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

Make Brain Project

Project Name

Make Brain Project

グループ名

グループ B

Group Name

Group B

プロジェクト番号/Project No.

21

プロジェクトリーダ/Project Leader

狩野政宗　Masamune Karino

グループリーダ/Group Leader

柳田陸斗　Rikuto Yanagida

グループメンバ/Group Member

柳田陸斗　Rikuto Yanagida
神田雄太郎　Yutaro Kanda
鈴木隼　Shun Suzuki
森久保孔明　Kohmei Morikubo
松下文太　Bunta Matsushita

指導教員

香取勇一　栗川知己　加納剛史

Advisor

Yuichi Katori　Tomoki Kurikawa　Takeshi Kano

提出日

2025 年 1 月 21 日

Date of Submission

January 21, 2025

概要

アニメーションキャラクターの音声を聞くだけで、ヒトはそのキャラクターの顔を予想することができる。本プロジェクトの目的は、この点に着目し、その情報処理を再現できる人工知能を開発することである。また、これにより、ヒトが音声からアニメーションキャラクターの顔をどのように予想しているのかを理解することを目指す。具体的には、アニメーションキャラクターの音声と顔画像の特徴を学習するため、特徴量抽出、クラスタリング、二分マッチングを用いる。そして、入力された音声の特徴に基づき、ユークリッド距離を利用して予想されるアニメーションキャラクターの顔画像を出力する人工知能を開発する。本プロジェクトの成果により、音声を入力としてアニメーションキャラクターの顔を予想する人工知能が実現することが期待される。さらに、この成果を基に、音声からアニメーションキャラクターの顔画像を生成する人工知能の開発にもつながる可能性がある。

キーワード　アニメ, マッチング, クラスタリング, 特徴量抽出, 深層学習

（※文責: 柳田陸斗）

目次

第 1 章　はじめに

1.1　背景

ヒトはある音声を聴いたとき、その発話者の身体的特徴をある程度予想することができる。なぜなら、骨格・体格・人種などの視覚情報は、ピッチ・抑揚・各言語に固有のリズム等の発声者の身体的属性と関連があると考えられるためである。具体例として、低い声は一般的に男性を連想させることが多く、高い声は女性や子供を連想させることがある。また、別の例として、アフリカ系アメリカ人特有のイントネーションやリズムが強調された英語（AAVE: African American Vernacular English）は、黒人コミュニティに属する話者を想起させる一方で、一般的に「標準英語」とされるアクセントは白人話者を連想させることがある。このような傾向をモデル化した研究として、Speech2Face がある [9]。Speech2Face は、音声からその話者の顔を予測するモデルであり、骨格や人種、性別といった身体的特徴を音声信号に基づいて推定することを目的としている。しかしながら、この推定の過程に類似した状況として、ヒトがアニメーションキャラクターの音声を聴いたときにそのアニメーションキャラクターの顔を予想できることに着目した手法は依然として確立されていない。アニメーションキャラクターの場合、視覚的な顔の特徴は現実の人間の骨格や体格に縛られることなくデザインされるため、音声との関連性は必ずしも直感的ではない。それでも、視聴者はしばしば声のトーンや話し方からそのキャラクターの性別、年齢、さらには性格をイメージする。このような現象を体系的に研究することで、音声とキャラクターデザインの結びつきに関する新たな洞察が得られる可能性がある。

1.2　プロジェクトの目的・意義

本プロジェクトの目的は、ヒトが音声から顔画像を予想する過程を再現し、音声とキャラクターデザインの結びつきに関する新たな洞察を得ることである。特に、この過程を人間の顔ではなくアニメーションキャラクターのような実世界に実在しない存在でも同様に再現できるかを実証するという点に新規性がある。

先行研究である Speech2Face は、音声から顔の特徴を推定するモデルを開発し、音声データを入力としてその音声の持ち主の顔を予想する手法を確立した。しかし、Speech2Face は実在する人間の顔を対象としている。アニメーションキャラクターはデフォルメされており、音声と顔の結びつきが現実の人物とは異なると考えられるため、従来のモデルをそのまま適用することは困難である。これに対処するため本プロジェクトでは、大規模なアニメーションデータセットを用いて、キャラクターの音声からキャラクター画像を推定するモデルを独自に開発した。

