

研究報告

AIによる歩行動作の潜在表現解析と可聴化および可視化 Latent Representation Analysis and Audio-Visual Rendering of Gait Motion Using AI

久世 空汰, 水谷 旭陽, 十河 悠真
Sorata KUZE, Asahi MIZUTANI, Yuma SOGO
滋賀大学
Shiga University

上田 尚史, 比嘉 怜菜, 濱野 峻行
Naofumi UEDA, Reina HIGA, Takayuki HAMANO
国立音楽大学
Kunitachi College of Music

概要

本研究は、歩行動作に内在する非言語的情報構造を深層学習により抽出し、その潜在表現を可聴化・可視化する枠組みを構築することを目的とする。MediaPipe Poseにより取得した33点骨格データ(99次元)を入力として、LSTM-AEを用いて8次元の潜在空間へ圧縮した。得られた潜在変数に対して相関解析を行い、一部の潜在次元が上半身・下半身など身体構造と対応する特徴を示すことを確認した。さらに、潜在変数の時間変化を音響パラメータへマッピングし、Cycling' 74 Maxによりリアルタイムでの可聴化を実装した。またレーダーチャート状および折れ線グラフ状の二種類の可視化手法を設計し、潜在構造の理解を支援した。これらにより「歩行動作→潜在表現→可聴化・可視化」という一連のパイプラインを実装し、身体動作を音・映像表現へ変換する枠組みの可能性を示した。

This paper aims to construct a framework that extracts non-verbal information structures in human gait by deep learning and renders its latent representation as both visual and auditory expressions. We use 33-point skeletal data (99 dimensions) obtained from MediaPipe Pose as input and compress it into an 8-dimensional latent space using an LSTM autoencoder. Correlation analysis between latent variables and input features shows that some latent dimensions correspond to specific body structures such as upper and lower limbs. We further map the temporal trajectories of latent variables to sound parameters in Cycling' 74 Max to implement sonification of gait. In addition, we design two visualization methods—a radar-chart-

like radial plot and a line-graph-like time series plot—to support the interpretation of latent structures. Through these components, we realize a pipeline from “gait motion” through “latent representation” to “audio-visual presentation,” suggesting a new framework for transforming bodily movement into sound and visual expression.

1. はじめに

人間の身体動作の中でも、本研究が対象とする歩行は、日常的に繰り返される動作でありながら、個人ごとの癖やリズム、姿勢など「その人らしさ」が現れやすい。私たちは、他者の歩き方から「楽しそう」「悲しそう」といった感情や、「かっこいい」「不格好」といった印象を無意識のうちに読み取っている。このことは、歩行が単なる移動手段にとどまらず、その人固有の非言語的情報を豊かに内包していることを示唆している。そこで本研究では、歩行動作の潜在的な特徴を抽出し、可聴化および可視化を通じてその差異を明らかにすることを旨とする。特に可聴化は、時間変化やリズムの違いといった動的な特徴を直感的に把握しやすく、微妙な変動を知覚しやすいという利点がある。音による提示は視覚情報へのアクセスが難しい場合にも有効であり、アクセシビリティの観点からも有益である。

本研究では、歩行動作データから潜在変数を抽出し、その非言語的情報構造を可聴化・可視化する枠組みを提案する。

- 歩行動作データを LSTM-AE により 8 次元の潜在空間に圧縮し、身体構造との対応関係を分析した。

- 潜在変数の時間変化を音響パラメータへマッピングし、歩行動作の潜在構造を可聴化するシステムを構築した。
- レーダーチャート状および折れ線グラフ状の可視化を設計し、潜在表現の構造理解を多面的に支援する枠組みを実現した。

さらに、本研究は滋賀大学と国立音楽大学との連携協定事業の一環として、両者の学生・教員が協働する共創的アートプロジェクトとして位置づけられる。2025年11月2日には国立音楽大学において作品展示および研究発表を行い、来場者からの反応を通して提案枠組みの評価も行った。

2. 関連研究

2.1. 動作解析と深層学習

姿勢推定に基づく動作解析は MediaPipe や OpenPose を用いた研究が盛んであり、AE や LSTM による潜在表現獲得も多く行われている。近年では、MediaPipe によるマーカーなし歩行解析も精度検証が進んでおり、Hii ら (Hii et al. 2023) は MediaPipe Pose により抽出した骨格点から歩行指標を自動推定し、Vicon 計測との比較によりその妥当性を報告している。しかし、歩行動作の抽象特徴を芸術的・音響的表現に接続する研究は少ない。

