B-02

2 次元キャラクタにおける音声生成モデルの検討

斉藤　彰吾†　　大井 翔‡　　佐野 睦夫‡
Saitou Syougo　　Sho Ooi　　Mutsuo Sano

1. はじめに

近年，電子書籍の市場規模は順調に伸びてきており，電子書籍では何かをしながら本を読んだりできるように，音声読み上げ機能があるものもある．しかし，音声読み上げのクオリティは低く，声優が音声を吹き込むオーディオコミックはコストがかかる．

人間の顔からおおよその声を人間は想像する事が可能だと考えられており，人の顔と声の関係性について調査している研究がある[1]．本研究の先行研究としてキャラクタの顔画像から年齢や性別の推定を行わずに，キャラクタの画像特徴量のみで音声を推定・生成するシステムを提案している[2]．しかし，この研究では音声を用いずキャラクタの顔画像のみを提示し，その顔画像から声の性別や声の年齢等を想像しアンケートしたものを元として音声学習データの主軸としていた．

そこで，本研究では，キャラクタにあった音声特徴量と画像特徴量の対応付けをするために，キャラクタに違和感のない音声の傾向を分析することを考えた．そこで，1 枚のキャラクタのイラストに対して，複数の音声を提示し，実験参加者に評価してもらう．これにより，人間の人間の知覚としてキャラクタにあった声の特徴を感じる傾向分析を行う．

2. 関連研究

キャラクタの顔画像のパーツに注目し人がキャラクタのどのパーツを見て声を想像しているかのメカニズムを用いる事で，よりシンプルなキャラクタの画像特徴から音声推定が可能かを検討した研究が行われている[2]．この研究では，初めに人がキャラクタの顔画像のイラストを見た際に，キャラクタの顔のどのパーツを見て声を想像しているのか調査を行っている．その結果，人は「目の形」「髪の形」「髪の色」が有力な特徴である事が判明した．この研究では図 1 に示すようにその三点の中でも特に有力な特徴である「目の形」を抽出している．画像推定から音声推定を行いキャラクタの画像から人の感性が想像する声に近い音声が生成可能か調査を行った[2].

また，他の研究ではキャラクタの声優のキャスティングに対して調査している研究もあり，キャラクタの声優が担当したいくつかの音声データを収集し，得られた音声データを元にして音響特徴量を算出する．その後得られた音響特徴量と印象値の関係を学習させ，活用する事で新しいキャラクタに対しても印象値与える事で図 2 に示すように適切な音響特徴量を推定するといった学習から音声生成を行った研究もあるが，音声生成は実際の人の音声で生成されているため，いわゆるアニメ声といった声の生成が難しいと考えられる．本研究ではアニメ等のキャラクタを演じられている方の音声を用いて音声生成したものを用いていくためよりキャラクタイラストから生成される音声としての精度が高いものの作成を狙う[3].

図 1　目の抽出方法

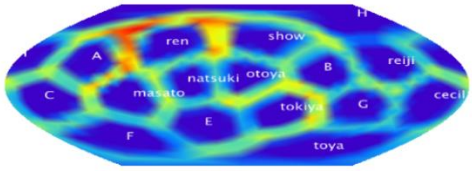

図 2　音響特徴量と印象値の学習結果

3. 提案手法

本研究の目的はキャラクタの画像に対して音声推定や生成する研究をより精度が高いものにする事を目指す．そのため，本稿ではキャラクタイラストと声の関係性を調査する事を目的とする．従来研究[2]ではキャラクタの顔画像の中で特に目の形が声を想像する際に有力だと判明しており，目の形の特徴量を用いて音声推定，音声学習を行われている．しかし，それらはキャラクタイラストのみを提示しその画像を見てどういった印象を持つかを調査し得られたもので，実際の音声を用いて調査が行われたものではない．そのため，本研究ではキャラクタイラストと音声生成を行った音声を一対比較法により重複しないランダムな組み合わせを提示し，評価を頂くことでより精度の高い音声学習の特徴モデルの生成を試みる．

本研究でキャラクタイラストと声の関係性を調査する流れは以下の図のとおりである．

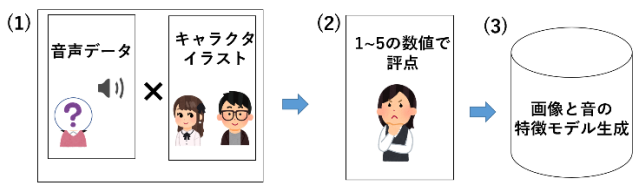

図 3　提案手法の流れ

†大阪工業大学大学院，Graduate School of Osaka Institute of Technology

‡ 大阪工業大学，Osaka Institute of Technology

(1) 自動生成された音声データとキャラクタイラストを一対比較法を用いて提示する．また，この際一度提示した組み合わせは再度表示されないようにする．
(2) 1~5 の数値でキャラクタと音声がどれだけ合致していると感じるのか実験参加者に評点を行ってもらう．
(3) 得られた評点データから画像と音の特徴モデルの生成を狙う．

