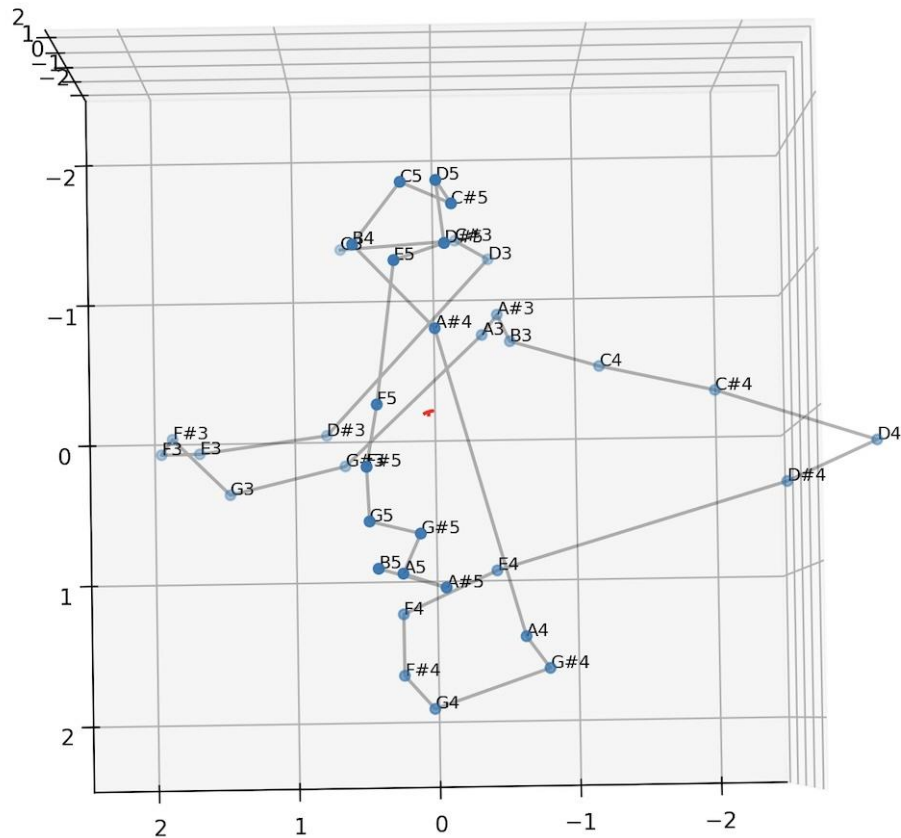
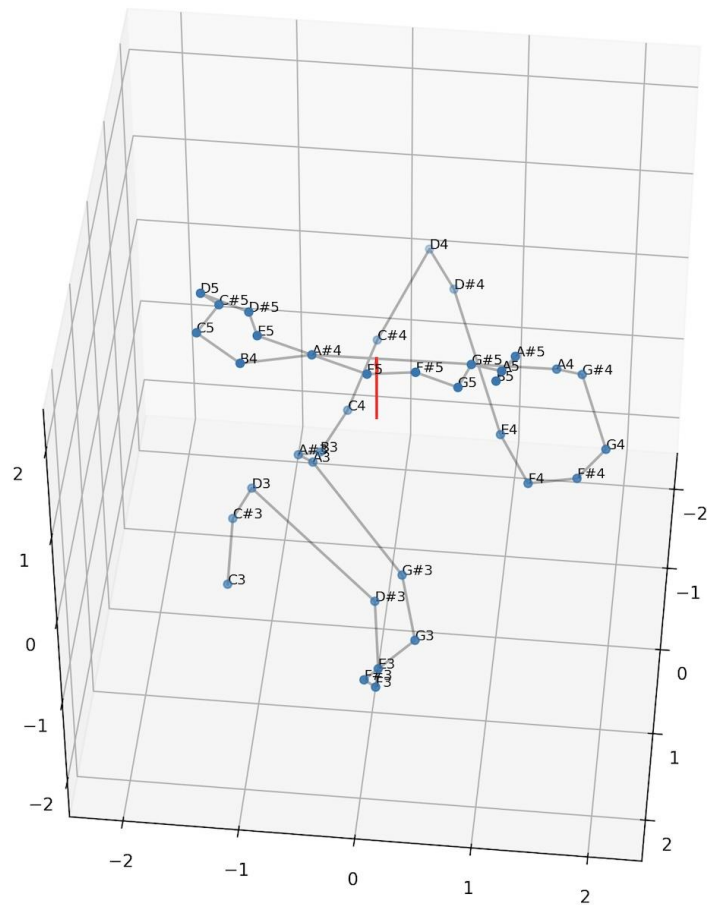


