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

大道昇[†1]　大井翔[†2]　佐野睦夫[†1]

概要：漫画や小説を読んでいる際に，キャラクタのセリフを音声として聞いてみたいことがある．そこで本研究では，最近の漫画の 8 割以上が電子書籍化されており，オーディオブックも増加していることに注目し，電子書籍の付加価値として，キャラクタのイラストからキャラクタに合った音声を生成することで，電子書籍にキャラクタに合った声をユーザが任意で再生できるシステムを検討する．今回の実験では，人がキャラクタのイラストを見た際にどのようにしてキャラクタに合った声を当てはめているかアンケートで調査することで，人が頭の中で声を当てはめる要因となる特徴を調べた．結果として，「目の形」，「髪の形」などの特徴から声を当てはめていることが分かった．

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, 2017 年には電子コミックの推定販売金額が, 紙のコミックス（単行本）を初めて上回るなど電子書籍の普及が発展しており，2014 年にはコミックの電子書籍化率が 8 割を超えている[1]．電子書籍はスマホやタブレットを使用して読まれることが多い．漫画や小説を読んでいる際に，キャラクタの声を聴きたい場合がある．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，音声読み上げでは，漫画などの「キャラクタごとのセリフの声が統一される」・「セリフを読み上げられない」などによって違和感を覚えてしまう．また，声優やナレータが本を朗読したオーディオブックというものも一部には存在するが，コストがかかってしまう．

関連研究として，音声インターフェースの自然な会話を実現するために，どのような顔が音声インターフェースに適しているかを定量評価しようとした研究や[2]，人の声から顔をまたは顔から声をある程度想像することができることから，その関係性について調査している研究もある[3]．また，機械学習を用いて顔と音声の関係性を学習し，顔画像から推定される埋め込みベクトルを用いた DNN 複数話者音声合成モデルの開発や[4]，デジタル化された漫画を入力した時，視覚的な印象と一致するスピーチを合成する研究がある[5]．

本研究では，以前我々が行っていた研究と同様に[6]，市場規模が伸びてきている電子書籍の付加価値として，図 1 に示すような漫画のキャラクタのイラストや小説などの挿入イラストからキャラクタに合った声を生成し，ユーザが自由にキャラクタの音声を聞くことができるシステムを検討する．

本研究ではこれまでに，アニメなどのキャラクタのイラストと音声がセットで得られるデータを用いて，キャラクタのイラストとそのキャラクタを担当する声優を対応付けて学習を行い，テスト用のキャラクタのイラストを用いてどの声優に近いか検証した[6]．3 人の声優が担当するキャラクタの画像を収集し，各声優クラスに 5 キャラクタ 250 枚のイラストを用いて実験を行った．学習に用いた声優が担当する未学習のキャラクタのイラストを分類したところ，5 キャラクタ中 2 キャラクタのみしか正しく分類することはできなかった．イラストを学習する際にキャラクタを顔だけになるよう切り抜いた画像を用いて学習を行ったが，キャラクタの顔全体から声を推定していたため，キャラクタの顔のパーツごとに学習を行うことによって分類精度が向上するのではないかと考えた．

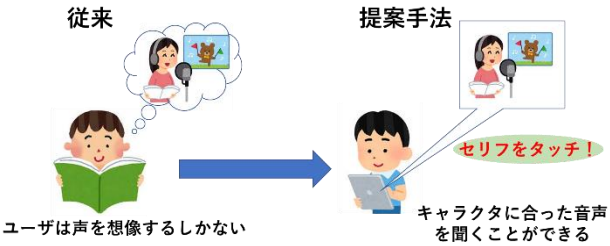

図 1　提案システムの概要図[6]

そのため，本研究ではキャラクタの声を推定するために必要な顔のパーツを調べる必要があると考える．そこで，アニメなどによる音声がついていないキャラクタのイラストを用いて，キャラクタのどのパーツを見てどのような声をすると思ったかアンケートを行うこととした．

2. 関連研究

2.1 顔と声の関連性に基づいた研究

人の声と顔には何らかの関係性があるという研究から

†1 大阪工業大学，情報科学研究科
†2 立命館大学，情報理工学部

[3], 顔と声の関連性に基づいた研究が行われている．Oh らの研究では[7]，声から性別や年齢・人種といった情報が判別できることから，人の声と話し方から話者の声を予想して画像を生成する AI を開発されている．ムービーから話者の年齢・性別・人種・話し方と声の関係性について学習を行うことで，顔の画像を予想して生成する手法が使われている．

