5R-01

ロボット教師における特定人物の音声の自動生成

寺尾 渉吾*　　康 鑫†　　任 福継‡

徳島大学

1. はじめに

人工知能技術とロボット工学が発展してきたことにより，それらを合わせた AI ロボットを利用した試みが盛んになってきている．その内の一つがロボット教師による人間の教師のサポートである．ロボット教師とは人工知能を搭載したロボットに授業や，授業のサポートをさせるという試みで，特に本研究が属するプロジェクトではスライドを用いた講義，学生からの質問への解答，テストの自動採点などの機能を搭載するための研究が進められている．言語学習用のロボット教師である「Elias」や「Musio」は既に実際の教育現場でも採用されており，どちらもネイティブでの自然な会話が可能で，レベルや目的に応じた問題が出せたり，学習状況をまとめ，可視化する機能などが搭載されている．本研究ではロボット教師としてアクトロイドを用いる．アクトロイドは高度に写実的な人間らしさを備えており，表情や口の動きも制御できるようになっている．そのため声においても外見等との齟齬が出ないように，使用するアクトロイドに合った声を生成する音声生成手法を提案する．

2. 提案手法

声質変換を用いて目標話者の声質を付与する方法と，ファインチューニングを用いて必要な音声・テキストデータペアを減らす方法を提案する．

声質変換の手法では，同一の文章を読み上げた入力話者と目標話者の音声を使い，入力話者の音響特徴量（メルスペクトログラム）が目標話者の音響特徴量に変換されるように学習を行う．そのベースのシステムとして Attention 付き Seq2Seq を用いた Text-to-Speech アルゴリズムである Tacotron2[2]を利用する．図 1 に構成を示す．Tacotron2 は TTS システムであり，テキストを入力することで音声が出力されるが，提案する声質変換手法では入力をテキストから音響特徴量に変える．テキストと比べて音声には非言語情報も含まれているため，自然なイントネーションを持つ音声を生成できるのではないかと考えた．

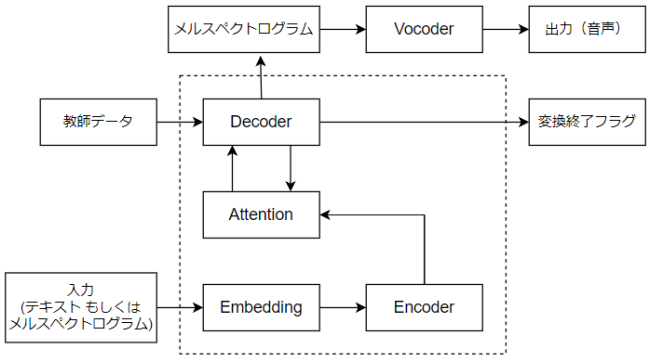

図 1 Tacotron2 の構成

ファインチューニングを用いた音声合成手法では目標話者の音声と，対応するテキストデータのペアを使い，テキストを音声に変換する．また同様にベースのシステムとして Tacotron2 を利用する．目標話者の声を一から生成する場合，目標話者の音声データと対応するテキストのペアが大量に必要になるが，あらかじめ別の話者で学習されたモデルを目標話者のデータで追加学習することで必要なデータを減らすことができる．

3. 実験条件

評価指標としてはメルケプストラム歪み(MCD)を用いる．これは出力と教師データの音響特徴量がどのくらい離れているかを表しており，次式で示される値である．

$$MCD = \left(\frac{10}{\log 10}\right)\sqrt{2\sum_{d}^{24}(mc_d^{conv} - mc_d^{tar})^2}$$

mc_d^{conv} と mc_d^{tar} はそれぞれ変換後のメルケプストラム，目標のメルケプストラムの d 次元目の特徴量である．

声質変換の実験には日本語音声コーパス CMU arctic より，入力話者として bdl，目標話者として rms を用いた．それぞれの話者ごとに学習には 1132 個，テストデータには 100 個のデータを使用した．また，比較する従来手法としては GMM で構