2.2. 身体動作と音楽表現の接続

ダンス動作の音響化やモーションキャプチャデータの音楽生成など、身体表現と音響表現を接続する試みはこれまでも行われている。例えば Bevilacqua らは、モーションキャプチャにより取得したダンス動作から特徴量を抽出し、Max/MSP を用いてサウンドトラックを生成する環境を構築している (Bevilacqua, Naugle, and Valverde 2001)。また Giomi は、ダンスにおけるインタラクティブ音楽システムとムーブメント・ソニフィケーションの関係を整理し、身体感覚に根ざした「ソマティック・ソニフィケーション」の枠組みを提示している (Giomi 2020)。さらに Landry らは、ダンス動作と感情を同時に伝達するインタラクティブな音楽的ソニフィケーション手法を提案している (Landry and Jeon 2020)。最近では、ダンス動画からマルチトラック音楽を自動生成する深層学習モデルも提案されている (Han et al. 2024)。しかし、これらの多くはダンス表現を対象としており、歩行に焦点を当て、潜在空間ベースの表現と音響・視覚表現を統合的に扱う研究は少ない。

3. 提案手法

本節では、本研究で構築した「歩行動作→潜在表現→可聴化・可視化」の一連の処理手順について述べる。

3.1. データ収集

本研究では、歩行動作の潜在表現解析および可聴化・可視化を行うために、計 12 名の被験者から歩行映像データを収集した。被験者のうち 8 名には、「嬉しい」「緊張」「通常」「悲しい」の 4 種類の感情を意識して歩行してもらい、感情による動作差異の取得を試みた。残りの 4 名については、自然な通常歩行のみを撮影した。

撮影には一般的なカメラ付きスマートフォン (iOS および Android 搭載端末) を使用し、フレームレート 30 fps で記録した。撮影距離は、被験者の体全体がフレーム内に収まるよう配慮し、撮影者を基準に正面 8 m 程度の位置から被験者が手前 3 m ほどの地点まで歩行する動きを取得した。

当初は「嬉しい」「緊張」「通常」「悲しい」など感情を意識した歩行を 4 種類の条件で収集し、感情差による動作の違いを取得することを試みた。しかし、撮影環境下で被験者が意識的に感情を表出しようとする影響もあり、4 条件間で明確な差異は得られなかった。この結果、後半では「感情を歩き分けてもらう」方針を撤回し、各被験者の自然な通常歩行を中心に、その人固有の歩き方が潜在空間でどのように表現されるかを解釈する方向へと方針を変更した。歩行回数は 4 回または 1 回とし、撮影時間・負担とのバランスを踏まえた運用上の設定とした。取得した映像に対しては、MediaPipe Pose (mediapipe 0.10.21) を用いて 33 点の骨格ランドマークを抽出し、それぞれの x, y, z 座標を取得した。これにより 99 次元 (33 点 \times 3 座標) の時系列データが得られた (図 1)。

3.2. 潜在表現解析

歩行動作データから潜在的な運動特徴を抽出するために、オートエンコーダ (Autoencoder, AE) を用いた潜在表現解析を実施した。本節では、最終的に採用した LSTM-AE に至るまでのモデル改良過程を含めて述べる。

3.2.1. 初期モデル：2次元座標 (66次元) による AE の課題

当初は、MediaPipe Pose により得られた骨格 33 点の x, y, z 座標のうち、 x, y 座標 (66次元) のみを入力として AE による潜在表現の抽出を試みた。これは、モデルが歩行動作の一連の流れを処理できるかを簡便に検証するための初期設定であった。しかし、被験者