（※文責: 柳田陸斗）

第 2 章　関連研究

2.1　先行研究

本プロジェクトテーマの決定にあたって特に参考にした先行研究として、Oh ら [1] による研究が挙げられる。概要については前章で述べた通りである。この研究では、図 2.1 のように YouTube 上の教育動画を使用して音声と顔の相関を学習し、年齢、性別、人種などの物理的な特徴を捉えた画像を生成する。画像生成にあたっては、まず人物が話している画像と音声をそれぞれ 4096 次元のベクトルに変換し、それらを誤差関数で評価するプロセスを通して Convolutional Neural Network (CNN) による自己教師あり学習を行っている。その後、Face Decoder でカノニカル（正面向き、中立表情の）顔画像を生成している。なお、画像のベクトル化については既存のモデル [2] を利用しており、Oh ら (2019）によって独自に提案されたのは音声のスペクトログラムをベクトル化するモデル部分である。

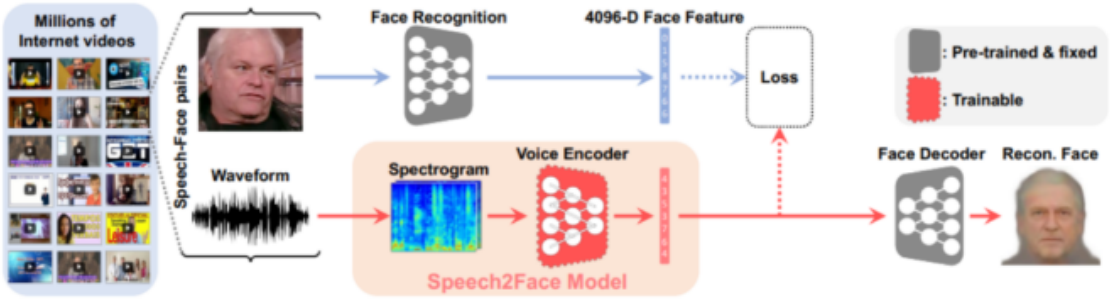

図 2.1　Speech2Face モデルとトレーニングパイプライン ([1] より引用)

2.2　関連研究

大規模なアニメーションデータセットとして、Cai（2024）らによるデータセット Anim400k[1] がある。、このデータセットは、自動吹き替え、同時翻訳、ガイド付きビデオ要約、ジャンル・スタイルの分類など、さまざまなビデオ関連のタスクの研究をサポートする目的で作られた。日本語と英語の 425,000 を超えるアニメーションビデオセグメントが含まれている。Speech2Face が YouTube 上に当時存在した大量の教育コンテンツをモデルの学習に用いている一方で、本プロジェクトではモデルの学習にこのデータセットを用いた。

2.2.1　本プロジェクトで必要なスキル

本プロジェクトで必要なスキルは主に 4 つある。1 つ目に Neural Netrwork（NN） や多層パーセプトロン (MLP) に関する基礎知識である。2 つ目に学習モデルに関する論文からモデル構築に応用できそうかを検討するスキルが必要である。特に、音声から顔画像を予測する関連研究や、アニメーションキャラクターのデータセット構築に関する研究を参考にし、どの手法を応用するかを

判断することが重要である。3 つ目に Python のコーディングスキルが必要である。今回は計算量が大きいため、Numpy や pandas、PyTorch などのライブラリを用いた高速化をするスキルも必要である。4 つ目にチーム開発をスムーズに行うため、git 等を用いたバージョン管理のスキルが必要である。

