

公立はこだて未来大学 2024 年度 システム情報科学実習 グループ報告書

Future University Hakodate 2024 Systems Information Science Practice
Group Report

プロジェクト名

Dynamics Insights: Unveiling Complex Patterns

Project Name

Dynamics Insights: Unveiling Complex Patterns

グループ名

音班

Group Name

Sound Group

プロジェクト番号/Project No.

19-B

プロジェクトリーダー/Project Leader

樋田悠馬 Yuma Hida

グループリーダー/Group Leader

寺崎僚将 Ryosuke Terasaki

グループメンバ/Group Member

寺崎僚将 Ryosuke Terasaki

関根拓斗 Takuto Sekine

加藤康介 Kosuke Kato

指導教員

加藤譲 義永那津人 栗川知己 ヴラジミールリアボフ リヴァーズダミアン

Advisor

Yuzuru Kato Natsuhiko Yosinaga Tomoki Kurikawa Volodymyr Riabov Damian Rivers

提出日

2025 年 1 月 21 日

Date of Submission

January 21, 2025

概要

「Dynamics Insights」では自然界で観察される複雑な現象やパターンに焦点を当て、様々な手法を用いて複雑系のダイナミクスに関する新たな知見を得ることを目的とする。私たちは、様々な現象やパターンの中で普段生活をしている。その生活の中で、常に無意識的に感じているものの一つに音というものがある。その音は、自分が発している音や自分の周りの人が発している音、人だけでなく木々の音や風の音など、その場に存在しているものから聞こえてくる。そのような生活をしているときに聞こえる音は環境音と呼ばれている。環境音は、様々な音が重なり合った複雑な音であり、私たち人の耳では具体的にどのような音が重なり合っているのかを詳細には判別できない。そこで本グループは、このような人の耳では詳細な判別が難しい環境音の分類について着目した。環境音の研究は、機械学習の発展によって、音の分析や合成が容易になってきたことをきっかけに多くの研究がされている。例えば、環境音をスペクトログラム化して画像として扱い、畳み込みニューラルネットワーク（CNN）を用いた環境音分析が多くされてきた。また、その他の機械学習の手法も環境音分析に用いられている。私たちは、様々な機械学習の手法を試行し、環境音の分類性能を向上させることを目標とする。そして、各環境音の科学的な距離を定義すること、分類結果から音の特徴を捉え、新たな環境音を生成することを目標とする。

キーワード 環境音分析, 環境音分類, 環境音生成, 異常音検知, 機械学習

(※文責: 寺崎僚将)

Abstract

“Dynamics Insights” focuses on gaining new insights into the dynamics of complex systems using various methods, emphasizing the complex phenomena and patterns observed in science and technology. We live our daily lives amidst various phenomena and patterns, and one of the things we constantly and unconsciously perceive is sound. These sounds include those we produce ourselves, those made by people around us, and sounds from our environment such as the rustling of trees and the blowing of the wind. The sounds we hear during our daily lives are called environmental sounds. Environmental sounds are complex, composed of various overlapping sounds, making it difficult for the human ear to distinguish the individual components in detail. Our group has focused on classifying these environmental sounds, which are challenging to distinguish in detail with the human ear. Research on environmental sounds has advanced significantly with the development of machine learning, which has made it easier to analyze and synthesize sounds. For instance, many studies have analyzed environmental sounds by converting them into spectrograms and treating them as images, using convolutional neural network (CNN). Other machine learning techniques have also been applied to environmental sounds analysis. We aim to improve the classification performance of environmental sounds by experimenting with various machine learning methods. Additionally, our objectives include defining the scientific distances between environmental sounds, capturing the characteristics of sounds from classification results, and generating new environmental sounds.