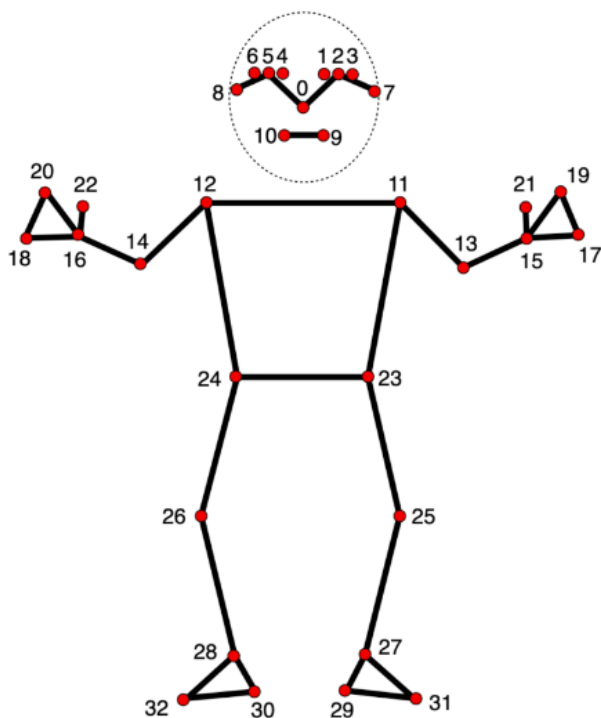


図 1: MediaPipe Pose より引用した骨格特徴点のイメージ

はカメラ方向へ前進しながら歩行しているため、以下の問題が生じた。

- x, y のみでは奥行方向 (z 軸) の情報が欠落する。
- 歩行に伴い x, y 座標が前進方向に引きずられるように変動し、不自然な増加が生じる。

この結果、AE は前進動作を適切に解釈できず、歩行動作の構造学習に困難をきたした。

3.2.2. 99次元入力 (x, y, z) への拡張

上記の問題を解決するため、骨格 33 点の x, y, z 座標 (99 次元) を入力とするモデルへ拡張した。MediaPipe Pose 自体は当初から x, y, z の 3 次元座標を出力しており、本研究ではそれらを全て用いることで、奥行方向を含む歩行動作の特徴を AE が適切に学習可能となった。

3.2.3. 静止フレーム学習の限界：時間構造を扱えない問題

この段階では AE の入力として各フレームを独立データとして扱っており、歩行動作に本質的な時間依存構造を捉えることができなかった。例えば、

- 足の交互運動、

- 腕振りの反対位相性、
- 体幹の上下動の周期構造

といった性質である。静止画的 AE では「動きそのもの」を表現しきれず、歩行の協調性や周期性など重要な構造を捉えられない点が課題として残った。

3.2.4. LSTM-AE の導入：時系列構造の学習

時間方向の動作構造を扱うため、エンコーダおよびデコーダ双方に LSTM (Long Short-Term Memory) を組み込んだ LSTM-AE を設計した。本研究では、30 フレームを 1 シーケンスとして入力し、歩行の時間変化を直接学習できるようにした。

この拡張により、モデルは以下のような歩行に内在する時系列的特徴を潜在空間に獲得できるようになった。

- 左足が前に出た後は右足が前に出る。
- 右手が上がると次の瞬間には左手が下がる。
- 上下動は一定の周期性を持つ。

このように、LSTM-AE を導入することで、歩行動作の時間構造を保持した潜在表現を獲得することが可能となった。

3.3. 可聴化

潜在変数の値を Cycling' 74 Max の音響パラメータに割り当てた (図 2)。特に、上半身系の潜在変数は高音域、下半身系は低音域へマッピングし、動作の構造が音響変化として知覚されることを意図した。本研究における可聴化の主目的は、音楽作品として完成度の高いサウンドを生成することではなく、歩行動作の潜在構造を聴覚的フィードバックとして提示することである。そのため、潜在変数の変化を、音高やワンポールフィルタの係数に反映させる素朴なマッピングを採用している。

3.4. 可視化

- レーダーチャート状可視化：8 次元潜在値を極座標上に配置し形状として提示するレーダーチャート型可視化を採用した。(図 3)。
- 折れ線グラフ状可視化：潜在変数の時間変化を折れ線グラフとしてプロットし、周期性や変化パターンの理解を支援する (図 4)。