例4

Acoustic_Grand_Piano, octave[1, 2, 3, 4, 5], Layer 36

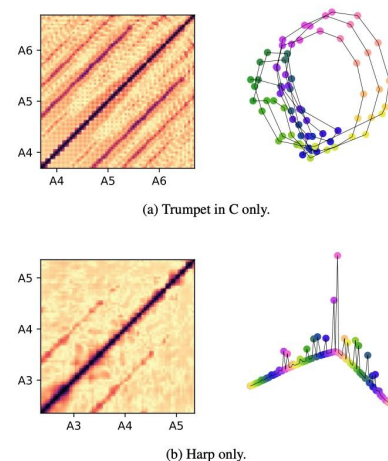
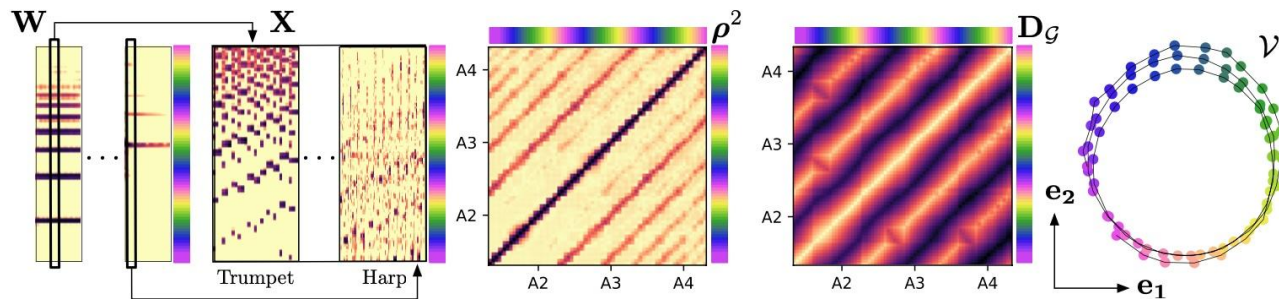