Keyword Environmental sounds analysis, Environmental sounds classification, Environmental sounds generation, Anomaly sound detection, Mechanical learning

(※文責: 寺崎僚将)

目次

第 1 章	はじめに	1
1.1	背景	1
1.2	目的	1
第 2 章	関連研究	2
2.1	若色・小坂（2022）による研究	2
2.2	松原ら（2019）による研究	2
第 3 章	プロジェクト学習の目標	3
第 4 章	目的を達成するための手法，手段	4
4.1	環境音分類課題	4
4.2	特徴量抽出課題	4
4.3	環境音生成課題	4
4.4	前処理	5
第 5 章	結果	6
5.1	環境音分類課題	6
5.2	特徴量抽出課題	6
5.3	環境音生成課題	8
第 6 章	考察	10
6.1	環境音分類課題	10
6.2	特徴量抽出課題	10
6.3	環境音生成課題	10
第 7 章	まとめ	12
	参考文献	13

第 1 章 はじめに

1.1 背景

私たちの日常は多様な音に囲まれている。例えば、音楽を鑑賞したり、音声を用いて他者とコミュニケーションを取ったりしている。それ以外にも、鳥の鳴き声を聞いてその存在を確認したり、自動車のクラクションや人の叫び声を耳にして危険を察知している。そのような音声や楽音に限らない音は環境音と呼ばれており、私たちはそれを聴取して身の回りの環境を理解している。

環境音分析とは、計算機を用いて環境音进行分析し、音の種類や音が発生した状況などを推定する技術であり、計算機を用いて環境音进行分析することは長い間検討が進められている。この技術は、音声や画像の認識技術と類似しており、異常音が発生したとき、その事象を検出する技術である異常音検知などへの活用がある。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。しかし、精度に改善の余地があり、精度向上に向けた研究が行われている。

環境音生成は環境音を人工的に生成する技術である。映画や音楽などのメディアコンテンツ制作において、環境音は状況説明などの観点から重要な役割を担っている。しかし、目的に合致する環境音をデータベースから探すことは、膨大な時間を要する。そのため、深層学習を用いて簡単に環境音を生成する研究が行われている。研究の課題として、満足できる生成ができていない、生成できる音がまだまだ限られていることが挙げられる。

(※文責: 加藤康介)

1.2 目的

これらの課題を解決するとき、音の持つ特徴を理解することで、分析精度の向上や特徴をとらえた生成を可能にすると考えた。これより、本グループでは環境音の時系列に潜む特徴を、分析や生成を通して捉えることを目的とした。

(※文責: 加藤康介)

第 2 章 関連研究

2.1 若色・小坂（2022）による研究

若色・小坂（2022）による音をクエリとする類似環境音検索システムの評価方法の検討がある [2]。若色らは、畳み込みニューラルネットワーク（CNN）による音響特徴量抽出を活用し、環境音データセットを用いた音データの分類と検索システムを提案した。このシステムでは、特徴量を 128 次元で抽出し、主成分分析により 20 次元に圧縮したうえで、類似度関数を用いてクエリ音との近似性を計算し、検索結果として最も近い音を提示する。この手法により、大分類の一致率が 80% 以上という成果を得ており、検索結果がユーザの期待に概ね合致することが示された。

（※文責: 加藤康介）

2.2 松原ら（2019）による研究

松原ら（2019）による CNN Autoencoder から抽出したボトルネック特徴量を用いた環境音分類がある [3]。松原らは、異なる収録条件を持つデータセットを用いて CNN Autoencoder から抽出されたボトルネック特徴量の有効性を検証し、従来の CNN ベースの分類手法と比較して性能の向上を示した。これにより、収録環境に依存しない環境音分類をすることが可能になる。

（※文責: 加藤康介）

第 3 章 プロジェクト学習の目標

音班は音の特徴を捉えるという課題について、3つのアプローチを試みた。1つ目は、環境音分類、2つ目は、特徴量抽出、3つ目は、環境音生成である。この3つのアプローチから、多角的に音の特徴を捉えることを目標とする。

(※文責: 寺崎僚将)

第 4 章 目的を達成するための手法，手段

本グループは，環境音の時系列に潜む特徴を捉え，それをもとに分類，生成を行うことで，実世界の音響環境を再現，分析することを考案した．また，環境音の分類，特徴量抽出，生成のプロセスを通じて，実録データへの適用の可能性を検証した．