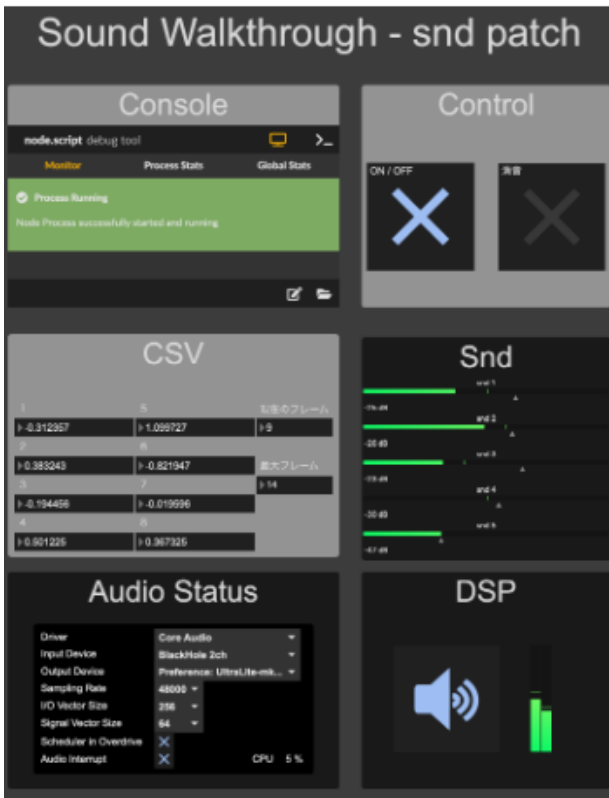


図 2: Max 上の UI 表示例

3.5. システム構成

本研究で構築したシステムは、ブラウザ上で動作する骨格推定・データ処理・可視化モジュール、Python による機械学習モジュール、および Max による音響生成モジュールの三つの主要コンポーネントから構成される (図 5)。

まずブラウザ側では、カメラ映像から MediaPipe によって骨格推定を行い、得られた 99 次元の座標データを HTTP POST リクエストとして Python 側へ送信する。Python 側はこのリクエストを受け取り、学習済みモデルに基づく推論処理を実行し、結果として 8 次元の値を JSON 形式でブラウザに返す。

受信した 8 次元データは、ブラウザ側で情報抽出処理を施したうえで CSV 形式へ変換され、可視化処理および後段の音響生成モジュールで用いるデータとして利用される。ブラウザと Max の通信には WebSocket を使用しており、ブラウザ側で推論データが更新されるたびに、その CSV 形式データが Max へ送信される。Max は受信した CSV データに基づき音響生成処理を行う。

以上により、ブラウザ発の骨格推定 → Python による推論処理 → ブラウザでの抽出・可視化 → Max での音響生成、という一連の処理がほぼリアルタイムで実行され、機械学習による時系列推論結果を視覚的・聴

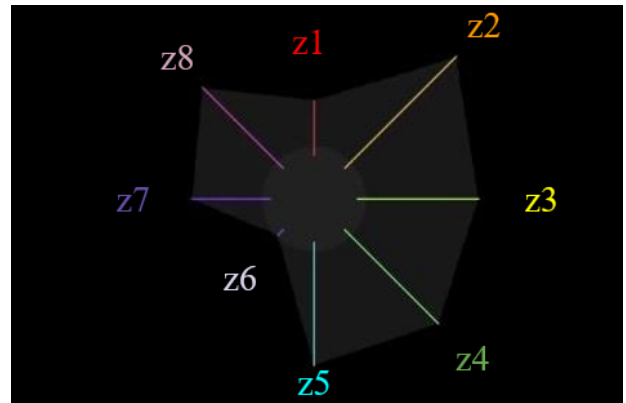


図 3: レーダーチャート状可視化の例

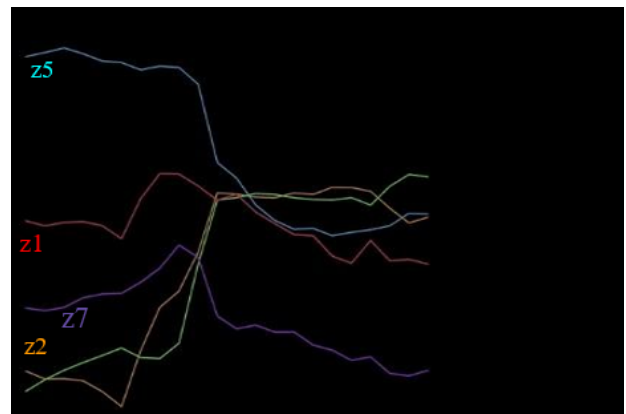


図 4: 折れ線グラフ状可視化の例

覚的に同時提示するインタラクティブなシステムを実現している。

4. 結果

本節では、相関分析に基づく潜在空間と身体部位の関係、および可聴化・可視化の観察結果について述べる。

4.1. 潜在空間と身体部位の関係

相関解析の結果、 z_1 が上半身の座標と比較的強い相関を持ち、 $z_6 \sim z_8$ が下半身座標と強い相関を持つ傾向が確認された (図 6)。一方で z_4 など一部の潜在次元は、全体として相関が低く、特定の身体部位と単純には対応しない抽象的な特徴を捉えている可能性が示唆された。