後藤らの研究では[4], Speech Encoder, Multi-Speaker TTS, Face Encoder の 3 つのモジュールから構成されるモデルを使用し，ある顔画像を入力としたとき，生成される音声とその人物の顔画像がどれだけ適合しているかの主観評価と生成された音声が自然に聞こえるかの評価を行っている．

2.2 声優の声に関する研究

現在の日本では多くのアニメやゲームが製作され，数多くのキャラクタも生み出されているため，酒井らの研究では[8]，キャラクタに合った声優を自動で推薦するシステムの開発を検討している．結果として，キャラクタの印象値から適切な音響特徴量を推定し，テキストデータを用いた主成分分析によって声優をキャラクタに割り当てる際の議論に活用することが可能となった．

林らの研究では[9]，音声から話者の情報を知覚する過程を明らかにするには，話者情報と個人性情報を切り分けて調べる必要があるとし，声優のキャラクタ演技音声に着目し，話者情報と個人性情報の切り分けを試みた．結果として，声優の演技音声がキャラクタによって様々な年齢・性別を持って知覚されていることを示す結果が得られた．

3. 提案手法

本研究では，電子書籍の付加価値として漫画のキャラクタのイラストや小説などの挿入イラストからキャラクタに合った声を生成し，ユーザが自由にキャラクタの音声を聞くことができるシステムを検討する．図 1 に示すように，従来の紙の漫画や小説だと，キャラクタの声を想像するか，アニメ化など映像化されるのを待つことが多かった．他にもオーディオブックと呼ばれるナレータや声優が朗読した本を聴くことができる機能がある本も存在する．しかし，声優やナレータによってオーディオブックを作製するには，時間とコストが必要となってくる．本システムの構想として，キャラクタのイラストに合った音声を自動生成することによって，ユーザが電子書籍のセリフをタッチすると，セリフに組み込まれたキャラクタに合った音声を出力されるものである．本システムによって，ユーザは自由に音声を聞くことができ，電子書籍のさらなる価値の向上に貢献できるのではないかと考える．

本研究の最終目標は，漫画や小説のテキストデータとキャラクタごとのイラストデータから，キャラクタに合った音声を生成し，セリフごとにユーザが自由にセリフごとの音声を再生できるようなシステムの構築である．

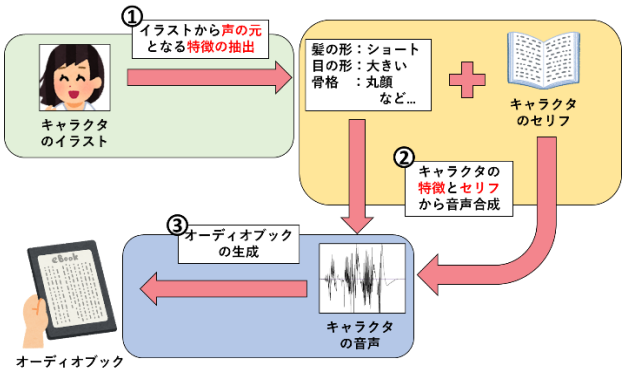

図 2　システム全体図

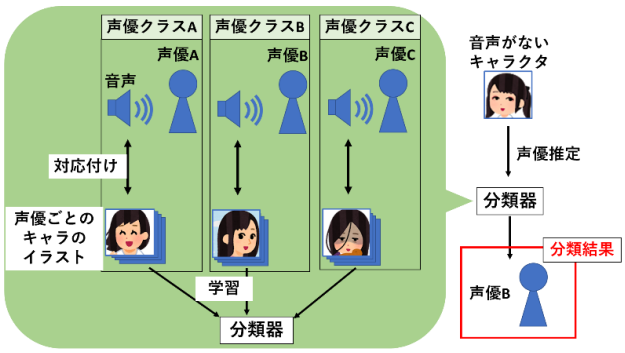

図 3　キャラクタのイラストから声優推定の流れ図

(1) キャラクタのイラストに合った音声を出力できるように，イラストと音声の対応関係を学習し，未知のイラストからそのキャラクタに合った音声特徴を取得できるようにする．
(2) キャラクタのセリフとイラストから得られた音声特徴の組み合わせを用いて音声を生成する．
(3) セリフの内容からキャラクタの感情を読み取り，生成するセリフの音声に感情を反映させる．
(4) ユーザの好みによって再生される音声のパラメータを調整できるようにする．