(※文責: 加藤康介)

4.1 環境音分類課題

環境音分類課題については，手法として畳み込みニューラルネットワークを用いた．CNN：Convolutional Neural Network（畳み込みニューラルネットワーク）とは，主に画像データの処理に用いられるディープラーニングのためのネットワークアーキテクチャである [4]．CNN は，入力層，出力層とその間にある多くの隠れ層で構成されており，データに特有の特徴を学習する目的でデータを変更する働きがある．各層での作業を数十から数百回繰り返し行われ，層ごとに異なる特徴を識別するように学習する．

(※文責: 加藤康介)

4.2 特徴量抽出課題

特徴量抽出課題については，手法として VAE と VGGish を用いた．VAE：Variational Autoencoder（変分オートエンコーダ）とは，生成モデルの一種で，観測データについて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従い生成する確率的なモデルである [5]．一般に，データの生成や特徴の潜在表現を得るために使用される．

VGGish とは，Google が開発した音声特徴量抽出のための深層学習モデルである [6]．このモデルは，画像分類で知られる VGG を基に設計されており，主に音声データや環境音データから高次元な特徴を抽出するために使用される．使用したモデルは github で参照できる [7]．

(※文責: 加藤康介)

4.3 環境音生成課題

環境音生成課題については，手法として VAE を用いた．また，使用したデータは既存の環境音データセットだけでなく，実録した環境音も用いた．

(※文責: 加藤康介)

4.4 前処理

これらの課題に取り組むにあたり、最初に環境音データの前処理として、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。

(※文責: 関根拓斗)

第 5 章 結果

本グループの目標は、様々なアプローチから実録した音も含めて、多角的に音の特徴を捉えることである。この目標を達成するために、環境音分類課題では CNN、特徴量抽出課題では VAE と VGGish、環境音生成課題では VAE を手法として用いた。

(※文責: 寺崎僚将)

5.1 環境音分類課題

まず、環境音分類課題について、CNN を用いて環境音の分類精度向上を試みた。CNN は主に画像認識で用いられる手法であるが、音声データをスペクトログラムという画像形式に変換することで、音声認識や環境音分類にも適用可能である。本グループでは、この特性を活かし、CNN を用いた環境音分類の性能向上を目指した。

これらの構造を適切に設計し、環境音データセットである ESC-50 を用いて評価を行った。ESC-50 データセットは、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 [8]。例えば、犬の鳴き声や雨の音、時計のアラーム音などの種類が収録されている。環境音分類の研究において広く用いられているデータセットである。本グループでは、このデータセットを用いてモデルをトレーニングし、分類精度の向上を図った。その結果、約 89% という高い分類精度を達成した。この精度は、環境音分類の分野において競争力のある成果であり、提案したモデルが環境音を効果的に識別できることを示している。

(※文責: 寺崎僚将)

5.2 特徴量抽出課題

次に、特徴量抽出課題について、主に VAE を用いて環境音の特徴を抽出した。今回使用した VAE のモデルでは、潜在空間の次元を 64 次元に設定した。この設定は、音声データの複雑な特徴を十分に捉え、計算負荷も抑えるためである。学習データとしては、音声をメルスペクトログラムに変換したデータを使用した。メルスペクトログラムは、人間の聴覚特性に基づく周波数スケールを持つ時系列データ [9] であり、音声や環境音の解析に適している。このメルスペクトログラムを VAE に入力し、エンコーダが音声データの特徴を潜在空間にマッピングするように学習を行った。その後、この潜在空間を独立成分分析 (ICA) を用いて 2 次元に次元削減することで、潜在空間を可視化した。山際 (2023) によると、ICA は単語や画像データにおいて人間が解釈しやすい成分を多く抽出できる [10] ことが示されており、本グループでは音声データへの適用を試みた。図 5.1 は、犬の鳴き声と赤ちゃんの泣き声を VAE に通した際の潜在空間を可視化したものである。この結果から、エンコーダが音声ごとの特徴の違いを捉え、適切に抽出していることが確認された。特に、異なる音声データが潜在空間上で明確に区別されることから、VAE を用いた特徴量抽出の有効性が示された。