また、本研究では潜在次元数を 8 次元に設定したが、相関の分布を詳細に確認したところ、主に 4 次元程度が入力特徴量と強い関係を持ち、残りは相対的に寄与が小さいことが観察された。このことは、潜在次元数の設定が過大であった可能性や、モデルがノイズ的な次元を抱えている可能性を示している。

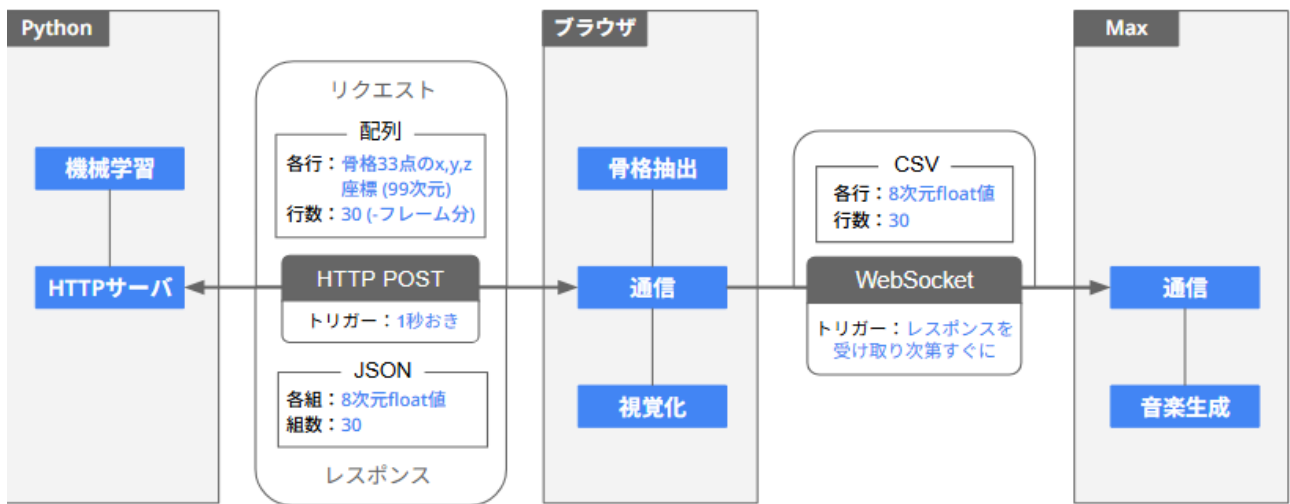


図 5: システム構成のイメージ

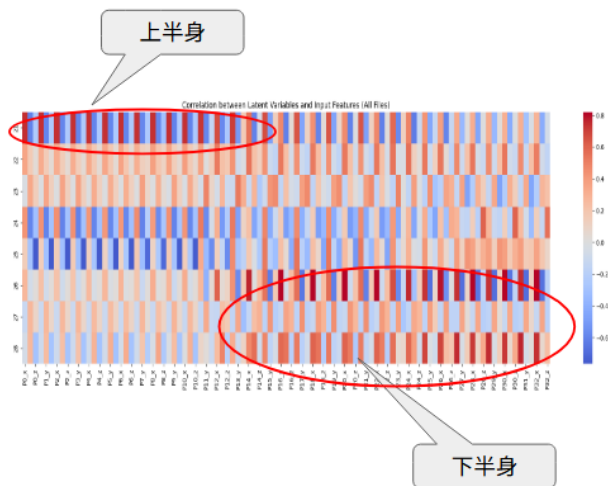


図 6: 潜在変数と入力特徴量の相関ヒートマップ

4.2. 可聴化・可視化の観察結果

本研究では、上半身系の潜在変数を高音域、下半身系を低音域にマッピングするなど、身体部位によって音響帯域を分ける設計を試した。しかし、実際に体験したところ、被験者ごとの歩行の違いが聴感上明瞭に区別できるほどの差異は得られなかった。一方で、潜在変数が変動することに伴い音高や視覚形状が確かに変化することは確認され、潜在空間の動きが視聴覚的フィードバックとして提示されること自体は有効であると分かった。