2.2.2 関連性の高い本学の専門科目

本プロジェクトと関連性の高い専門科目として「ニューロコンピューティング」「データサイエンス入門」がある。ニューロコンピューティングは、生物の神経系を模倣した人工ニューラルネットワークを用いて、学習や情報処理を行うための理論や技術を学ぶ学問分野である。本プロジェクトにおける音声と顔画像からの特徴量抽出は、まさにニューラルネットワーク、特に深層学習モデルの応用と言える。さらに、本プロジェクトで扱った特徴量を用いたクラスタリングは、「データサイエンス入門」でも学ぶデータ分析の基礎的な手法の一つである。クラスタリングにあたっては、適切なアルゴリズムの選択やパラメータチューニングが重要となるが、データサイエンス入門で学ぶ様々なクラスタリング手法（k-means、階層的クラスタリングなど）や評価指標（シルエット係数など）に関する知識が役立つと考えられる。

（※文責: 柳田陸斗）

第 3 章　プロジェクト学習の目標

3.1　本プロジェクトの目標

本プロジェクトの目標は、アニメーションキャラクターの音声を入力として、そのキャラクターの顔画像を予測する人工知能を開発することである。具体的には、音声と顔画像の特徴量を効果的に抽出し、クラスタリングおよびマッチング手法を通じて、入力された音声に最も適合する顔画像をデータセットから選出・出力することを目指す。この過程を通じて、ヒトが音声からキャラクターの顔を予測する際の情報処理過程を模倣し、その仕組みを再現することを目的とする。

さらに、この人工知能の開発を通じて、ヒトが音声から視覚情報をどのように連想しているのか、そのメカニズムを解明する手がかりを提供することが期待される。具体的には作成した人工知能が出力する予測結果を分析することで、音声のピッチや抑揚、リズムなどの音声的特徴が、顔の特徴をどのように結びつ得るかを検証し、ヒトの情報処理のメカニズムに対する新たな知見を得ることを目指している。

（※文責: 鈴木隼）

第 4 章　手法

使用したデータセットである Anim400k においてキャラクターと顔画像、キャラクターと音声の対応付けは行われていない。もし、対応付けが存在していれば、今回作成した重心を用いた計算や、Generative Adversarial Networks（GAN）[5] 、Variational Autoencoder（VAE）[7] などの手法を用いて目的のモデルを作成することができる。そのため、顔画像のクラスターと音声のクラスターの一対一対応を先に作成し、その対応を元に音声から顔画像を推論するプログラムを作成した。

4.1　顔画像のクラスターと音声のクラスターの一対一対応の作成

図 4.1 は一対一対応の作成の全体像である。

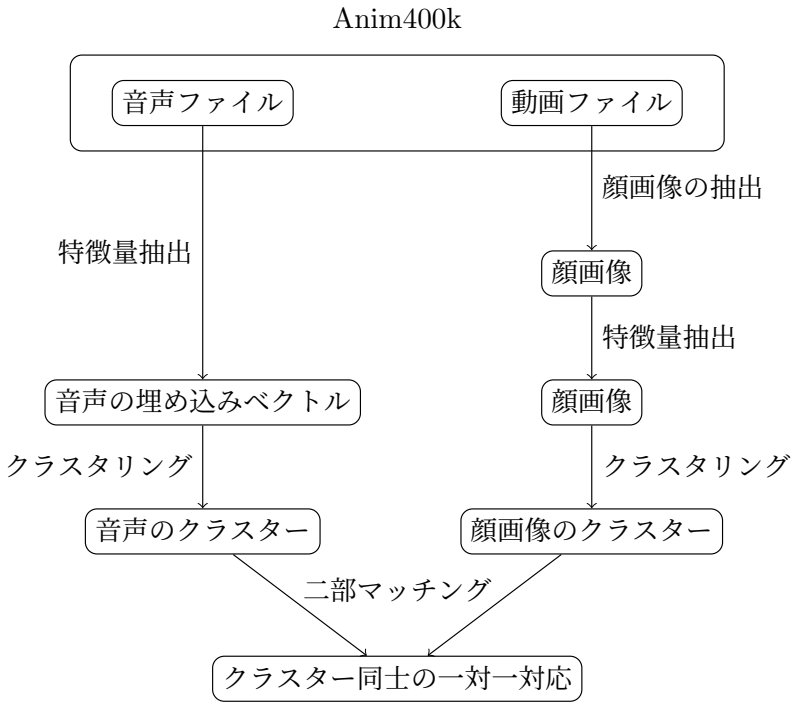

図 4.1　一対一対応作成の流れ

本節では次の順に説明する。

1. 顔画像の前処理
2. 顔画像と特徴量抽出
3. 音声の特徴量抽出
4. 顔画像と音声の特徴量のクラスタリング
5. 顔画像のクラスターと音声のクラスターの対応付け