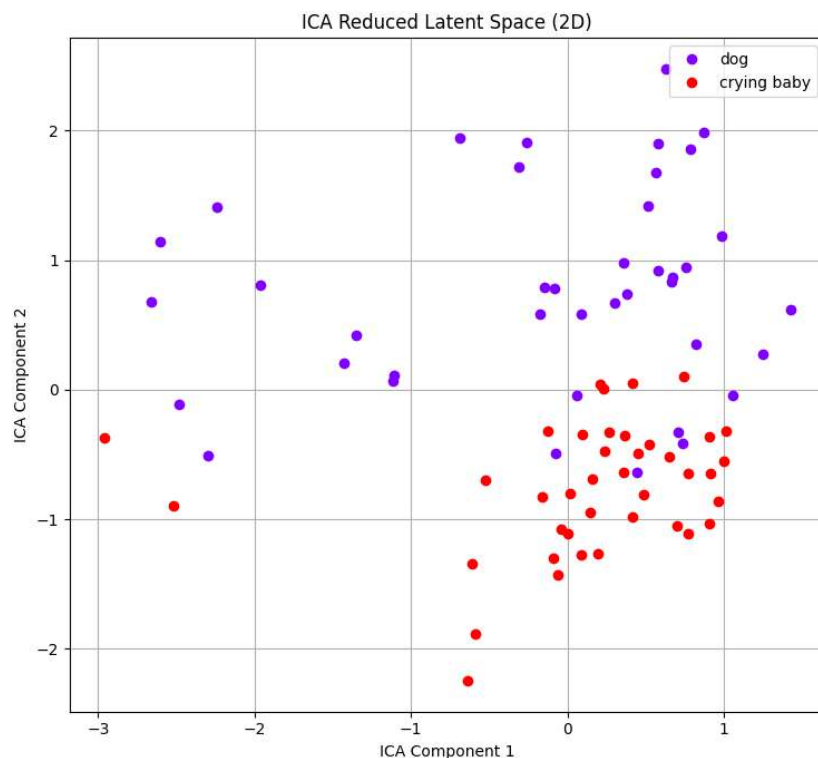


図 5.1 潜在空間

さらに、本グループでは、VAE 以外の特徴量抽出方法として Google が開発した VGGish モデルも活用した。VGGish を使用することで、音声データを固定次元（128 次元）の特徴ベクトルに圧縮することが可能であり、これにより、音声データの処理と解析が効率化された。

特徴量抽出の後、VGGish で得られた 128 次元の特徴ベクトルも ICA を用いて 2 次元に次元削減し、0～19 の 20 種類の Components に分けて可視化を行った。この可視化では、VAE と同様に、異なる音声データが特徴空間上で明確に分離されていることが確認された。

