キャラクタに適合した合成音声の生成と人間によるイメージとの一致検証

斉藤 彰吾[1,a)]　大井 翔[2,b)]　佐野 睦夫[2,c)]

概要：本研究の目的は，オーディオブック等のキャラクタの画像に対して音声が付いた際に発生する違和感を解消する事である．また，従来の研究ではアニメキャラクタの目の画像から既存の声を推定する手法を用いていたが良い結果が得られなかった．その為，本研究ではアニメキャラクタにマッチする声の特性を分析しその分析データに基づいて学習データの割合を制定し合成音声の生成を行う．また，本稿では生成された合成音声とそれに対応したキャラクタ画像の一致具合を確認した実験を行う．

1. はじめに

何かをしながら本を読む手法の一つとして，音声読み上げ機能がある．また，近年では公益社団法人全国出版協会が発表したデータによると図 1 に示すように電子書籍の市場規模は順調に伸びてきていると見られる [1]．しかし，音声読み上げのクオリティは現状低いものが多く，声優を起用したオーディオコミックはコストがかかる．また，コミック等が映像化した際にキャラクタに音声が付いた際に思っていた声を違ったといった違和感が生じる事がある．また，先行研究としてキャラクタの画像特徴量のみを用いて音声を推定・生成するシステムを提案しているものがある [2]．しかし，この研究では人の顔画像の中で最も人の声を連想させているとされる目の画像を用いて音声推定・生成を行ったが生成された音声は人間が推定した場合に比べてキャラクタの性別が一致していなかったり結果としては不十分なものとなっていた．そこで，本研究では人間が推定した際と機械が声を推定したい際に乖離が発生しない様に，人の顔と声の傾向データを用いた学習データ割合調整による音声合成によって解決する事が出来ないかと考えた．そこで，人が顔を見た際に声を想像するメカニズムはキャラクタにも当てはまると考えられている [3]．また，声からも顔を想像出来ると考えられており [4]，そのメカニズムを利用する．より具体的には，何人かの実験参加者にランダムに提示したキャラクタの顔画像と音声データを 1 種類ずつ提示を行いそれらの一致度を収集し，得られた傾向データを用いて学習データの合成割合を制定し tacotoron2 [5] と waveglow [6] を用いて音声学習・音声生成を行う．また，本研究では傾向データを用いて生成された合成音声がキャラクタと一致しているかの確認実験を行った．

[1] 大阪工業大学大学院情報科学研究科
[2] 大阪工業大学情報科学部
[a)] m1m22a15@oit.ac.jp
[b)] sho.ooi@outlook.jp
[c)] mutsuo.sano@oit.ac.jp

図 1　出版年毎の電子書籍化率

2. 関連研究

過去の研究では，人間がキャラクターのどの要素に注目して声を想像しているかについての調査が行われ，その結果を利用して，画像特徴からの音声推定をより簡素化する可能性が模索されている [2]．この研究では，最初に人が顔画像を見た際に，顔のどの部分に着目して声を想像しているかに関する調査が行われた．研究の進行に伴い，「髪の形」や「髪の色」，「目の形」などが声を想像する際に重要な要素であることが分かった [2]．特に，図 2 に示すように「目の形」が最も声を想像する上で重要な要素とされ，次い

で「髪の形」も抽出された．この研究では，抽出された目と髪の画像を対応するキャラクターボイスに関連付け，これを利用して分類器と合成音声を生成しています．その後，未知のキャラクターに対して生成された分類器を用いて合成音声を割り当て，イメージに近い音声を生成することを試みている．しかしながら，この手法では音声割り当ての際に画像のみを使用しており，生成された合成音声が人間の期待通りのものでない可能性がある．そのため，本研究ではキャラクターの画像と音声を組み合わせ，人間が知覚する通りの音声を特徴分析モデルの傾向データとして取り込む．この特徴モデルを用いて音声生成を行うことで，画像だけでなく音声も活用する新たな手法を提案し，従来の手法よりもよりイメージ通りの音声を生成できる可能性があると考えている．