本システムの簡単な全体図を図 2 に示す．最終目標ではセリフから感情を読み取る・ユーザの好みを反映させることも含めているが, まずは図 2 のシステムの構築を目指す．システムの流れは，①イラストから声のもととなる特徴の抽出，②キャラクタの特徴とセリフから音声合成，③オーディオブックの生成，である．我々が行ってきたこれまでの研究では[6]，図 2 における①イラストから声のもととなる特徴の抽出を行おうとした．アニメなどのキャラクタのイラストと音声がセットで得られるデータを用いて，図 3 に示すようにキャラクタのイラストとそのキャラクタを担

図 4　アンケートに使用したキャラクタのイラスト
イラストの利用許諾についてはページ末尾の
脚注1,2参照

当する声優を対応付けて学習を行い，テスト用のキャラクタのイラストを用いてどの声優に近いか検証していた．しかし，実験を行ったところ，分類精度は非常に悪い結果となった．そこで本研究では，キャラクタのイラストにおける顔のパーツのうち，どのパーツがキャラクタの声を推定するために必要なのかを厳選し，各々で学習する必要があると考えた．今回は，キャラクタのイラストについて顔のどのパーツからどのような声を想像したかのアンケートを行うことで，キャラクタの声を推定するために必要なパーツが何であるか検証する．

4. 実験

4.1 概要

本研究の今回の実験は，キャラクタのイラストについて顔のどのパーツからどのような声を想像したかのアンケートを行うことである．人が声を想像するうえで影響を与えているパーツをアンケートから求めることで，人が声を想像するメカニズムを再現し，キャラクタのイラストからキャラクタの声を分類する際の精度向上に期待する．

4.2 アンケート

アンケートに使用した画像を図 4 に示す．アンケートではイラストの雰囲気による違いが生まれるか検証するために，PC で作成したようなイラスト，アニメ風のイラスト，漫画風のイラストをそれぞれ男女 3 キャラクタずつ使用した．使用した画像はすべて 2 次利用可能な画像である．1 行目のイラストは cre8tiveAI の SAI によって作成したイラストである[10]．2 行目のイラストはアニメ「こうしす！」公式サイト[1]から引用した画像である．また，3 行目のイラストは漫画「ブラックジャックによろしく」[2]から引用した画像である．

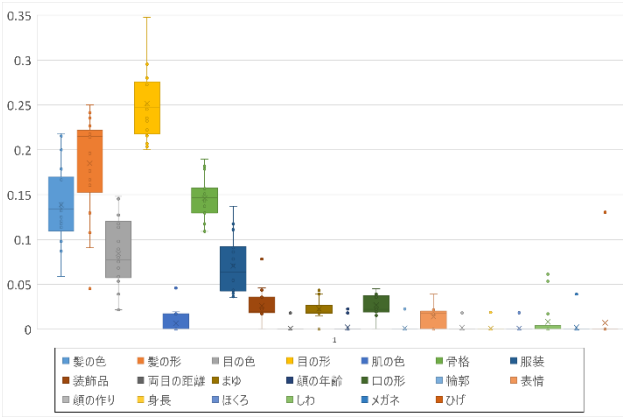

図 5　アンケートの結果（パーツ別）

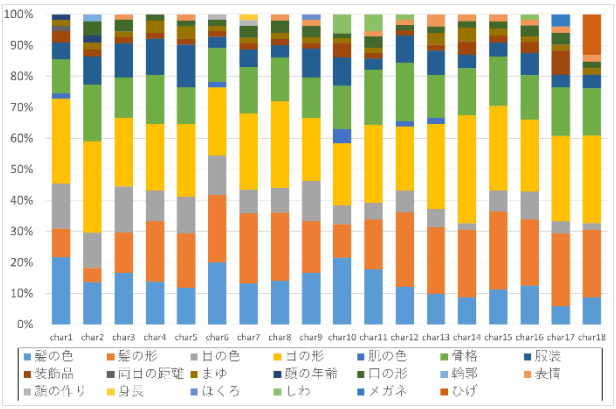

図 6　アンケートの結果（キャラクタ別）

実験参加者は 20 から 24 歳の男女 17 人である（男性:13 人，女性:4 人）．実験参加者の方には図 4 のイラストを見て，イラストのどこのパーツを見てイラストの声を想像したかアンケートで答えていただいた．アンケートには，「髪の色」「髪の形」「目の色」「目の形」「肌の色」「骨格」「服装」「装飾品」「その他」の項目があり，複数選択可能とした．「その他」の項目には，実験参加者がほかの選択項目がある際に自由に記述していただけるようにした．