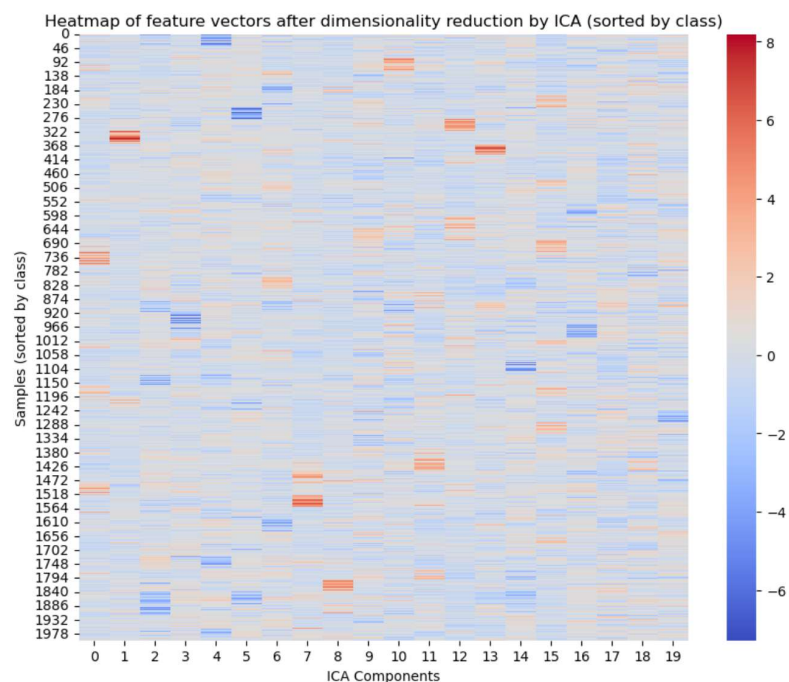


図 5.2 VGGish の特徴空間

例えば, Component 0, 1, 2 では, 以下の音がそれぞれ抽出された.

```
Classes with high feature values in Component 0:  
brushing_teeth: 1 samples  
cat: 17 samples  
chirping_birds: 1 samples  
crow: 5 samples  
crying_baby: 13 samples  
dog: 26 samples  
door_wood_creaks: 4 samples  
frog: 10 samples  
hen: 16 samples  
laughing: 5 samples  
pig: 1 samples  
rooster: 23 samples  
siren: 1 samples  
sneezing: 1 samples
```

図 5.3 VGGish Component 0 の抽出音

```
Classes with high feature values in Component 1:  
airplane: 3 samples  
car_horn: 7 samples  
engine: 2 samples  
helicopter: 2 samples  
sheep: 1 samples  
siren: 7 samples  
train: 12 samples  
washing_machine: 1 samples  
wind: 1 samples
```

図 5.4 VGGish Component 1 の抽出音

```
Classes with high feature values in Component 2:  
brushing_teeth: 1 samples  
chainsaw: 1 samples  
clapping: 1 samples  
door_wood_creaks: 1 samples  
door_wood_knock: 1 samples  
glass_breaking: 2 samples  
hand_saw: 13 samples  
toilet_flush: 5 samples  
train: 13 samples  
vacuum_cleaner: 25 samples  
washing_machine: 7 samples
```

図 5.5 VGGish Component 2 の抽出音

VGGish の特徴量は, 事前学習モデルの利点を活かしており, 特に小規模なデータセットでも高い特徴抽出能力を発揮することが示された.

(※文責: 寺崎僚将)

5.3 環境音生成課題

最後に, 環境音生成課題について, VAE を用いて環境音の生成を行った. 本グループでは, 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

ペクトログラムを出力するプロセスを設計した。

図 5.6 は、左が実際の鶏（rooster）の鳴き声をメルスペクトログラムに変換したもので、右がその鶏の鳴き声を VAE によって再構成したメルスペクトログラムである。この再構成結果は、元の音声と類似しており、再構成精度は平均二乗誤差（MSE）で約 0.0038 という値を記録した。これは、モデルが入力データの特徴を正確に学習し、それを再現する能力に優れていることを示している。