（※文責: 森久保孔明）

4.1.1 顔画像の前処理

アニメビデオクリップから 20 フレームごとに画像を抽出した。その抽出した画像から lbpcascade_animeface [8] を用いてアニメ顔を検出し、トリミングを行った。lbpcascade_animeface とは、アニメや漫画のキャラクターの顔を検出するためのカスケード分類器ファイル (lbpcascade_animeface.xml) を提供するライブラリである。これは、OpenCV の機能を利用して、アニメ画像や動画からキャラクターの顔領域を自動的に検出することを目的としている。このライブラリは、Local Binary Pattern（LBP）特徴量とカスケード分類器を組み合わせた手法を用いており、アニメ特有の顔の特徴を捉えるように設計されている。具体的には、OpenCV の CascadeClassifier クラスを使用して、画像内の顔領域を検出する。

（※文責: 神田雄太郎）

4.1.2 顔画像の特徴量抽出

顔画像から特徴量を抽出するために、DeepFace ライブラリ [10] を使用した。DeepFace は、複数の事前学習済みモデルを利用できるフレームワークであり、その中から Facenet512 モデルを使用した。Facenet512 は、512 次元の埋め込みベクトルとして顔画像の特徴量を生成する深層学習モデルである。

具体的には、対象となる顔画像を Deepface に読み込ませることで、画像は自動的に標準化・整形され、モデルの入力形式に適合するように加工される。その後、Facenet512 モデルを用いて処理された顔画像から、512 次元の特徴ベクトルが出力される。

（※文責: 鈴木隼）

4.1.3 音声の特徴量抽出

学習済モデルである spkrec-ecapa-voxceleb[3] を用いた。このモデルに mp3 ファイルを入力することで特徴を表す 192 次元の埋め込みベクトルが出力される。このモデルを用いて Anim400k の音声データからそれに対応した埋め込みベクトルを作成した。

4.1.4 特徴量のクラスタリング

顔画像・音声のクラスターの形状が球体に近くなるように手法を選択した。

顔画像･音声のクラスタリングではともに階層的クラスタリング手法（AgglomerativeClustering）を採用した。キャラクター数が不明のため、sklearn.metrics の silhouette_score を用いて最適なクラスター数を探索し、データの形状に合わせて決定するようにした。

他には DBSCAN[4] も検討したが、分類不可な要素が出てくることや期待するクラスターの形状と合わないことから不採用とした。

4.1.5　顔画像のクラスターと音声のクラスターの対応付け

アニメのシーズンごとに顔画像のクラスターと音声のクラスターを頂点とした二部グラフとして最大マッチングを行った。顔画像のクラスター $\mathfrak{f}$ と音声のクラスター $\mathfrak{a}$ を繋ぐ辺の重み $E_{\mathfrak{f},\mathfrak{a}}$ は以下の式で定義した。

$\mathfrak{f}$: 顔画像のクラスター
$\mathfrak{a}$: 音声のクラスター
C: ビデオクリップ
C_{F}: ビデオクリップ C に登場する顔画像のクラスターの集合
C_{A}: ビデオクリップ C に登場する音声のクラスターの集合
$\mathcal{C}$: ビデオクリップ全体の集合

$$W_{\mathfrak{f},\mathfrak{a}}(C) \coloneqq \begin{cases} \dfrac{1}{|C_{\mathrm{F}}| \cdot |C_{\mathrm{A}}|} & \text{if } \mathfrak{f} \in C_{\mathrm{F}} \text{ and } \mathfrak{a} \in C_{\mathrm{A}} \\ 0 & \text{else} \end{cases}$$

$$E_{\mathfrak{f},\mathfrak{a}} = \sum_{C \in \mathcal{C}} W_{\mathfrak{f},\mathfrak{a}}(C)$$