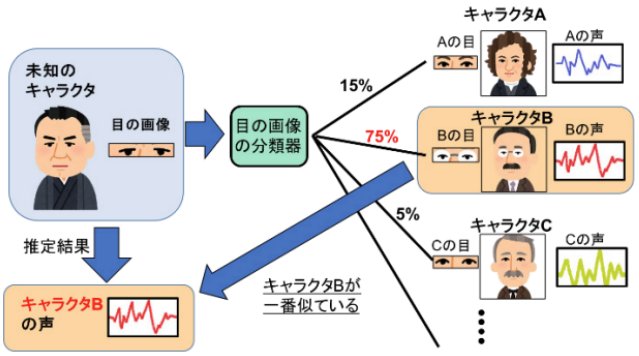

図 2 音声推定の流れ

3. 提案手法

ここでは，主にキャラクタのイメージ通りの音声を生成する手法となる学習データ割合の調整手法についての提案を行う．

3.1 画像と音声の一対比較法による傾向データ調査

本研究の目的は，キャラクターの画像に対する人間が感じるイメージに近い音声推定および音声生成だ．その為，この研究ではキャラクターのイラストと声の相互関係をより詳細に調査することを提案する．従来の研究 [2] では，特に目の形が音声を想像するのに有用であることが示され，その特徴を使用して音声推定と学習が行われている．ただし，これらのアプローチは，キャラクターのイラストに対する視覚的印象に基づいており，実際の音声を使用した調査では無かった．その為，本研究では，キャラクターのイラストとそれに合う声データの配分を見つけ，その配分で音声学習，音声生成を行えば未知キャラクタのイメージ通りの音声を生成可能だと提案する．しかし，具体的にどの様な事を行えばイメージ通りの音声となる学習データの配分を見つけられるのか，といった課題が残る．そこで，我々は図 3 に示すように何人かの人間に対して一枚の画像と複数の音声を用いた傾向データ分析実験を提案する．簡潔に説明すると，画像と声を人間が同時に見てどれくらい一致しているかを数値として入力し続けてもらい入力データを元に音声学習データの個数の配分を決定し配分通り合成する事でそのキャラクタ画像に対して適切な音声学習データの割合が算出され，その学習データ数に従って音声学習，音声合成を行う事でイメージ通りの声が作れると考えた．