この時，(2)の評点は低い点数である程キャラクタと音声の組み合わせが一致していないといった評価になり，高い点数程キャラクタと音声の組み合わせが一致していると感じているとする．続いて，実験に用いるデータに関しての説明を行う，用いたキャラクタイラストは以下の図のものを用いた．また，キャラクタイラストに区別を行うために今回は左上のキャラクタから右に向かって male1，male2，male3，male4 とし，左下のキャラクタから右に向かって female1，female2，female3，female4．と名称付ける．

図 4　使用したキャラクタイラスト
cre8tiveAI (https://cre8tiveai.com/sc)

続いて，音声は 8 種類用いて実験を行い先行研究[2]のシステムを用いて音声生成された音声を用いた．音声の内容は発話内容によるアンケート結果の差異を生み出さないため「こんにちは」の音声のみを用いて行った．

4. 実験

本研究の実験は，一対比較法を用いたキャラクタイラストと音声の提示を行い，それらを実験参加者にどれ程一致しているのかを直感的に評価してもらう事で，人間の知覚としてキャラクタイラストにあった声の特徴を感じる傾向分析が可能だと考える．また，得られた傾向から今後の画像と音の特徴モデルの生成を行っていく．

本実験で用いた音声は，声優の「田村ゆかり」「松風雅也」「沢城みゆき」「鈴村健一」「花澤香菜」「逢坂良太」「平田広明」「小野友樹」の音声を Tacotoron2 を用いて学習させ Waveglow にて音声生成を行ったものを用いる．実験ではキャラクタイラスト 8 種類と音声 8 種類の組み合わせを重複無しでランダムに提示したため 64 種類の組み合わせで提示を行った．

本実験を行う際に提示用プログラムを制作した．実験では「押下するとランダム画像表示」のボタンを押下し提示されたキャラクタイラストと音声の組み合わせを評価する．その後，1~5 のラジオボタンを選択し選択後に押下する「評価値選んでね」のボタンを押すことで評価値が決定される仕様とした．また本研究では，評点値 1 を選んだ時キャラクタ画像と音声が一致していると全然感じない，2 の時は感じない，3 の時はどちらでもない，4 の時は感じる，5 の時はとても感じるとした．

以下の図に実験で用いた提示する際に用いた実験用プログラム画面の表示を行う(図 5).

図 5　実験用プログラムの画面

本実験は，21~24 歳の男子大学生・大学院生 7 名を対象に行った，実験では評点に加え，「どういった基準でその評価を選定したか，教えてください」の質問をアンケート終了後に行った．