上記のグラフに対し、scipy.optimize の linear_sum_assignment を使用して、顔画像のクラスターと音声のクラスターのマッチングを行った。

これにより、顔画像のクラスターと音声のクラスターの一対一対応を作成できた。

4.2　顔画像の推論プログラム

推論プログラム内では、図 4.2 の流れに沿って処理を行う。

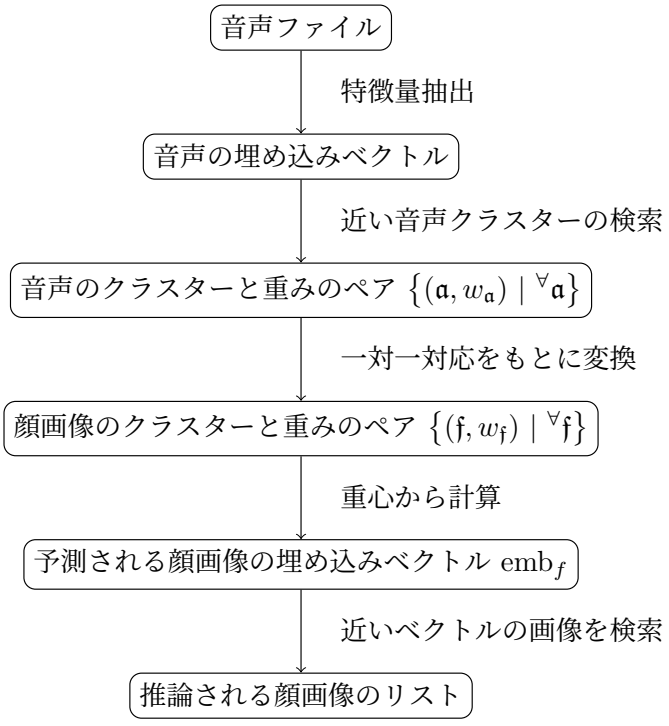

図 4.2　推論の流れ

これは音声の特徴量から、節 4.1.4 のクラスタリングの結果と節 4.1.5 の一対一対応を用いて、顔画像の埋め込みベクトルを予測し、それによって近い画像を検索するという流れである。

本節では次の順に説明する。

1. 音声の特徴量抽出
2. 音声のクラスターと重みのペアの作成
3. 顔画像のクラスターと重みのペアへの変換
4. 顔画像の埋め込みベクトルの予測
5. 予測した埋め込みベクトルに近い埋め込みベクトルを持つ顔画像の検索

4.2.1 音声の特徴量抽出

音声の特徴量抽出については節 4.1.3 と同じ手法を用いる。spkrec-ecapa-voxceleb によって入力された音声データから 192 次元の埋め込みベクトルを作成する。

4.2.2 音声のクラスターと重みのペアの作成

音声の埋め込みベクトルから音声のクラスター $\mathfrak{a}$ と重み $w_{\mathfrak{a}}$ のペア $(\mathfrak{a}, w_{\mathfrak{a}})$ の集合に以下の方法で変換する。音声のクラスター $\mathfrak{a}$ の重心 $c_{\mathfrak{a}}$ と節 4.2.1 で抽出した埋め込みベクトルとの距離を $d_{\mathfrak{a}} \in \mathbb{R}$ として、重み $w_{\mathfrak{a}} = d_{\mathfrak{a}}^{-1}$ とする。そして、すべての音声のクラスター $\mathfrak{a}$ について、同じように処理をし、$(\mathfrak{a}, w_{\mathfrak{a}})$ の集合を作成する。最後に、$w_{\mathfrak{a}}$ の降順に並べ、上位 20 個に絞り込む。この重みの計算式と絞り込む数については改善の余地があり、今回は重みに距離の逆数を、個数に 20 を選んだ。