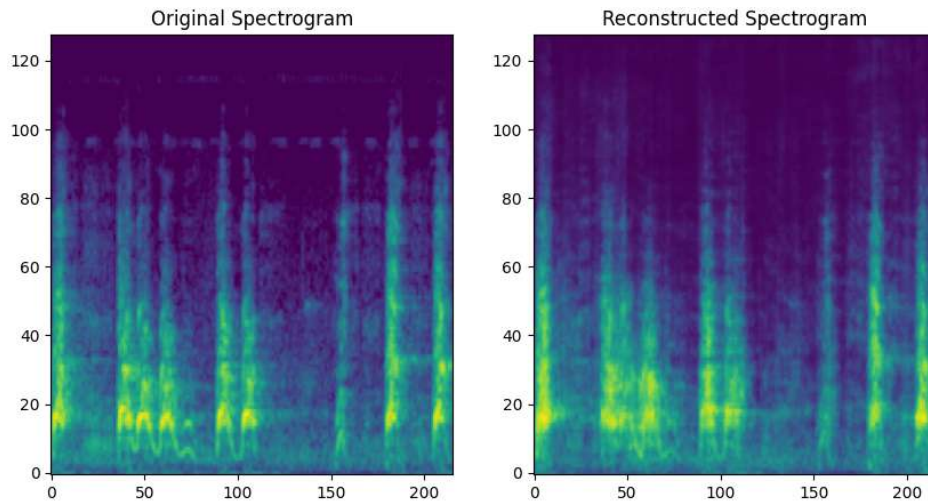


図 5.6 元データと再構成後のメルスペクトログラム

VAE の学習段階では、潜在空間に特徴量がプロットされる。この潜在空間は正規分布に従うように生成されており、音声データの本質的な特徴を圧縮した表現となっている。この潜在空間の特徴は、新たな音声の生成にも利用できる。具体的には、プロットされた特徴量の値をわずかに変化させ、それをデコーダに入力することで、元の音声と類似しながらも異なる特徴を持つメルスペクトログラムを生成することが可能である。この手法により、元の音声データを基に多様な音声バリエーションを生成できる点が確認された。

また、潜在空間における応用として、異なる 2 種類の音の特徴量の間に内分点を取る操作を行った。この内分点の値をデコーダに入力することで、両方の音の特徴を持つ新たな音の生成に成功した。しかし、再構成後の音声と生成した音声のどちらにも、ノイズが入ってしまっていた。

さらに、実録した音も用いて生成を行った。対象の音は、公立はこだて未来大学の敷地内で録音した児童の遊び声を用いた。録音された音には対象外の音が含まれているため、前処理としてノイズ除去を実施した。その後、処理済みの音データを VAE に入力し、潜在空間にプロットされた複数の特徴量の間を選択してデコーダに入力することで音の生成を行った。生成された音は ESC-50 のデータセットに基づいて出力された生成結果に比べ、ノイズが目立つ傾向が確認された。

（※文責: 寺崎僚将）

第 6 章 考察

6.1 環境音分類課題

ニューラルネットワークに基づく環境音分類の従来法としては、CNN, RNN, CRNN に基づく手法などが提案されている。また、Transformer に畳み込みモジュールを組み込んだ Conformer を用いた手法も提案されており、高い性能を示すことが報告されている [11]。これらの手法を用いることで、今回の分類性能をより向上させることが可能であると考ええる。

今回は ESC-50 という既存のデータセットを用いて学習を行った。このデータセットはラベル以外の音が除去されており、音の発生タイミングが明確である。しかし、実環境では目標以外の音も含まれており、さらに音の発生タイミングが不明確なためタイムスタンプ情報ラベルが必要となる。これらの問題を解決するために、弱ラベル学習を行う必要があると考える。弱ラベル学習とは、「音データの中に何の音が含まれるかはわかるがどのタイミングでなっているかはわからない、というゆるいラベルを用いた学習である」と述べられている [12]。弱ラベル学習を行うことで、実環境での環境音分類にも適用できるような頑健なモデルの構築が可能であると考ええる。

(※文責: 加藤康介)

6.2 特徴量抽出課題

結果より、Component 0 では図 5.3 の通りであった。猫 (cat) や犬 (dog)、鶏 (rooster) などの音が多く抽出されているため、Component 0 では生物が発する音を含む特徴をうまく抽出できるといえる。また、Component 1 では図 5.4 の通りであった。車のクラクション (car horn) やサイレン (siren)、電車 (train) などの音が多く抽出されているため、Component 1 では一般に騒音と認識されるような音を含む特徴をうまく抽出できているといえる。さらに、Component 2 では図 5.5 の通りであった。片手のこぎり (hand saw) や掃除機 (vacuum cleaner)、洗濯機 (washing machine) などの音が多く抽出されているため、Component 2 では連続的で周期的な音を含む特徴をうまく抽出できているといえる。