一方で、現在の単純なマッピングでは、音高・音程の選択において調性や機能音声に基づく制約を設けていない。そのため、スペクトル的あるいは伝統的音声

理論上の協和度が低い音程関係を含む音響も多く生成された。2025年11月2日に国立音楽大学で実施した展示においても、来場者からは「歩行を音として聴く発想が面白い」という肯定的な意見とともに、「人ごとの音の違いがまだ分かりにくい」、「不協和音が多く聞き心地はよくない」といった感想が寄せられた。

可視化に関しては、レーダーチャート状可視化がその瞬間における潜在構造の偏りを直感的に示し、折れ線グラフ状可視化が歩行周期の波形やリズムの安定度を視覚的に確認するのに有効であった。特に、複数被験者の結果を比較した際には、潜在変数の振幅や周期の違いが視覚的に現れ、歩き方の個性の一端を示す手がかりとなった。

5. 考察

5.1. 潜在表現の解釈性と次元数の妥当性

相関分析の結果、潜在変数の一部は上半身・下半身といった特定の身体部位と対応する傾向を示した。これは、LSTM-AEが歩行動作に内在する周期性や身体部位ごとの振る舞いを潜在空間へ反映していることを示唆する。特に、 z_1 (上半身) や $z_6 \sim z_8$ (下半身) のように明確な偏りが見られた潜在変数は、身体構造に沿った抽象特徴を自動的に抽出しており、AEモデルが一定の表現獲得能力を有していることを示している。

一方で、 z_4 のように相関が弱い潜在変数も存在し、これらは身体部位と直接的に結びつかない抽象的特徴(身体全体の微小な揺れ)を捉えている可能性があるが、現段階では解釈が難しい。また、全体としては8次元

のうち実質的に寄与の大きい次元が限定されており、潜在次元数の設定はヒューリスティックであったと言える。今後は、情報量基準やベイズ推定などを用いて、潜在次元数の妥当性を定量的に評価する必要がある。

5.2. 可聴化の役割と限界

潜在変数を音響パラメータへマッピングする可聴化手法は、動作の変化を聴覚的に理解する手段として一定の効果を持った。上半身系の潜在変数を高音域に、下半身系を低音域に割り当てた設計により、身体性の違いが音響的な帯域差として反映され、歩行動作との対応が直感的に把握しやすくなった。

ただし、本研究の目的は「音楽的に調和した作品」を生成することではなく、歩行動作の潜在構造をそのまま音としてフィードバックすることである。そのため、和声的な制約を設けず、潜在変数の推移に比較的忠実な音高変化を採用した結果、不協和音が含まれること自体は設計上許容している。一方で、展示における来場者の意見から、「人ごとの音の違いが分かりにくい」という課題も明らかとなった。これは、潜在空間がまだ十分に個体差を表現できていないこと、およびマッピング設計が差異を強調するよう最適化されていないことに起因すると考えられる。

今後は、(1) 個体差・感情差がより明確に分離される潜在空間の設計と、(2) 人の違いが聴感上も分かりやすくなる音響マッピングを組み合わせることで、「自分の歩行から自分固有の音が生成される」という体験を強化していく必要がある。

5.3. 可視化の意義

レーダーチャート状可視化および折れ線グラフ状可視化は、潜在空間の構造とその時間的推移を理解するために有効であった。レーダーチャート状可視化は各瞬間における「潜在空間のかたち」を直観的に把握するのに適し、身体の偏りや周期的揺らぎが視覚的に表れた。一方、折れ線グラフ状可視化は潜在変数の時間変化を連続的に示した。

さらに、可視化は可聴化を補完する役割も果たし、音響変化と動作構造の対応関係を視覚的に理解する手段として有用であった。潜在変数の解釈性が十分でない状況において、可視化は潜在表現の意味を理解するうえで不可欠な手段であると言える。

5.4. 本研究の限界

本研究には以下のような制約が存在する。

- データ量の不足 (12名) : 個人差や感情差を潜在空間へ十分に反映するには不十分である。

- 感情を意識した歩行の差異が小さい: 歩行は日常的動作であり、感情や人それぞれによる変化幅がダンス等に比べて小さく、潜在空間に明確なクラスタ構造を形成しにくい。
- 潜在変数のノイズ性・解釈性の課題: 特に相関が弱い潜在次元の説明性が低く、次元数の設定もヒューリスティックにとどまっている。

