B-10

人間の認知特性による 2 次元キャライラストの音声推定手法の検討
Study on Estimation Method of Character Voice based on Cognitive Traits

大道 昇†　　大井 翔†　　佐野 睦夫†
Noboru Omichi　　Sho Ooi　　Mutsuo Sano

1. はじめに

人は人間の顔からおおよその声を想像することが可能であると考えられており，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]．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が提案されている [2]．この研究では，顔画像とその顔の持ち主の声を対応付けたデータセットから未知の顔画像に対して音声を生成することが検討されている．他に，現実の顔と声の関係性を 2 次元キャラクタのイラストに生かして，デジタル化された漫画を入力した時に視覚的な印象と一致する音声を合成する研究がある [3]．この研究では，キャラクタの顔画像から年齢と性別を推定し，キャラクタの音声を推定している．

上述した人やキャラクタの音声生成を試みる研究では，顔画像を入力として年齢や性別などの特徴や画像認識から得られた特徴を用いて音声特徴の推定を行っていた．しかし，顔画像から年齢や性別を推定する際の分類精度の誤差や画像認識の誤りから最終的に生成される音声に影響を与える可能性がある．特に漫画などのキャラクタの顔画像は，作者によるキャラクタの描き方は作者ごとに異なっており，一概に年齢や性別を判定することは困難であると考えられる．

そこで本研究では，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て，これまでの研究よりもシンプルなキャラクタの画像特徴から音声推定が可能であるか検討する．我々はこれまでに，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するか調査を行った [4]．その結果，特に声を想像する際に見ている顔のパーツとして「目の形」「髪の形」「髪の色」が有力な特徴であるとわかった．そこで本研究では図 1 の提案手法に示すように，キャラクタの顔画像のイラストに対して「目の形」「髪の形」「髪の色」を抽出し音声の推定する手法を検討する．

2. 関連研究

近年，人の声と顔とを組み合わせた研究が盛んに行われている．人の声と顔には何らかの関係性があると考え，人の顔画像とその人の音声を対にして学習して音声を生成する研究や [2]，人の顔画像からランドマーク検出を行い，各ランドマークの関係性から人の声を推定する研究がある [5]．一方で，ある人の音声からその人物の顔を推定する研究もある [6]．

また，人ではなくキャラクタに対して調査している研究もあり，キャラクタの声優のキャスティングに対してキャラクタの性格推定を行い、適した声質を推薦するシステム

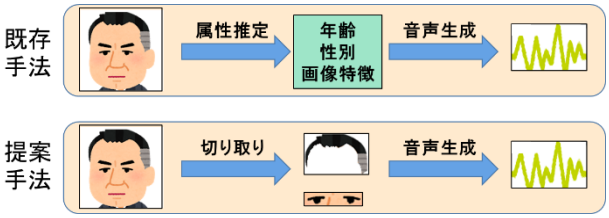

図 1　既存手法と提案手法の比較

や [7]，キャラクタの顔画像に対して年齢や性別の推定を行い音声特徴の抽出をする研究もある [3].

3. 提案手法

本研究では，キャラクタの顔画像のイラストに対して，人の認知特性を考慮した顔パーツの抽出を行い，シンプルな画像特徴から音声が生成できるか検討する．図 1 に示すように，これまでの研究では顔画像をそのまま入力として使用するか，年齢や性別などの属性を推定して音声生成を行っていた．しかし実際に人が顔画像から声を想像する際の過程を考慮していないため，本研究の調査で人が声を想像する際に有力な特徴であるとわかった「目の形」「髪の形」「髪の色」をシンプルな画像特徴として音声生成を検討する．

本研究でキャラクタの顔画像から音声を生成するまでの課題は図 2 に示すように以下のとおりである．

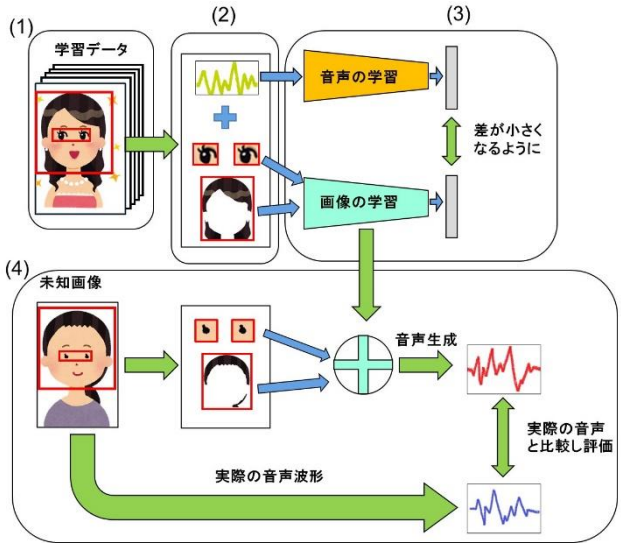

図 2　提案手法の流れ

†大阪工業大学，Osaka Institute of Technology

(1) ウェブサイトからキャラクタの画像を収集し，顔画像のみを抽出する
(2) 顔画像から「目」と「髪」を抽出する
(3) キャラクタの音声とキャラクタの「目」・「髪」の画像を用いて未知のキャラクタからそのキャラクタに合った音声を生成できるように学習する
(4) 生成された音声と実際のキャラクタの声とを比較する

本研究での音声性は学習データに使用するキャラクタの画像と音声を 1 対 1 であらかじめ対応付けて置く．そして，画像から抽出した「目」・「髪」のパーツとキャラクタ音声で特徴量抽出の学習を行う．その際に，顔パーツの学習器と音声の学習器で出力結果の差が小さいように学習を行う．その結果，顔パーツの学習器の結果が音声学習器の結果に近づけることができ，この学習器に未知キャラクタの画像を入力とすると，出力結果が音声特徴量として扱うことができると考える．また，アニメなどの音声を対応付けられる未学習のキャラクタ画像を学習済みのパーツ学習器に入力すると，生成された音声と実際の音声で比較することによって，このシステムを評価することができると考える．

