プロジェクト報告書（最終）Project Final Report

Dynamics Insights : Unveiling Complex Patterns

提出日 1 月 21 日
樋田悠馬 Yuma Hida

1　背景

本プロジェクトではダイナミクスから複雑なパターンを探すことで複雑な現象を簡単にとらえることを目的に活動した．私たちが着目したダイナミクスは馬の動きと環境音である．

まず，馬の動きはアニメーション作成において，馬を描くことが難しい．馬を描くことが難しいのは，馬が動いた時に，動かしている部位が多いことが主な要因であり，さらに，馬の大群ともなると，人の力だけで作画するには，大きな負担となる．そこで，数理モデル化された動く馬の骨組みのようなものを製作すれば，馬を描く時の補助になると考えた．

次に，環境音とは，私たちの身の回りから日常的に聞こえてくる音のことである．例えば，自然界の鳥のさえずりや風の音，都市部の交通騒音，家庭内の生活音，人の声などがある．この他の雑音も含めて環境音と呼ばれる．近年，環境音の分析や生成は様々な側面で活用されつつある．環境音分析は，都市計画や防災システムの異常音検知という社会的利用や，映画や演劇に向けて人工的に作った音声など芸術創作での文化的利用があり，その有用性は高まっている．例えば，神戸新聞 NEXT（2023/3/3）によると,「兵庫県加古川市は，街頭で夜間の悲鳴や怒声などの異常音を検知し，音声と回転灯で警告する人工知能（AI）搭載カメラの設置を始めた.」という [1]．異常音を検知し，対象者に注意を促すことにより，犯罪や事故の防止につなげることができるとして活用されている．

2　課題の設定と到達目標

2.1　馬班

本グループでは，数理モデル化の題材として馬の動きに注目した．課題は，馬の動きの数理モデルの作成方法と，モデルと実際の馬の動きの比較方法である．そのために、馬のデータを取得することが課題になった。到達目標として，全体の動きを再現するのは難しいと考え，まずは馬の前脚の動きの数理モデル化することを目指す.

2.2　音班

本グループでは，日常的に聞こえる環境音に注目した．課題は，環境音の特徴をどのようなアプローチで捉えるか，環境音データをどのように収集するかである．到達目標としては，環境音の特徴的な要素を分析・抽出し，それに基づいて新たな環境音を生成することを目指す.

3　課題解決のプロセスとその結果

3.1　馬班

まず馬の全体の動きを数理モデル化するのは難しいと考え，馬の足の動きに絞って考えることにした．荒井ら（2011）は，人間の足を，粘性減衰を含む単振り子の運動方程式を用いて再現し，人間の足の動きを振り子で数理モデル化する妥当性を示した [2]．したがって，馬の足の動きを数理モデル化するのに，二重振り子の運動方程式をベースに改良を加えていった．

また，課題の一つであったデータの取得について，ハイスピードカメラによって撮影された，馬が走る映像 [3] を用いて，座標を取得し，それぞれの角度を計算するシステムを作成した．

最後に，馬の動きに近づける方法については，作成した数理モデルをシミュレーションしたものと実データを，比較して，数理モデルを改良するプロセスを繰り返す方法を取った．比較する方法として，実データを x，その許容誤差を ϵ としたときに，$x \pm \epsilon$ 内にシミュレーションの値がいくつ入っているか数を数え，パーセントで表示した．また，図 1 のように，二重振り子の角度 θ_1 と θ_2 を，実データとシミュレーション，それぞれ計算し，比較した．

結果として，$\epsilon = \pi/8$ のとき，θ_1 は約 94 パーセント．θ_2 は約 55 パーセントまで近づけることができた．(図

2) 数理モデルの構造としては，元になった二重振り子の数理モデルに，sin 波をいくつか加えたものになっている．これにより，馬の前脚をモデル化し，シミュレーションすることができた（図 3）．