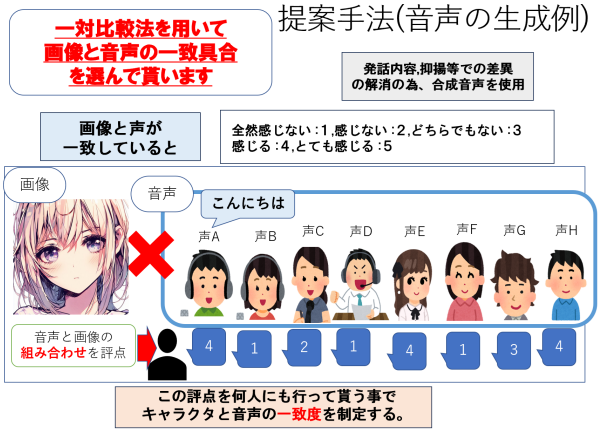

図 3 音声の生成例

3.2 トレーニングデータ数の決定

最後に学習データ数の決定手法についての提案を行う．先程算出された image に対してのそれぞれの声の割合を参照し学習データ数を算出する．今回例として挙げた場合だと声 A の学習データは 4%，声 B の学習データは 19%，声 C の学習データは 5%．といった風に合成割合が制定されている為，仮に学習データがそれぞれ 100 個あると考えた場合の学習データ数は声 A の学習データからは 4 種類，声 B の学習データからは 19 種類，声 C の学習データからは 5 種類．といった風にそれぞれから音声データを取得し，その後得られた音声学習データを用いて音声学習，音声生成を行えば image のイメージに近い音声が生成されると考える．また，得られるデータは割合な為，本研究では音声を機械学習させる際それぞれの割合を 3 倍した個数ずつ音声学習データを用意する．以下に学習データ数の決定の流れを図 4 に示す．また，本研究ではよりイメージ通りの声を創造する為，男女による声データの分類分けは行っていない．その為，男性キャラクタの画像に対しても女性キャラクタの音声データが一部学習に混ざるといったパターンが存在する．また，音声学習には tacotoron2，音声生成には waveglow を用いる．本研究では以上の傾向データ取得，それを元にした学習データ配分制定，制定した学習データ数による音声学習・音声生成を提案する．

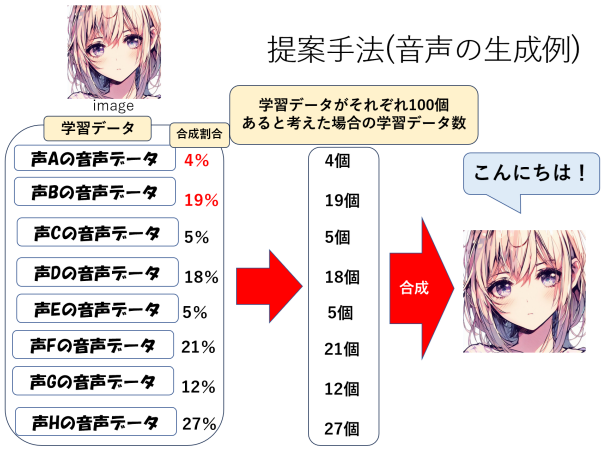

図 4 音声の生成例

4. 実験

4.1 合成音声割合制定の実験

実験では，図 5 に示す Bing Ai にて生成した 8 種類のキャラクタイラストと研究利用可能なキャラクタを演じた音声のコーパスから 8 種類の音声からそれぞれ 3 種類ずつ無作為に抽出し，24 種類の音声データを用意した．また，実験では重複しないランダムな画像と音声を提示するプログラムを作成し実験参加者に対して画像と音声をそれぞれ 1 種類ずつ提示する事を一人の実験参加者に対し 192 パターン行った．実験では画像と音声の組み合わせが一致していると，とても感じない場合は 1，感じない場合は 2，どちらでもない場合は 3，感じる場合は 4，とても感じる場合は 5 を選択する事とした．また，実験参加時に年齢や性別による感じ方の差を調査する為に年齢と性別の入力を行わせた．また本実験では 10 23 歳の男性 10 人に対して実験を行った．実験ではディスプレイはノートパソコンとディスプレイ 1 台，ワイヤレスマウスを用いて行った．また，実験の際には雑音によって実験の阻害を防ぐため実験用ワイヤレスイヤホンを用いた．

図 5 音声の生成例

4.2 生成された音声がイメージ通りかの確認実験

本研究では合成音声割合制定実験から得られた傾向データを用いて生成した合成音声がキャラクタの顔画像とイメージ通りの音声になっているかの確認実験を行った．実験では音声生成元となるキャラクタ画像 1 枚とそれを元にして生成された合成音声を同時に提示を行い，それに対してキャラクタから生成された音声はイメージ通りですか？といった質問を提示する．また，質問に対する回答としてキャラクタと声が一致していると，とても思う，思う，どちらでもない，思わない，とても思わない．の 5 段階の評価によってキャラクタと声が一致しているかの確認を行う，またこの際どうしてその評価にしたかを実験参加者に記述して頂いた．本実験では 20 35 歳の男性 11 人に対して実験を行った．実験ではノートパソコンとディスプレイ 1 台，ワイヤレスマウスを用いて行った．