4.2.3 顔画像のクラスターと重みのペアへの変換

音声のクラスターと重みのペア $(\mathfrak{a}, w_{\mathfrak{a}})$ を顔画像のクラスターと重みのペア $(\mathfrak{f}, w_{\mathfrak{f}})$ に変換する。具体的には、節 4.1 で作成した一対一対応によって、音声のクラスター $\mathfrak{a}$ を対応する顔画像のクラスター $\mathfrak{f}$ に変換する。重み $w_{\mathfrak{a}}, w_{\mathfrak{f}}$ については、ラベルを $\mathfrak{a}$ から対応する $\mathfrak{f}$ に変えているだけで、値は変わらない。

4.2.4 顔画像の埋め込みベクトルの予測

顔画像のクラスター $\mathfrak{f}$ について、そのクラスターの重心を $c_{\mathfrak{f}}$ とする。この $c_{\mathfrak{f}}$ は節 4.1.2 で抽出した埋め込みベクトルと同じ 512 次元のベクトルである。

$c_{\mathfrak{f}}$ に対して、$(\mathfrak{f}, w_{\mathfrak{f}})$ のペアの集合を用いて重み付き平均を取り、それを予測される顔画像の埋め込みベクトル emb_f とする。つまり、emb_f を次の式によって計算する。この emb_f も 512 次元のベクトルである。

$$\mathrm{emb}_f = \frac{\sum_{\mathfrak{f}} c_{\mathfrak{f}} w_{\mathfrak{f}}}{\sum_{\mathfrak{f}} w_{\mathfrak{f}}}$$

4.2.5　予測した埋め込みベクトルに近い埋め込みベクトルを持つ顔画像の検索

節 4.1.2 で抽出した顔画像の埋め込みベクトルと節 4.2.4 で予測した埋め込みベクトルの次元は同じため、予測したベクトルに近い埋め込みベクトルの画像を検索する。

本プロジェクトで作成したプログラムでは、Python の Faiss ライブラリを用いて emb_f に近い埋め込みベクトルを持つ画像を検索する。

（※文責: 森久保孔明）

第 5 章　結果

音声と顔画像の特徴量抽出およびクラスタリングを行い、それぞれのクラスタ間の対応付けを最大マッチング手法を用いて実施した。結果については、主に目視による定性的な評価を行った。以下に出力例を示した。入力音声は効果音ラボ [6] の音声素材の中から、穏やかな口調の男性ボイスと語気の強い男性ボイスを用いた。ファイル名の詳細については付録に記載した。前者の音声を入力した場合、図 5.1 のように丸みを帯びたキャラクターデザインの顔画像が出力される傾向が確認された。後者の音声を入力した場合、鋭い目元や顎の特徴を持つキャラクターデザインの顔画像が多く出力されることが確認された。

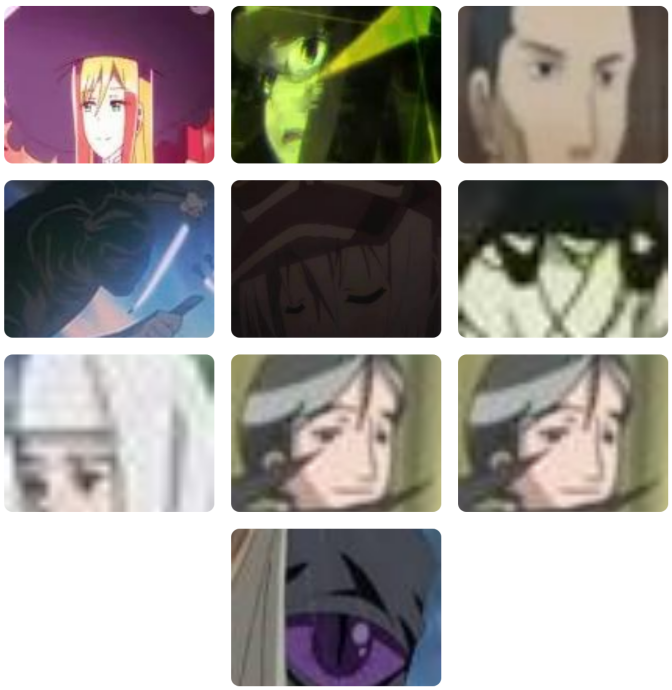

図 5.1　穏やかな口調の男性ボイスを入力したとき

また、一つのアニメシーズンの音声を例に、顔画像クラスタおよび音声クラスタの分布を、主成分分析を用いて 3 次元にプロットして可視化を行った。クラスタリング結果の可視化の一例を図 5.3 に示した。使用したアニメシーズンは「Sword Art Online Alicization War of Underworld」、使用した手法はウォード法で、クラスタ間の距離をユークリッド距離で計算した。クラスタ数は 5 に設定した。

