C-07

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

A Study on Speech Generation based on 2D Character Features for Automatic Generation of Audiobooks

大道 昇† 大井 翔‡ 佐野 睦夫†
Noboru Omichi 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、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．ところで，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 に示すように，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．

2. 関連研究

2.1. 音声と顔の対応関係の研究

音声のみで人とやり取りを行う音声インターフェースは様々な種類で普及しているが，人間同士の自然な会話の域には達していないと考えられる．自然な会話に近づける手法として音声だけでなく，インターフェースに話者のイラストを示すことがある．しかし，これらは既に性質が決められているものに話者のイラストを割り当てるか，話者の顔がすでに決められているものに声質を付与させる方法があり，顔と性質を適切に組み合わせる必要が生じるため大杉らの研究では[2]，Gaussian Mixture Model(GMM)に基づいてある顔に印象的に対応した声を推定する手法を提案している．同著者の先行研究と比較し，評価する顔画像によって有効に働く手法に違いがあることが分かった．

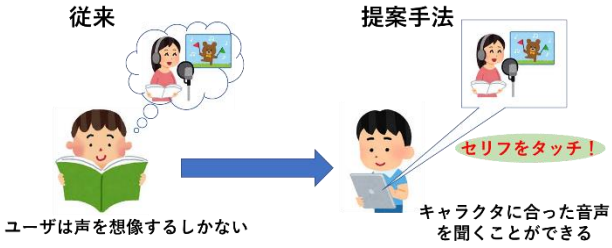

図 1　提案システムの概要図

2.2. 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[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[6]，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．

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．本システムでは，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

†大阪工業大学，Osaka Institute of Technology
‡ 立命館大学，Ritsumeikan University

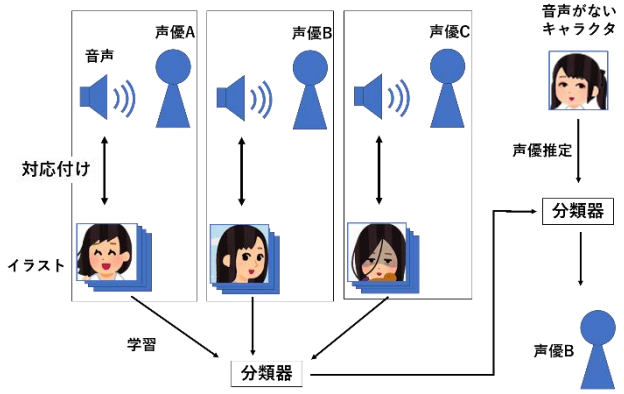

図 2　実験の流れ

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．

システム構築までの課題は次の通りである．

(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．

今回はイラストと音声の対応関係を学習できるかを検討する．まず，アニメやゲームなどのキャラクタのイラストと音声がついているキャラクタから，そのイラストと担当している声優を対応付けておき，声優ごとにキャラクタのイラスト画像をクラス分けして学習を行う．そして，学習に使用していない声優が割り当てられていない未知のキャラクタの，イラスト画像を入力として時に画像の特徴から，入力されたキャラクタの音声はどの声優クラスに適しているか比較することによって，入力されたキャラクタの音声はどの声優の声に近いかを確かめることができる．未知のキャラクタの音声を一番近い結果となった声優に割り当て，声優の音声でセリフを読み上げるように自動で出力することで，未知キャラクタが入力したセリフを話しているように音声を出力できるのではないかと考える．

4. 実験

本研究の実験では，あらかじめ声優の音声とイラスト画像を対応付けしておき，イラストの学習からどの声優の音声が一番近いのかを求める．

実験の流れを図 2 に示す．まず，キャラクタの分類器を学習するために，声優ごとにクラス分けされた画像を用意する．この時使用するイラストは，すでにクラス分けする声優の音声として付与されている複数のキャラクタとする．声優ごとにクラス分けした画像集を畳み込みによって学習を行い，分類器を作成する．作成された分類器に未だ音声が付与されていないキャラクタのイラストを入力したとき，分類結果からどの声優に対応付けられたイラストに近いか求められ，どの声優の音声が適しているかがわかる．

実験で用いる声優のクラス数は 3 クラスとし，それぞれのクラスで画像を 300 枚程度用いる予定である．

5. まとめ

本研究では市場規模が伸びてきている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た。ユーザが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することによってユーザが自由に音声を聞くことができるようにするためのシステムを計画し，音声が付与されていないキャラクタに対して，声優と対応付けられたイラストを用いた分類器で音声推定を行い，キャラクタに適している声優を求める実験を今後行う予定である．また，声優のクラス数を増やすことによるキャラクタの音声推定の向上や，セリフの内容に適した感情の音声を出力できるような Text To Speech（TTS）を作成していく予定である．

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] 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.