2 次元キャライラストの顔パーツにおける音声生成手法の検討

大道昇† 大井翔† 佐野睦夫†

概要：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，人の顔から音声の生成を試みる研究が存在する．その中で，キャラクタの音声を推定する研究では，年齢や性別を推定したのちに音声を生成しているが，年齢や性別の推定誤差により生成される音声に影響を与える可能性がある．我々はこれまでに，アンケートから人はキャラクタの「目の形」から声を想像することが分かった．本研究では，「目の形」から声を生成する手法について，Tacotron2 と WaveGlow を用いて音声を生成する手法を提案する．結果として、学習したキャラクタの特徴を含んだ音声を生成することができた．

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 Deep Neural Network（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，近年普及しているオーディオブックに注目し，オーディオブックのキャラクタの音声を自動で生成することを目的として，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が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．しかし，同じキャラクタでもそのキャラクタたらしめる特徴はかけ離れたものになるとは考えにくく，かけ離れている場合，別のキャラクタであると考えられる．そのため，我々はキャラクタの音声を推定するにはキャラクタの画像から年齢や性別を推定したのちに音声を推定するのではなく，キャラクタの顔画像から音声を推定するために有力なパーツを抽出し，顔画像のパーツ画像から音声を推定する手法を検討する.

我々はこれまでに，キャラクタの声を推定する手法を検証してきた．まず，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[4]．結果として，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．実験では，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.

そこで我々はキャラクタの顔のうち，どの顔のパーツがキャラクタの音声を推定するための特徴なのか検証する必要があると考えた．方法として，実験参加者に対してキャラクタの顔画像を見た際にどこを見て声を想像しているかアンケートを行った[5]．その結果，キャラクタの顔のパーツのうち，「髪の形」「髪の色」「目の色」が声を想像するために有力な顔のパーツであることが分かった.

これまでの研究を踏まえ，本研究ではキャラクタの顔画像のイラストに対して「目の形」を抽出し音声の推定する手法を検討する. アンケートの結果では「髪の形」「髪の色」も声を想像するために有力な顔パーツであった．しかし，キャラクタの画像は実際の人の写真画像と比べて，髪の色や形がかけ離れており，実際の人で用いられている髪の抽出手法を適用することが難しいと考えた．そのため，本研究では，「目の形」のみを用いて音声の推定を行う．キャラクタの「目の形」を画像特徴量とし，未知のキャラクタからそのキャラクタに合った音声が生成できるように学習す

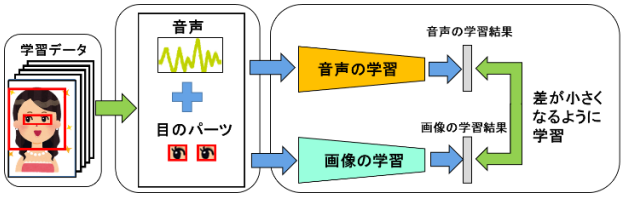

図 1 システムの概要図

†大阪工業大学

る．システムの流れを図 1 に示す．

本来は，キャラクタごとに様々な声を生成できることが目標であるが，本研究では未知のキャラクタが学習したどのキャラクタに近いか判定を行い，一番画像特徴量が近いキャラクタの声を割り振ることとする．

2. 関連研究

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，人の顔と声とを組み合わせた研究が盛んにおこなわれている

2.1 人の声から顔に対するアプローチ

人の声には，声を発した人の性別や年齢などあらゆる情報が含まれていると考えられており，声から声を発した人の顔を推定・生成する研究が行われている[6]．またテキストからの音声合成に加えて，生成された音声に合わせて顔画像の唇が動いているような映像を生成する研究もある[7]．

2.2 人の顔から声に対するアプローチ

人の顔にも様々な情報が含まれている．その中でも，人はある人物の顔画像を見たとき，その人のおおよその声を想像できると考えられており，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 [8]．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や [9]，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，キャラクタの音声を推定する手法を検討する．我々はこれまでに，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[4]．しかし，5 キャラクタ中 3 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．

そのため我々はキャラクタの顔のうち，どの顔のパーツがキャラクタの音声を推定するのに必要なのか検証する必要があると考えた．まず，我々は人がキャラクタの画像を見た際に，キャラクタの顔画像のどこを見て声を想像しているのか確かめる必要があると考えた．そこで，実験参加者に対してキャラクタの顔画像を見た際にどこを見て声を想像しているかアンケートを行い[5]，キャラクタの顔のパーツのうち，「髪の形」「髪の色」「目の色」が声を想像するために有力な顔のパーツであることが分かった．

そこで本研究ではキャラクタの顔画像のイラストに対して「目の形」を抽出し音声の推定する手法を検討する．アンケートの結果では「髪の形」「髪の色」も声を想像するために有力な顔パーツだとわかった．しかし，キャラクタの画像は実際の人の写真画像と比べて，髪の色や形がかけ離れており，実際の人で用いられている髪の抽出手法を適用することが難しいと考えた．そのため，本研究では，「目の形」のみを用いて音声の推定を行う．キャラクタの「目の形」を画像特徴量として学習し，未知のキャラクタが学習したどのキャラクタに近いか判定を行い，一番画像特徴量が近いキャラクタの声を割り振ることとする．