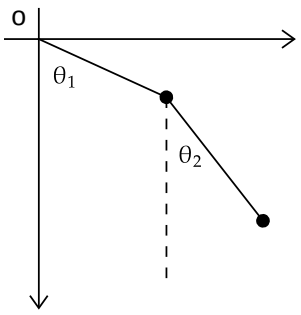

図 1: 取得した角度のイメージ

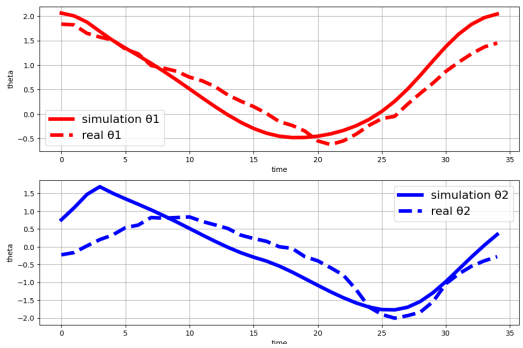

図 2: 作成したモデルと，実データの比較

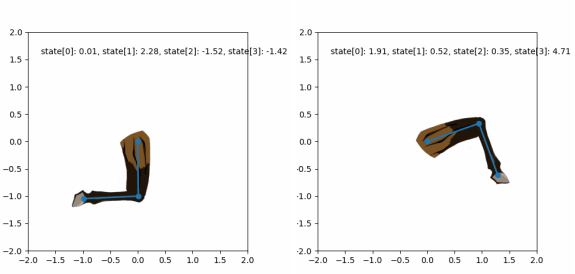

図 3: 完成したモデルのシミュレーション

3.2 音班

音班は音の特徴を捉えるという課題について，3 つのアプローチで試みた．1 つ目は，環境音分析，2 つ目は，特徴量抽出，3 つ目は，環境音生成である．3 つのアプローチを用いた理由は，多角的に音の特徴を捉えるためである．また，環境音データについては，50 種類の環境音が各クラス 40 個，5 秒間ずつあり，合計 2000 個のデータセットである ESC-50[4] を用いた．これに加えて，実際に収録した環境音も用いた．

3.2.1 前処理

高速フーリエ変換（FFT）を用いて音声波形を周波数領域に変換した．これにより，時間軸上の音声信号を各周波数成分に分解することが可能となる．Python の numpy ライブラリを使用して FFT を実行し，音声データをスペクトル表現に変換した．

実録した環境音の前処理には，作編曲ソフトである Studio One 6 artist を用いた．まず，不要な低周波数や高周波数を除去するため，各音声データに適切な値のハイパスフィルターとローパスフィルターを適用した．これにより，目的の音がよりクリアに際立つようになった．また，音声データ内の不要な高音域の歪みを軽減するため，ディエッサーを使用した．さらに，目的の音を際立たせるためにノーマライズとコンプレッサーの処理を行った．ノーマライズでは音声全体の音量を一定の範囲に調整し，コンプレッサーでは音量のダイナミクスを制御して聞き取りやすさを向上させた．これらの前処理により，分類タスクでの音声認識精度の向上が期待できるデータセットを生成することができた．

3.2.2 環境音分析

環境音分析とは，日常生活に存在する多様な音を分析することである．そして，環境音分析は環境音認識または環境音分類技術に応用されている．環境音認識は異常音検知などの防災システムや自動運転技術など，多岐にわたる分野で活用されている．環境音分類では，畳み込みニューラルネットワーク（CNN）を用いて分類精度の向上を試みた．CNN は主に画像認識で用いられる手法であるが，音声をスペクトログラムという画像形式に変換することで音声認識にも適用が可能となる．環境音認識と分類の共通課題として，音の種類が多様であるため，識別や分類が難しいことがある．また，実録した環境音を適用する場合，認識対象外の音も含まれることがある．そのため，ノイズ除去や音源分離などの技術が必要となる．CNN は，畳み込み層，プーリング層，全結合層で構成され，特徴抽出と分類を行う．今回はデータとして ESC-50 を用いた．分類結果として，約 89%の分類精度を達成した．