（※文責: 松下文太）

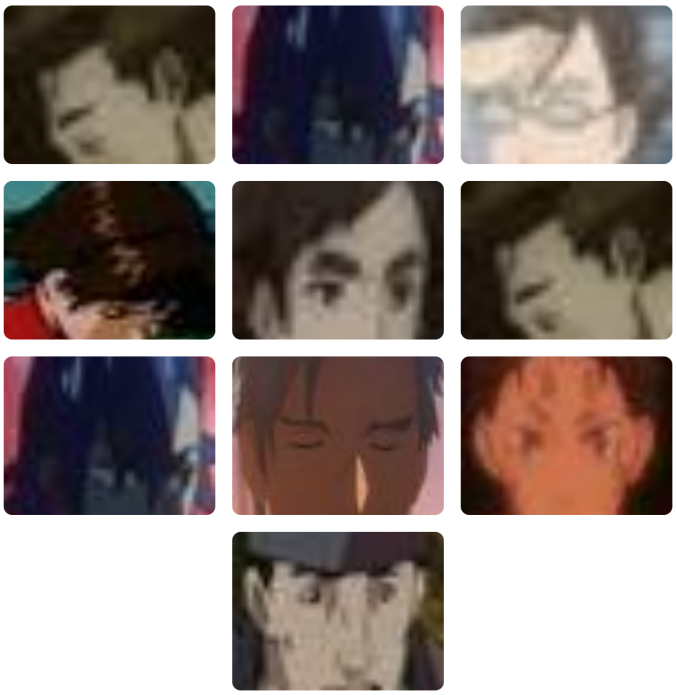

図 5.2　語気の強い男性ボイスを入力したとき

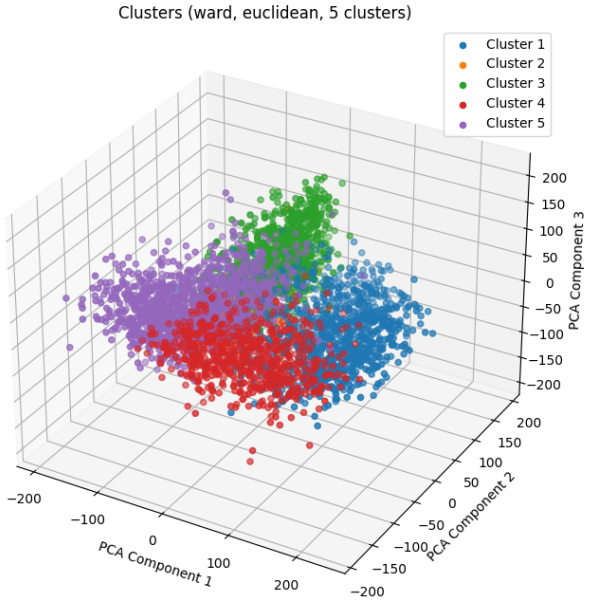

図 5.3　1 つのアニメシーズンに対する階層的クラスタリング結果

第 6 章　考察

6.1　考察

このプロジェクトでは、音声の特徴量とアニメーションキャラクターの顔画像との間に潜在的な関係性があるのかを検証した。プロジェクトを通じて、音声の性別や声の定性的な高さがキャラクターデザインに反映される傾向が確認できた。これは、人間が音声から視覚情報を想起する際に、音声と顔の相関性を無意識に利用している可能性を示唆している。

このプロジェクトにおけるクラスタリング・マッチング結果の評価は、時間上の制約から目視による定性的なものに留まるに至ったが、理想的には定量的な評価手法を取り入れる必要がある。例えば、別クリップで発話している同一人物が、同一のクラスターに分類されているかどうかをスコアとして採用すること等が挙げられる。