本研究では，現在キャラクタの画像から「目」のみを抽出するまでの手法は確立しているが，「髪」の抽出方法は検討中である．「目」を抽出するまでの流れは図 3 に示すように以下のとおりである．

(I) 収集した画像に対して顔画像を抽出し，ランドマーク検出を行う
(II) ランドマークから「目」の中心と「目」の周囲の座標を取得し，左右の「目」の中心点を結んでキャラクタが水平になるように回転処理を行う
(III) 「目」の周囲の座標点を用いて「目」を抽出する

まず，収集したキャラクタのイラストに対して顔画像のみを切り出すために，OpenCV のアニメ顔検出のカスケードファイルである“lbpcascade_animeface”を使用する [8]．この際に，収集されたキャラクタの顔は水平から回転していることがあるため，カスケードで検出された顔画像の領域から高さと幅を 2 倍ずつ大きくした領域を確保しておくことで，顔の水平回転の処理をした後に顔領域の切り取りで，必要な部分まで欠けてしまうことを防ぐことができる．

顔画像の検出後，高さと幅を 2 倍ずつ大きくした領域に対して顔のランドマーク処理を行う．使用するランドマークは事前にアニメの顔のランドマーク検出として学習されたモノを使用し，ランドマークの点の数はキャラクタの顔に合わせるように 24 個となっている [9]．ランドマーク処理を行った結果の画像を図 4 に示す．検出された顔画像に対して図 4 における点 11 から点 20 までの左右の「目」の位置関係が，ありえないことになっている画像を除外することによって，顔検出で誤って取得した画像を除外することが可能となる．

顔画像から「目」を取得する際に，ランドマークによって得られる「目」の周囲の点の座標から得られる「目」の領域は，イラストの顔が傾いていると正しく取得できないので，事前に顔が水平になるように回転させる必要がある．顔画像の回転には，図 4 における点 15 と点 20 である左右

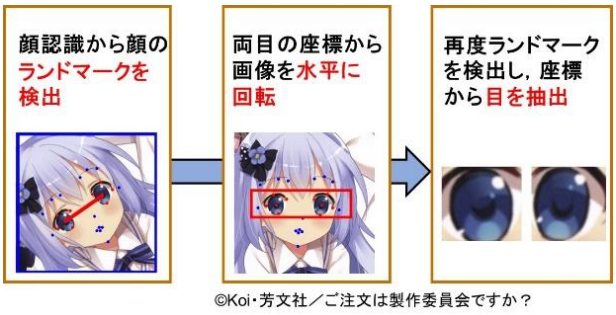

図 3　「目」の抽出手法

図 4　顔画像に対するランドマーク処理の結果
「目」に関する点を赤色，それ以外の点を青色としている．
画像は cre8tiveAI (https://cre8tiveai.com/sc)にて作成した画像を使用している．

の「目」の中心点を結んだ直線と画像の水平線から角度を求め，画像を回転させる．

顔画像の回転処理が済んだ画像に対して，再度ランドマーク検出によって「目」の座標を再計算し，図 4 で示す右「目」は点 11 から 14，左「目」は点 16 から 19 を用いて「目」の領域を抽出する．

4. 実験

本研究の実験は，図 5 に示すようにアニメなどの音声を対応付けられる未学習のキャラクタ画像を学習済みのパーツ学習器に入力し，生成された音声と実際の音声で比較することで，このシステムを評価することができると考える．また，実験に使用するキャラクタのイラストに対して，生成された音声が想像する声に近いかアンケートを行うことで，人の感性から想像されるキャラクタの音声を再現することができるか検証を行うことができると考える．

また，アンケートから得られた有力な各特徴が与える影響のパラメータの調整や顔画像から取得する特徴の数によって結果に変化があるか検証する必要もあると考えている．

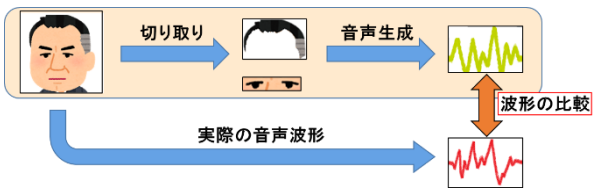

図 5　実験手法

5. まとめ

本研究では，人は人間の顔からおおよその声を想像することが可能であると考えられており，顔画像から音声を生成する研究がある中，シンプルな画像特徴からキャラクタの音声を生成する手法を提案した．本研究では，これまでの調査から得られた有力な特徴である「目の形」「髪の形」「髪の色」のうち，「目」の抽出手法について述べた．また，顔のパーツから音声を生成する流れを提案した．

今後は，「髪」の抽出手法について検討するとともに，キャラクタから取得したシンプルな特徴から音声を生成し評価していく予定である．

参考文献

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

[2] 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.

[3] Wang, Yujia, et al, “Comic-guided speech synthesis,” ACM Transactions on Graphics (TOG) 38.6 (2019): 1-14.

[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.

[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] Oh, Tae-Hyun, et al, “Speech2face: Learning the face behind a voice,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[7] 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.

[8] nagadomi, “ nagadomi/lbpcascade_animeface, ” *https://github.com/nagadomi/lbpcascade_animeface*, （参照日：2021/0718）.

[9] kanosawa, “ kanosawa/anime_face_landmark_detection, ” *https://github.com/kanosawa/anime_face_landmark_detection*, (参照日:2021/07/18).