実験結果を以下の図 6 から図 13 に提示する．実験結果はそれぞれのキャラクタイラストごとに得られた 1~5 の評価値の割合を示したものとなる．

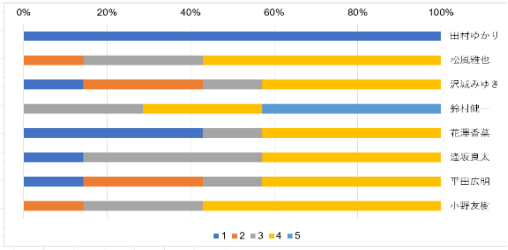

図 6　male1 の実験結果

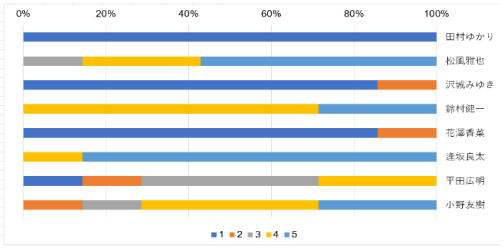

図 7　male2 の実験結果

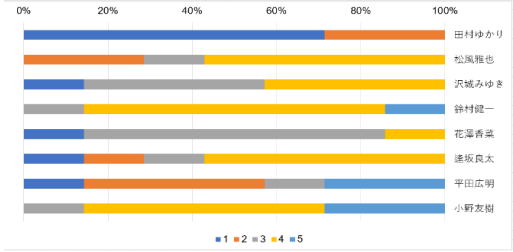

図 8　male3 の実験結果

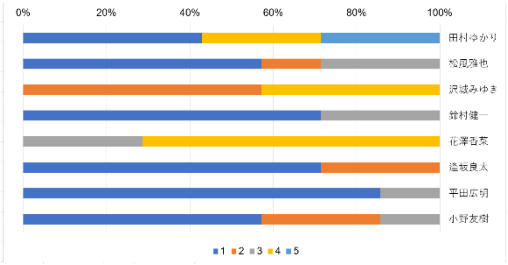

図 9　man4 の実験結果

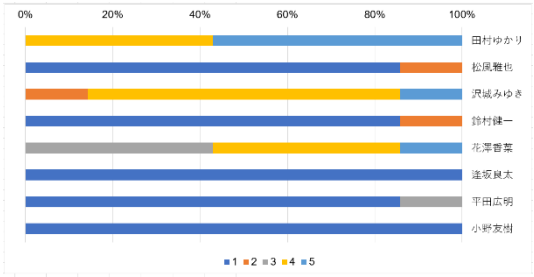

図 10　female1 の実験結果

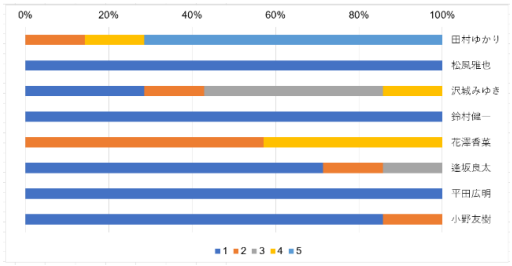

図 11　female2 の実験結果

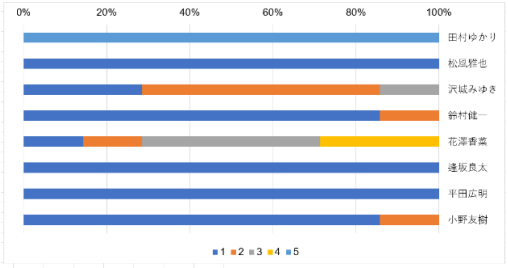

図 12　female3 の実験結果

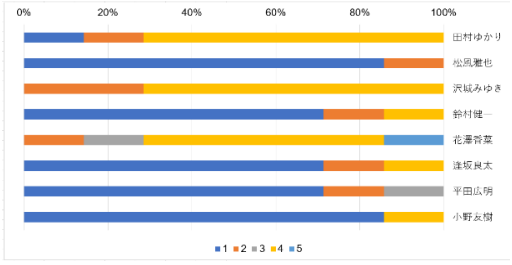

図 13　female4 の実験結果

実験結果の考察を行うと図 6 の時 male1 の画像に対して松風雅也さんと小野友樹さんの音声を用いて音声生成を行った音声が顔画像と一致していると見受けられる．逆に田村ゆかりさんの音声を用いて音声生成を行った音声とはあまり一致していないと見受けられる．続いて，図 7 の時 male2 の画像に対しては逢坂良太さんの音声を用いて音声生成を行った音声が顔画像と特に一致していると見受けられる．逆に田村ゆかりさんの音声を用いて音声生成を行った音声とは一致していないと見受けられる．続いて，図 8 の時 male3 の画像に対して平田広明さんと小野友樹さんの音声を用いて音声生成を行った音声が顔画像と一致していると見受けられる．逆に田村ゆかりさんの音声を用いて音声生成を行った音声とは一致していないと見受けられる．続いて，図 9 の時 male4 の画像に対して逢坂良太さんの音声を用いて音声生成を行った音声が顔画像と特に一致していると見受けられる．逆に田村ゆかりさんの音声を用いて音声生成を行った音声とは特に一致していないと見受けられる．続いて，図 10 の時 female1 の画像に対して田村ゆかりさんと沢城みゆきさんの音声を用いて音声生成を行った音声が顔画像と一致していると見受けられる．逆に逢坂良太さんと小野友樹さんの音声を用いて音声生成を行った音声とは特に一致していないと見受けられる．続いて，図 11 の時 female2 の画像に対して田村ゆかりさんと花澤香菜さんの音声を用いて音声生成を行った音声が顔画像と一致していると見受けられる．逆に平田広明さんの音声を用いて音声生成を行った音声とは特に一致していないと見受けられる．続いて，図 12 の時 female3 の画像に対して田村ゆかりさんの音声を用いて音声生成を行った音声が顔画像と特に一致していると見受けられる．逆に逢坂良太さんと平田広明さんの音声を用いて音声生成を行った音声とは特に一致していないと見受けられる．続いて，図 13 の時 female4 の画像に対して花澤香菜さんと沢城みゆきさんの音声を用いて音声生成を行った音声が顔画像と一致していると見受けられる．逆に小野友樹さんの音声を用いて音声生成を行った音声とは一致していないと見受けられる．