5. 結果

アンケートによる結果を図 5,6 に示す．図 5 には実験参加者が「その他」で自由に記述していただいた声を想像するために必要なパーツを含めたパーツ別の割合を示している．図 5 は，1 つのキャラクタのイラストの内でそれぞれのパーツの割合を求め，18 キャラクタ分をパーツ別に箱ひげ図で作成した．図 5 の縦軸は 1 キャラクタあたりのパーツの総数を 1.0 とした時の，パーツごとの割合である．図 5 を見ると，「目の形」「髪の形」「服装」「目の色」の割合が高

1 ©2012-2020 OPAP-JP contributors.
(利用許諾:http://creativecommons.org/licenses/by/4.0/ 、改変して使用)
https://opap.jp/contributors

2 タイトル：ブラックジャックによろしく　著作者名：佐藤秀峰
サイト名：漫画 on web

くなっていることがわかる.

図 6 は図 5 同様に「その他」のパーツを含めた割合をキャラクタ別に棒グラフで作成した. 図 6 を見ると「目の形」はどのキャラクタでも一定の割合を占めているが，図 5 で割合の大きかった「髪の形」や「髪の色」はキャラクタによって低くなっているところがあることがわかる．また，実験参加者が「その他」で追記していただいたパーツのうち，一部のキャラクタは「ひげ」や「しわ」という独自のパーツを持っており、一定の割合を占めているため，キャラクタによって声を想像するために影響を与えるパーツはキャラクタによって増減する可能性があると考えられる.

6. まとめ

本研究では，電子書籍の付加価値として漫画のキャラクタのイラストや小説などの挿入イラストからキャラクタに合った声を生成し，ユーザが自由にキャラクタの音声を聞くことができるシステムを検討している．しかし，キャラクタの顔イラストの中で，どのパーツがキャラクタの声を推定するために必要なのかを検証する必要があったため，アンケートを用いて実験を行った．その結果「目の形」「髪の形」「服装」「目の色」の割合が高くなっていることが分かったが，キャラクタによって「髪の形」や「髪の色」はキャラクタによって低くなっていたり，独自のパーツが一定の割合を占めていることが分かった.

今後の展望としては，キャラクタの声を想像するのに必要なパーツとして高い割合を占めていた「目の形」「髪の形」「服装」「目の色」を用いてキャラクタの声を推定精度が向上するか検証していこうと考えている.

謝辞　アンケート実験にご協力していただいた実験参加者の皆様に謹んで感謝の意を表する.

参考文献

[1] 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, https://www.soumu.go.jp/main_content/000637255.pdf (参照 2020-06-18)
[2] 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, 情報処理学会研究報告(Web), Vol.2017-SLP-115, No.3, pp.1-6 (WEB ONLY), 2017 年 02 月 10 日.
[3] H. M. Smith, A. K. Dunn, T. Baguley, and P. C. Stacey. Matching novel face and voice identity using static and dynamic facial images. Attention, Perception, & Psychophysics, 78(3):868–879, 2016.
[4] 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.
[5] YUJIA WANG, WENGUAN WANG, WEI LIANG, LAP-FAI YU. Comic-Guided Speech Synthesis, ACM Trans. Graph. 38, 6, Article 187 (November 2019), 14 pages.（SIGGRAPH Asia 2019）
[6] 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に基づく音声生成の検討, 2020 年度情報処理学会関西支部 支部大会, 2020 年 9 月
[7] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik. Speech2face: Learning the face behind a voice. In CVPR, 2019.
[8] 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, 第 9 回データ工学と情報マネジメントに関するフォーラム(DEIM), 2017 年 03 月
[9] 林　大輔. 声優のキャラクター演技音声を用いた音声知覚に関する実験研究, 愛知淑徳大学論集―人間情報学部篇　第 9 号　2019 年 3 月，pp.49―62
[10] RADIUS5 Inc. “cre8tiveai”. https://cre8tiveai.com/sc (参照 2020-12-19