「目の形」を抽出する手法として，ランドマーク検出を用いる．ランドマーク検出から，「目」を抽出する流れを図 2 に示す．まず，収集したキャラクタから顔画像のみを抽出し，ランドマーク検出を行う．この際に，検出されたランドマークから目のみを抽出すると，顔が水平になっていない場合，目の抽出がうまくできないことがある．そのため，検出したランドマークのうち，両目の中心座標を用いて顔画像を水平に回転する．その後，再度ランドマーク検出を行い，目の抽出を行う．

また，抽出された目の画像から，「目の形」を抽出するために，エッジ検出を行おうと考えている．エッジ検出されたキャラクタの目の画像を「目の形」として，キャラクタの声の推定に使用しようと考えている．

キャラクタに割り当てる音声はあらかじめ 1 キャラクタ当たり複数の音声から，Tacotron2 を用いてキャラクタの音声を学習する[10]．この際に，キャラクタごとの音声の学習には，音声データとセリフのテキストデータが必要になる．しかし，キャラクタの音声とセリフのデータをキャラクタごとに多く集めることは困難である．また、Tacotron2 は英語の音声に対する音声特徴量を抽出する手法である．そのため，キャラクタの音声特徴量の学習を行う前に，あらかじめ日本語音声による転移学習を行う必要がある．その後，

図 2　ランドマークによる目の抽出

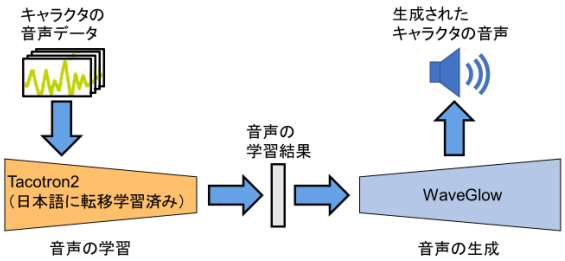

図 3　音声生成の流れ

日本語で転移学習を行った学習結果からキャラクタごとの音声を転移学習することで，少数のキャラクタの音声とテキストデータだけでキャラクタの音声特徴量の学習を行うことが可能となる．音声生成は WaveGlow を用いる[11].

4. 実験と結果

本研究の実験では，図 4 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分類器に入力し，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．

実験には同一のゲームやアニメ・漫画内のキャラクタから男女それぞれ 5 キャラクタ程度を使用しようと考えている．同一にする理由として，同一のゲームやアニメ・漫画だと収録環境が統一されていると考えられ，生成された音声のクオリティに違いを生まないためである．また，本実験の対象キャラクタとして，人型のキャラクタとし，獣人やマスコットのようなキャラクタは対象外とする．

我々は前実験として，日本語に転移学習した Tacotron2 を用いてキャラクタの音声の推定を行った．1 キャラクタ当たり約 100 文章を学習に用いた．結果として，生成された音声には聞き逃せないほどのノイズが含まれていたが，元のキャラクタの音声と比べて，ある程度特徴をとらえた音声を生成することができた．ノイズが含まれた原因として，学習量が少なかったのと，学習に使用したテキストデータの内容に特殊な固有名詞が含まれていたため，学習用にテキストデータの変換を行う際に誤変換されてしまった事であると考える．

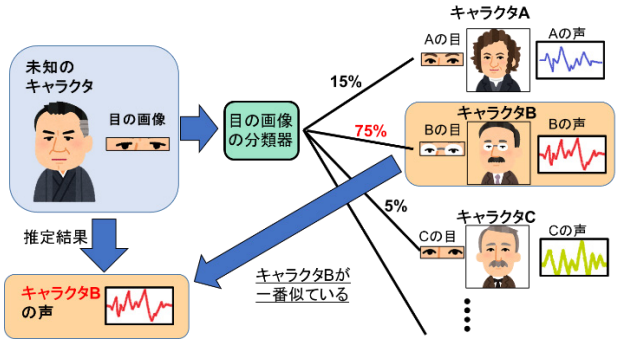

図 4　音声推定の流れ

5. まとめ

本研究では，キャラクタの顔画像に対して音声を推定・生成する手法を検討した．音声の推定には顔画像からそのまま推定することは難しいことがこれまでの我々の研究で分かったので，本研究ではキャラクタの顔画像のパーツである「目の形」に注目して音声の推定・生成を行おうと考えた．目の抽出にはランドマーク検出を行い，エッジ検出を用いて「目の形」を抽出しようと考えた．

実験では Tacotron2 を用いてキャラクタの音声推定を行った．推定された結果から WaveGlow で生成した音声はノイズが含まれていたが，ある程度キャラクタの特徴を捉えることができた音声を生成できた．

今後の課題として，キャラクタから抽出された「目の形」を用いてキャラクタの音声推定を行い，キャラクタに適した音声を生成することができるか検証する必要があると考えている．

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 講演論文集 2020 (2020).

[5] 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.

[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] Schroeter, Juergen, et al. "Multimodal speech synthesis." 2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532). Vol. 1. IEEE, 2000.

[8] 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.

[9] 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

[10] Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[11] Prenger, Ryan, Rafael Valle, and Bryan Catanzaro. "Waveglow: A flow-based generative network for speech synthesis." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.