例3 Jukebox top 47層 pc1-pc2-pc3



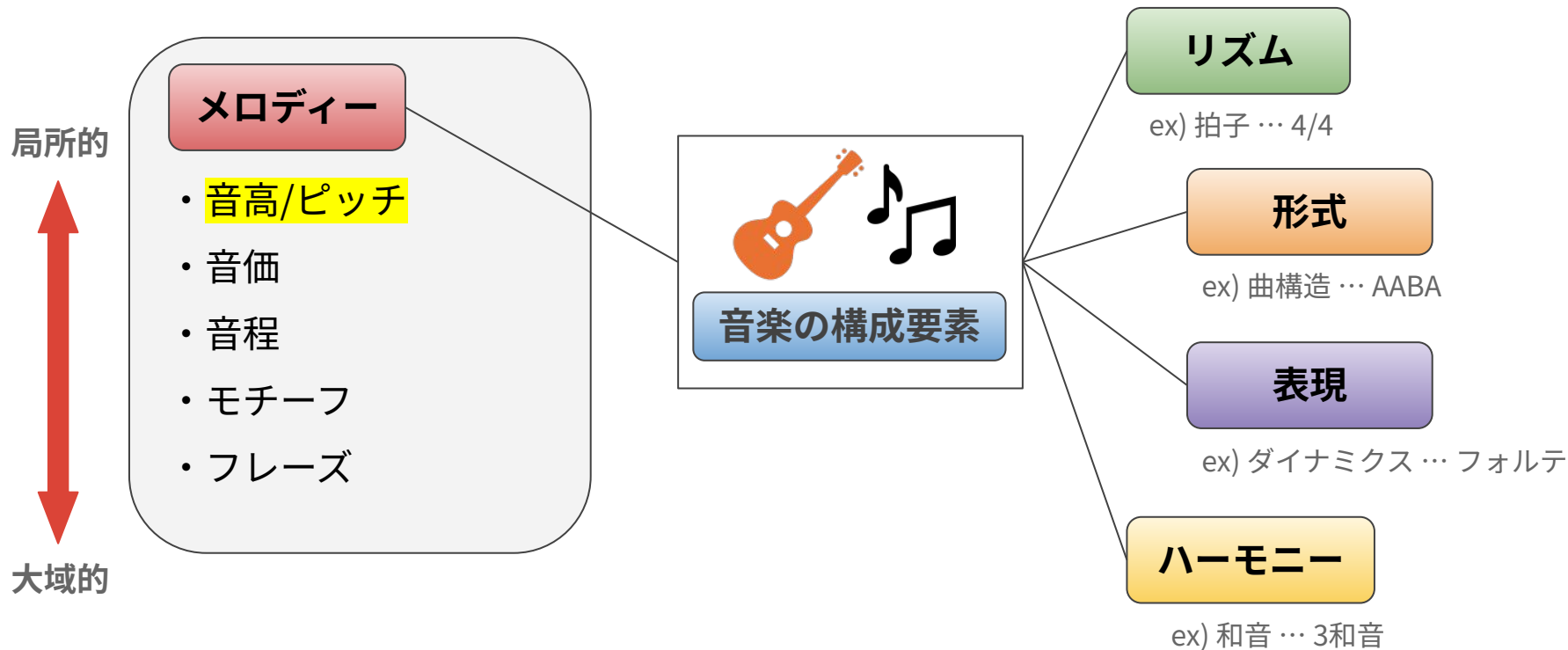
音高の螺旋表現（オクターブ等価性）

音響信号から音高螺旋構造を可視化 [Vincent, 2020]



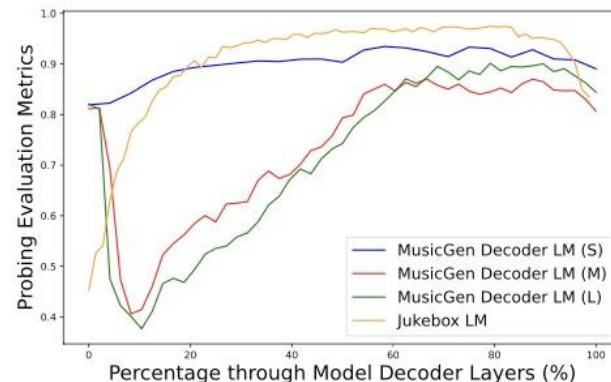
- 音響信号から音高螺旋構造を自動発見する手法
- 様々な楽器の単音データにこの手法を適応した結果，綺麗な螺旋を形成
- 音色によって構造が変わる
 - トランペットのような倍音が豊かな楽器は，綺麗な螺旋が形成される
 - ハープのような倍音が弱い楽器では，隣接音との相関が強く，直線的な構造

音楽の構成要素



音楽基盤モデルの解析

- 多くの音楽理論タスクで高性能
- モデルが音楽の**基本概念を中間表現にエンコードしている**とわかる



	Notes	Intervals	Scales	Chords	Chord Progressions	Tempos	Time Signatures	Average
Jukebox LM	0.951	0.995	0.978	0.997	0.971	0.993	1.000	0.984
MusicGen LM (S)	0.897	0.995	0.949	0.990	0.942	0.969	0.911	0.950
MusicGen LM (M)	0.851	0.983	0.863	0.989	0.870	0.956	0.883	0.914
MusicGen LM (L)	0.866	0.972	0.905	0.989	0.901	0.965	0.905	0.929
MusicGen Audio Codec	0.729	0.965	0.383	0.879	0.330	0.947	0.677	0.701
Mel Spectrogram	0.712	0.995	0.897	0.988	0.723	0.785	0.827	0.847
MFCC	0.467	0.822	0.370	0.863	0.872	0.923	0.688	0.715
Chroma	0.954	0.820	0.989	0.994	0.869	0.847	0.672	0.878
Aggregate Handcrafted	0.941	0.997	0.972	0.992	0.868	0.947	0.833	0.936