5. 結果

5.1 合成音声割合制定の実験

合成音声割合制定の実験を行った結果を表 1 に示す．本研究ではこの結果から合成音声割合を制定し音声合成を行う，例として male1 のキャラクタの場合，voice1 の音声学習データを仮に全体で 100 個の学習データを用いる場合，6%である 6 個学習させる，voice2 の音声学習データの場合だと表 1 より 6%となる為 6 個学習させる．といった風に音声学習・音声生成を行う．

5.2 生成された音声のイメージ通りかの確認実験

本実験では得られた傾向データのうち female1 を元にして生成された合成音声を用いて確認実験を行った．結果を図 6 に示す．結果から考察すると一致していると思うが 63．6% といった結果が得られた．この結果から生成された音声は半数以上にはイメージ通りだと感じられたが完全にイメージ通りの音声になってはいないと考えられる．理由として，音声学習用いた傾向データの収集不足や用いた音声学習データの種類の少なさ等が考えられる．また，評価を選定して頂いた際にどうしてその評価にしたか？といった意見を頂いたが，「清楚で可愛らしい感じのキャラにマッチしてる音声だと思ったから」「声が震えて聞こえるという問題点はあったが画像の通り可愛い女の子のような声がした」といった肯定的な意見を散見されたが「見た目的にもう少しクールな感じで考えていた」「もう少し，低めの感じの声があっているかと思う」といったイメージとは少し違うといった意見も散見された．

6. おわりに

本研究では様々なキャラクタイラストと音声を組み合わせキャラクタに沿った音声の特徴分析を行い得られた傾向

表 1 傾向データ取得実験の結果

	male1	male2	male3	male4	female1	female2	female3	female4
voice1	6	6	6	14	17	17	16	17
voice2	6	6	7	14	22	20	17	20
voice3	20	18	16	8	5	5	7	5
voice4	6	6	6	12	15	18	14	16
voice5	29	25	23	11	6	4	8	5
voice6	11	12	17	16	8	8	11	8
voice7	6	6	6	15	21	23	19	22
voice8	18	21	19	10	5	5	8	7

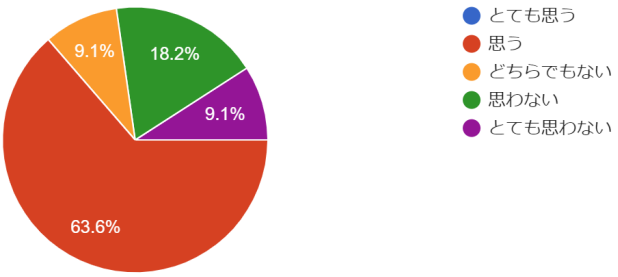

図 6 female1 を用いた確認実験の結果

データから学習データ割合を制定し音声学習を行い，生成された音声と学習元となったキャラクタイラストととの一致具合について調査を行った．今後は生成される合成音声をよりイメージ通りの音声とする為，機械学習元の学習データや傾向データ量を増やしより人間が感じるイメージに近い音声を生成を目指す．

参考文献

[1] 鷹野凌・堀正岳，日本における電子書籍化の現状（2020年版）．春秋合同研究発表会，2020.
[2] Noboru Omichi, Sho Ohi, and Mutsuo Sano. ”Study on Feature Extraction Method from 2D Character Illustration based on Human’s Cognitive Characteristics for Automatic Voice Estimation” AICCC’21, (2021).
[3] Smith, Harriet MJ, etal. “Concordant cues in faces and voices: Testing the backup signal hypothesis,” Evolutionary Psychology 14. 1 (2016): 1474704916630317.
[4] 栗津俊二・浅野遥．”静止画における未知人物の音声から外見の推定”，日心第 72 回大会，2008
[5] 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.
[6] Prenger, et al. ”Waveglow: A flow-based generative network for speech synthesis. ” IEEE International Conference on Acoustics, Speech and Signal Processing IEEE, 2019.