これらの結果から考察を行うと，男性キャラクタを用いた傾向調査では小野友樹さんと逢坂良太さんの音声を用いて音声生成を行った音声は人間が感じる傾向として適しているといった傾向が見受けられ，田村ゆかりさんの音声を元にして音声生成を行った音声は男性キャラクタの音声には適していない傾向が見受けられた．続いて女性キャラクタを用いた傾向調査では田村ゆかりさん，沢城みゆきさん，花澤香菜さんの 3 名の音声を元として生成を行った音声が良い結果が出たが，その中でも田村ゆかりさんの音声を元とした音声が傾向として良い結果が現れた．また，女性キャラクタの画像と小野友樹さんの音声は適していないといった傾向が見受けられた．

また，実験終了時に実験参加者に評価基準の選定方法についてのアンケートを行った．アンケートで得られた回答として「キャラ画像を見て抱いたイメージに近いかどうかで判断した」「声質が性別に合っているかどうか，またその声が顔からイメージできる顔かどうか」などといったものが得られた．また，他のアンケート結果も総合して多く見受けられたのがキャラクタを顔から声をイメージしたり雰囲気で感じ取って判断している部分が多いと考えられた．また，性別を判断して声をある程度想像しているといった意見も散見され，総合的にキャラクタの外見のイメージから人間は声質を決定している面があると考えられる．

5. まとめ

本研究では，複数枚のキャラクタのイラストに対して複数の音声と組み合わせたものを提示した，これによって人間の知覚としてキャラクタにあった声の特徴を感じる傾向分析を行った．本実験では 7 名の実験参加者に実験を行い 8 種類のキャラクタと音声の組み合わせによって実験を行い，人間の知覚としての特徴データが得られた．しかし，実験を行うにあたって傾向分析するデータ量が少なく十分な調査結果が得られなかったという点が考えられる．また，実験終了時に評価基準の選定方法についての調査を行った結果キャラクタの雰囲気やイメージ，性別で声を選定した声が多いという結果が得られた．今後は画像の枚数や音声データの数を増やし特徴モデルをより深めた上でキャラクタの顔画像から音声学習・生成を行う事を目指す．

参考文献

[1] Smith，Harriet MJ,etal，“Concordant cues in faces and voices: Testing the backup signal hypothesis，” Evolutionary Psychology 14.1 (2016): 1474704916630317.

[2] 大道昇，大井翔，佐野睦夫，“オーディオブックス自動生成のための 2 次元キャラクタ特徴と声の関係性の調査，”情報処理学会関西支部，支部大会，2021.

[3] 酒井えりか，伊藤彰教，伊藤貴之．“ゲームキャラクタと声質の傾向分析，”(可視化，キャラクタアニメーション，映像表現・芸術科学フォーラム 2016). " 映像情報メディア学会技術報告 40．11．一般社団法人 映像情報メディア学会，(2016).

[4] 大道昇，大井翔，佐野睦夫．“オーディオブックス自動生成のための 2 次元キャラクタ特徴と声の関係性の調査，” 情報処理学会 インタラクション 2021，(2021).

[5] 大杉 康仁，齋藤 大輔，峯松 信明，“Eigenvoice と CLNF を用いた顔から声への統計的対応付けの検討，”研究報告音声言語情報処理 (SLP) 2017．3 (2017): 1-6.

[6] 一般社団法人 電子出版政策・流通協議会 "平成 30 年度 電子書籍等の情報アクセシビリティの現状等に関する調査研究 報告" ↓

[7] https://www．soumu．go．jp/main_content/000637255．pdf (参照 2022-06-18)

[8] 後藤 駿介，大西 弘太郎，齋藤 佑樹，橘 健太郎，森 紘一郎，“顔画像から予測される埋め込みベクトルを用いた複数話者音声合成，” 日本音響学会 2020 年春季研究発表会 講演論文集，2-Q-49，pp．1141--1144，2020 年 3 月．

[9] 大杉 康仁，齋藤 大輔，峯松 信明，“Eigenvoice と CLNF を用いた顔から声への統計的対応付けの検討，”研究報告音声言語情報処理 (SLP) 2017．3 (2017): 1-6.

[10] Wang,Yujia, et al，“Comic-guided speech synthesis，” ACMTransactions on Graphics (TOG) 38．6 (2019): 1-14