3.2.3 特徴量抽出

主に環境音の特徴を抽出するために Variational Autoencoder（VAE）を用いた．VAE とは，通常のオートエンコーダにエンコーダから取得した平均と分散を用いて正規分布にしたがって生成された潜在空間を加えたものである [5]．今回使用した VAE のモデルは，潜在空間の次元を 64 次元とした．エンコーダにメルスペクトログラム入力することで学習を行い，特徴量を抽出する．この潜在空間を独立成分分析（ICA）を用いて 2 次元へ次元削減することで，潜在空間を可視化することができる．山際（2023）によると，単語や画像では，ICA は人間が解釈しやすい成分を多く取り出すことができると述べられている [6]．そのため，今回は音でも ICA を用いて次元削減をすることにした．図 2 は，犬の鳴き声と赤ちゃんの泣き声を VAE に通したときの潜在空間を可視化したものである．この結果から，エンコーダが種類ごとの特徴の違いを捉え，抽出できていることが示された．

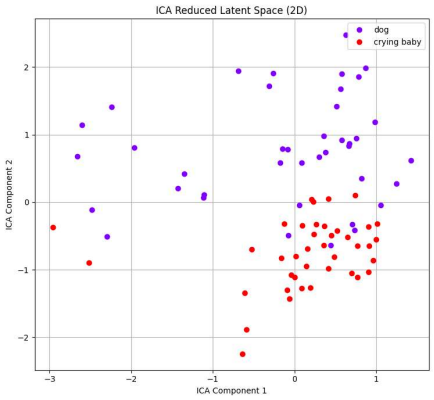

図 4: 潜在空間

VAE 以外の特徴量抽出方法として，Google が開発した VGGish モデル [7] も活用した．VGGish は，音声データをメルスペクトログラムに変換し，その特徴を VGG ベースのニューラルネットワークで抽出する手法である．メルスペクトログラムとは，スペクトログラムの周波数軸を人間の聴覚システムに近似するメル尺度に変換したものである [8]．このモデルは，事前学習された音声特徴抽出モデルであり，特に環境音や一般的な音声解析タスクにおいて有効である．VGGish を使用することで，高次元の音声データをコンパクトな固定次元の特徴ベクトルに変換することが可能となる．特徴量抽出においては，128 次元に圧縮された特徴量を ICA を用いて 2 次元へ次元削減を行い，可視化した．

3.2.4 環境音生成

VAE を用いて環境音生成を行った．VAE のエンコーダに，メルスペクトログラムを入力することで学習を行い，デコーダから再構成されたメルスペクトログラムが出力される．図 3 は，実際の鶏（rooster）の鳴き声のメルスペクトログラムである．図 4 は，その鶏の鳴き声を再構成し，出力されたメルスペクトログラムである．精度は平均二乗誤差（MSE）約 0.0038 であった．

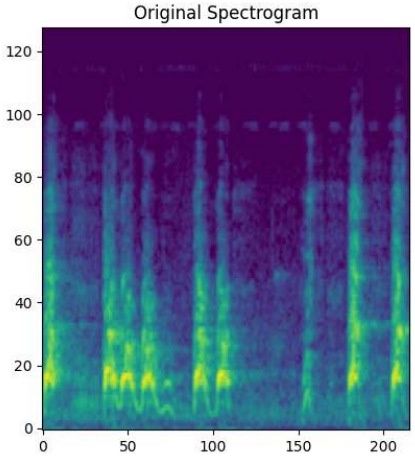

図 5: 元データのメルスペクトログラム

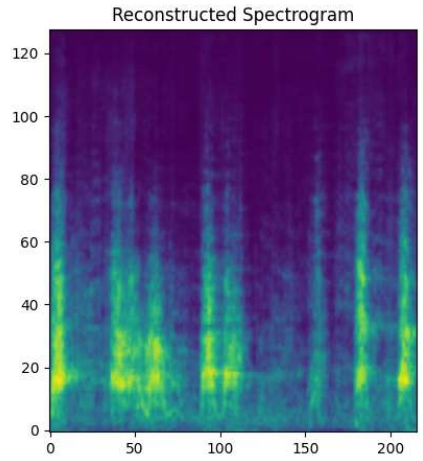

図 6: 再構成後のメルスペクトログラム

学習段階で潜在空間に特徴量がプロットされる．これは正規分布に従うように生成される．プロットされた特徴量から少し値を変えたものをデコーダに入力することで，元の音声と類似した特徴を持つメルスペクトログラムが出力された．