(※文責: 寺崎僚将)

6.3 環境音生成課題

結果より、MSE が 0.0038 であることは、モデルが入力データの音響的特徴を忠実に再現する能力を示しており、再構成モデルとして十分な精度を有していると考えられる。しかし、再構成後の音声にはノイズが入ってしまうため、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 [13] を用いて位相の推定をしたが、完全に位相の再構成をすることはできなかったと考えられる。

また、潜在空間の特徴量を操作することで、多様なバリエーションの環境音を生成できたことは、環境音生成モデルの有用性を示している。しかし、画像生成の領域と比較したときに、生成結果は見劣りするものとなった。これは、生成モデルとして改善の余地があると考えられる。この課題を解決するために、GAN や Transformer などの他のモデルを試行することや、生成結果についての客観的な評価指標を設けることが挙げられる。

実録データの生成結果より、データセットを使用した生成と比べてノイズが目立つ傾向が確認された。これは収録データに対象外の環境音が多く含まれるため、生成結果に直接的な影響を与えたと考えられる。前処理としてノイズ除去を実施したが、完全に対象外の音を排除することは困難であり、これが生成された音声の品質低下に寄与したと考えられる。

(※文責: 加藤康介)

第 7 章 まとめ

本グループでは、環境音の分類、特徴の抽出、音の生成について、それぞれの手法が効果的に機能することを確認した。一方で、生成した音にノイズが含まれる問題や、実環境で使用するためにはモデルをより強化する必要があるといった課題も明らかになった。

今後の方向性としては、音声生成の質をさらに向上させるために、他のモデルとして GAN や先進的な Transformer といったモデルを導入することが考えられる。また、実環境でモデルを活用できるようにするためには、弱ラベル学習という手法を用いることで、より柔軟に対応する必要がある。さらに、生成した音の品質を客観的に評価できる基準を導入することで、モデルの改良を進めることが可能である。

本グループの成果は、環境音分析における新たな可能性を示している。本研究を基盤とし、さらに発展させることで、将来的には環境音を活用したさまざまな分野への応用が期待される。

(※文責: 寺崎僚将)

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] 若色夏美, 小坂直敏 (2022) 音をクエリとする類似環境音検索システムの評価方法の検討. 2022 年 電子情報通信学会総合大会, p.59
- [3] 松原拓未, 原直, 阿部匡伸 (2019) CNN Autoencoder から抽出したボトルネック特徴量を用いた環境音分類. マルチメディア, 分散協調とモバイルシンポジウム 2019 論文集, 2019 : 339-346
- [4] Mathworks (2025) 畳み込みニューラル ネットワーク. <https://x.gd/NdkGS> (2025/01/18 アクセス)
- [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] 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 (ICASSP)* , pp.131-135
- [7] GitHub1 (2024) VGGish. <https://x.gd/zThnx> (2025/01/20 アクセス)
- [8] Karol J. Piczak. (2015) Esc: Dataset for environmental sound classification. *Proceedings of the 23rd ACM international conference on Multimedia*, 1015-1018
- [9] Hugging Face (2023) Introduction to audio data. https://huggingface.co/learn/audio-course/chapter1/audio_data (2025/01/15 アクセス)
- [10] Hiroaki Yamagiwa, Momose Oyama, Hidetoshi Shimodaira (2023) Discovering Universal Geometry in Embeddings with ICA. *EMNLP2023*, pp.4647-4675
- [11] 井本桂右, 川口洋平 (2020) 環境音分析・異常音検知の研究動向. 電子情報通信学会 基礎・境界ソサイエティ Fundamentals Review, 15(4) : 268-280
- [12] 小松達也 (2020) 環境音認識のコンペティション DCASE2020 で世界 1 位を獲得しました. <https://engineering.linecorp.com/ja/blog/dcase2020-challenge> (2025/01/20 アクセス)
- [13] Papers With Code (2020) Griffin-Lim Algorithm. <https://x.gd/2wMl5> (2025/01/15 アクセス)