このプロジェクトでは、データセットのアノテーションファイルから話者が単一のものであるとわかる音声・動画クリップを使用し、話者と動画のフレームに映っているキャラクターが同一のものであるという仮定のもとクラスタリングを行った。そのため、フレームに映っているキャラクターと実際の話者に相違がある場合に、クラスタリングのノイズになっていた可能性がある。また、定性的には妥当とはいえない顔画像が出力されることがある。これらの問題点を解決するためには、アノテーションの精度向上、データセットの精査、そしてモデルの改良といった多角的なアプローチを検討する必要がある。まず、アノテーションの精緻化として、現在の「話者単一」というアノテーションに加え、「フレーム内に映るキャラクターと音声データの話者が一致する」という新たなラベルを付与することが考えられる。これにより、キャラクターと話者が一致しないデータを除外もしくは別途扱うことで、クラスタリングの精度を向上させることができる。さらに、クラウドソーシングなどを活用し、複数人によるアノテーションを行い、その一致率を評価することでアノテーションの信頼性を高めることも有効である。次に、データセットの拡張とフィルタリングを行うことも考えられる。現在のデータセットに加え、キャラクターと話者が確実に一致するデータセットを新たに収集もしくは作成する必要がある。また、既存のデータセットに対しては、音声と顔画像の特徴量の相関を分析し、閾値を設定することで、不適切なデータをフィルタリングすることも挙げられる。例えば、音声から推定される性別と顔画像から推定される性別が一致しないデータなどを排除することで、データの質を自動的な処理によって向上させることができる。

6.2　大学カリキュラムとの関連性

本プロジェクトは、「ニューロコンピューティング」「データサイエンス入門」等の授業で学んだ知識を応用する実践的な機会となった。データ分析の手法、Python を用いたプログラミング、Git 等のバージョン管理ツールの利用など、様々なスキルを実践的に活用し、関連する科目への理解を深めることができた。特に scikit-learn を用いたクラスタリングや既存の機械学習モデルの使用においては、機械学習の基本的なアルゴリズムや手法についての理解が、モデルの選定や実装に貢献した。

6.3　新たなプロジェクト学習のテーマ

このプロジェクトの拡張として、生成タスクへの応用が挙げられる。クラスタリングではなく、顔画像から音声特徴量を生成、あるいは音声から顔画像を生成するタスクに挑戦することで、より自然でリアルな結果を期待することができる。顔と声の両方の特徴量の誤差を用いた深層学習モデルは、この方向性における有効なアプローチとなると考えられる。

（※文責: 神田雄太郎）

付録 A　画像の出力例に使用した音声ファイル

ファイル名	カテゴリ	備考
「よろしく頼む」.mp3	声素材/ゲームの戦闘	穏やかな男性ボイス
「受けてみろ！奥義！」.mp3	声素材/ゲームの戦闘	語気の強い男性ボイス

表 A.1　画像の出力例で使用した音声ファイルの一覧

参考文献

[1] Kevin Cai, Chonghua Liu, and David M Chan. Anim-400k: A large-scale dataset for automated end to end dubbing of video. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11796–11800. IEEE, 2024.

[2] Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman. Synthesizing normalized faces from facial identity features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3703–3712, 2017.

[3] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In Helen Meng, Bo Xu, and Thomas Fang Zheng, editors, *Interspeech 2020*, pp. 3830–3834. ISCA, 2020.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, Vol. 96, pp. 226–231, 1996.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, Vol. 63, No. 11, pp. 139–144, 2020.

[6] Killy. 効果音ラボ. `https://soundeffect-lab.info/`, 2013. 閲覧日 2025 年 1 月 21 日.

[7] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[8] Nagadomi. lbpcascade_animeface. `https://github.com/nagadomi/lbpcascade_animeface`, 2018. 閲覧日 2025 年 1 月 21 日.

[9] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T. Freeman, Michael Rubinstein, and Wojciech Matusik. Speech2face: Learning the face behind a voice. *CoRR*, Vol. abs/1905.09773, , 2019.

[10] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pp. 1–4. IEEE, 2021.