さらに，潜在空間において異なる 2 種類の音の特徴量の間に内分点をとり，それをデコーダに入力することで，どちらの特徴ももつ新たな音の生成を達成した．

4　考察と今後の課題

4.1　馬班

今回のプロジェクトでは，実データとシミュレーションの比較を行った結果，θ_1 が 94 %に対し，θ_2 が 52 % と，θ_1 は高かったものの θ_2 はそこそこといった結果になった．このような結果となったのは，下の振り子が上の振り子の影響を強く受けるためである．上の振り子は，単振り子とほとんど変わらないため，実際の馬の周期に合わせるのは簡単だが，下の振り子は，上の振り子の周期の影響を受けながら，別の周期で動かなければならないため，実際の馬に近づけるのが困難であった．そのため，馬の全身の動きを再現する際は前脚のモデルに変数を追加する手法では困難であると予想される．ゆえに，それぞれ独立した馬のパーツをつなぎ合わせて再現する手法が良いと思われる．ただ，それで自然な馬の動きになるかは，わからないため実践して試す必要がある．

また，今回高い精度で再現できた動きは 1 通りの動きであり，機械的で生物のようにはいかない．そこで，ノイズを加えてランダムな動きのシミュレーションを生成し組み合わせることで，より生き物のようになり，大量の馬が一緒に走ったとしても自然になると思われる．しかし，二重振り子の複雑さゆえに，小さいノイズでも大きく動きが変わるため，解決策を模索する必要がある．

4.2　音班

本グループは，高精度の環境音分類や VAE を用いた環境音の再構成，生成を可能にした．しかし，精度について平均二乗誤差より完全に再構成することはできなかった．その原因として，VAE の情報圧縮における点と，音声とスペクトログラムとの変換における点が考えられる．VAE について，エンコーダ部分が十分に特徴圧縮を行えなかったことが考えられる．また，音声とスペクトログラムとの変換について，変換時に位相情報が失われてしまったことが原因だと考えられる．これを解決するために，短時間フーリエ変換の冗長性に基づいた位相再構成手法である Griffin-Lim アルゴリズム [9] を用いて位相の推定をしたが，完全に再現することはできなかったと考えられる．今後の課題として，VAE のファインチューニングが挙げられる．VAE は画像データの処理を得意とする手法である．そのため，音声のような時系列データの処理にも対応したチューニングを行う必要がある．また，振幅情報と位相情報の両方を失うことなく変換する手法の検討が挙げられる．この他にも，VAE の潜在空間への入力を調整することで，VGGish から抽出した特徴量を後から VAE の潜在空間へ入力し，環境音生成の精度向上を図ることも挙げられる．

参考文献

[1] 神戸新聞（2023）夜間の異常音検知.「監視中です」音声と回転灯で警告加古川市が AI カメラ 150 台設置．https://x.gd/1WnO5（2025/01/08 アクセス）

[2] 荒井美沙子, 高橋亜佑美, 美坐地一人（2011），歩行に関する人体の数理モデル化研究. 日本大学生産工学部第 44 回学術講演会概要, 1053-1056p.

[3] NHK.“走る馬 スピードカメラ”. NHK アーカイブス,2008．https://www2.nhk.or.jp/archives/movies/?id=D0002060199_00000

[4] Karol J. Piczak.（2015） Esc: Dataset for environmental sound classification. *Proceedings of the 23rd ACM international conference on Multimedia*, 1015-1018

[5] 我妻幸長（2020）はじめてのディープラーニング 2. SB クリエイティブ株式会社，東京，pp.222-231

[6] Hiroaki Yamagiwa, Momose Oyama, Hidetoshi Shimodaira（2023）Discovering Universal Geometry in Embeddings with ICA．*EMNLP2023*, pp.4647-4675

[7] S. Hershey, S. Chaudhuri, D.P.W. Ellis, J.F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R.A. Saurous, B. Seybold, M. Slaney, R.J. Weiss, and K. Wilson.（2017）CNN Architectures for Large-Scale Audio Classification. Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing*（*ICASPP*）, pp.131-135

[8] Hugging Face（2023）Introduction to audio data．https://huggingface.co/learn/audio-course/chapter1/audio_data（2025/01/15 アクセス）

[9] Papers With Code（2020）Griffin-Lim Algorithm. https://paperswithcode.com/method/griffin-lim-algorithm（2025/01/15 アクセス）