5.5. 今後の課題と展望

本研究では、歩行解析と音楽生成のプロセスを切り分け、歩行から得られる統計的特徴量を音のピッチといった音楽的統語要素へ対応づける可聴化枠組みを構築した。これにより、歩行特性の個性を音響として提示する基盤は整ったといえる。しかし、歩行動作に内在する「個性」や「情動的価値」を、音楽がもつ「意味」構造や「情動的効果」と結びつけるためには、単に機械学習モデルの高度化や特徴量の精緻化を図るだけでは不十分である。

音楽は、音高・リズム・強弱といった表層的な統語レベルに加え、文化的象徴性、音画性、情動価値など、多層的な表象構造を備えている。本研究で実現した可聴化は、主として統語レベルとの対応付けに基づくものであり、歩行が有する身体性・情動性と音楽の意味レベルとの連結には、なお大きな検討余地が残されている。

この課題を克服するには、歩行と音楽の表象を直接的に関連付けるための新たな中間表現の設計や、両者を共通の表現空間にマッピングする機械表現の開発が不可欠である。同時に、マッピングの妥当性を聴覚的観点から反復的に検証する姿勢が重要となる。すなわち、歩行特徴量がどのような音響的変換に対して知覚的意味をもつのか、どのような音楽的操作が情動的価値を付与し得るのかを、聴取者の経験を基軸として評価し続ける必要がある。

さらに、可聴化を“作品”あるいは“経験”として成立させるには、音楽の時間構造や聴覚特性を積極的に活用した音響設計が求められる。技術的に変換が可能であることと、そこから得られる音が体験として意味を持ち得ることのあいだには大きな隔りがある。したがって、「技術的に実現した可聴化」から「聴取体験として成立する音響表現」へと橋渡しする創造的かつ構成論的アプローチが、本研究の延長線上にある重要な課題である。

6. 結論

本研究では、MediaPipe Poseにより取得した歩行骨格データをLSTM-AEによって潜在空間へ圧縮し、その結果を可視化・可聴化することで、「歩行動作から潜在表現へ、そして視聴覚提示へ」という一連のパイプ

ラインを構築した。相関解析の結果、一部の潜在次元は上半身・下半身の座標と対応する傾向を示し、モデルが歩行に内在する身体構造や周期性を捉えていることが確認された。一方で、相関が弱く解釈が困難な潜在次元も存在し、次元数の設定や表現効率については今後の検討が必要である。

可聴化では、潜在変数の変化を音高などにマッピングすることで歩行構造を聴覚的に提示できたが、被験者ごとの差異が明瞭に知覚されるには至らず、調性や音響設計を考慮しない素朴なマッピングの限界も確認された。可視化については、レーダーチャート状および折れ線グラフ状の表示が潜在構造の把握に有効であり、音響変化と動作構造の対応理解を補助する視覚的手がかりとして機能した。

課題としては、データ数の不足、感情条件間の差異の小ささ、潜在次元の解釈性の問題などが挙げられる。今後は、より多様なデータ収集、潜在空間の再設計（次元選択、VAE や Transformer 系モデルの導入）、調性・音色・和声を含む音響マッピングの改善により、「歩行の個性が音として立ち上がる」体験を強化していく必要がある。また、身体動作と音楽の意味構造を橋渡しする中間表現の設計に取り組むことで、身体性と音楽表現を統合する新たな創作・研究の可能性が開かれると期待される。

参考文献

- Hii, J., B. L. Tan, W. M. Aizat, J. Yip, S. M. Ayu, and S. H. Salleh. 2023. "Automated Gait Analysis Based on a Marker-Free Pose Estimation Model." *Sensors* 23(4):2054.
- Bevilacqua, F., L. Naugle, and I. Valverde. 2001. "Virtual Dance and Music Environment Using Motion Capture." *MTAC Proceedings*.
- Giomi, A. 2020. "Somatic Sonification in Dance Performances. From the Artistic to the Perceptual and Back." In *Proceedings of the 7th International Conference on Movement and Computing (MOCO'20)*.
- Landry, S., and M. Jeon. 2020. "Interactive Sonification Strategies for the Motion and Emotion of Dance Performances." *Journal on Multimodal User Interfaces* 14:167-186.
- Han, B., Y. Li, Y. Shen, Y. Ren, and F. Han. 2024. "Dance2MIDI: Dance-Driven Multi-Instrument Music Generation." *Computational Visual Media* 10:791-802.