Generation of a Specific Person's Voice in the Robot Teacher
*Shogo Terao: Tokushima University, Graduate School of Sciences and Technology for Innovation
†Xin Kang: Tokushima University, Graduate School of Technology, Industrial and Sciences
‡Fuji Ren: Tokushima University, Graduate School of Technology, Industrial and Sciences

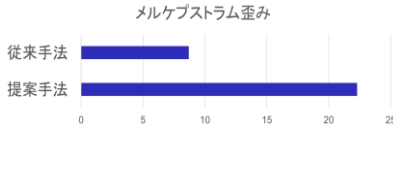

図 2 声質変換実験の MCD

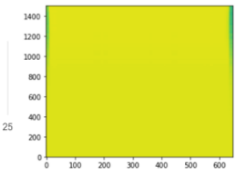

図 3 声質変換実験のアテンション出力

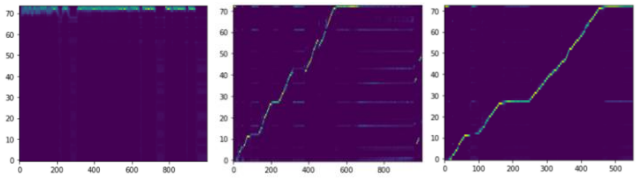

図 4 音声合成実験のアテンション出力
（左から事前学習なしで CSS10 を 6400 文，事前学習+CSS10 を 25 文，事前学習+CSS10 を 600 文使用したもの）

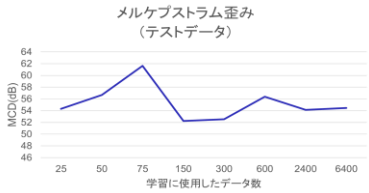

図 5 音声合成実験の MCD（CSS10 でファインチューニングした際の使用データ数による変化，図 4 の中央図・右図に対応）

成された sprocket[1]を用いた．

音声合成の実験では目標話者である日本語音声コーパス CSS10 のみで学習したモデルと，英語音声コーパス LJ Speech と日本語音声コーパス JSUT により事前学習されたモデルを CSS10 でファインチューニングしたモデルを比較した．LJ Speech は 13100 文，JUST は 7196 文，CSS10 は CSS10 のみで学習する場合 6400 文を使用している．テストデータには 100 文使用した．

4. 実験結果

図 2 及び図 3 に声質変換実験の結果を示す．図 2 のメルケプストラム歪みの比較より，提案手法のほうが目標との誤差が大きくなった．図 3 のアテンション出力は縦軸がエンコードされた隠れ状態，横軸がデコードされた隠れ状態を表している．各列のピクセル値は学習された重みであり，これが疎であれば（明るい線がはっきりとしていれば）モデルはアライメント情報を十分に学習したと言える．今回の声質変換実験においてアテンション出力は密になっており，十分に学習できていないことが分かる．この理由として，学習データの数が十分でなかったことや，構築したモデルに不具合があったことなどが考えられる．

図 4 及び図 5 に音声合成実験の結果を示す．図 4 は音声合成実験におけるアテンション出力であり，左の図が目標話者 CSS10 の音声テキストペアデータを 6400 文使って学習したもの．真ん中の図が LJ Speech 13100 文で学習したあと JUST 7196 文で学習し，最後に CSS10 を 25 文使ってファインチューニングしたもの．右の図は，最後の CSS10 のファインチューニングに 600 文使ったモデルのテストデータにおけるアテンション出力である．この 3 つの図を見比べると，左の図が入力の後半にしか注目されておらずアライメントがうまく学習されていないのに対して，事前学習を行った真ん中と右の図は入力と出力の関係を学習できていることが見て取れる．また，その中でもファインチューニングに多くのデータを用いた右の図のほうが線がはっきりとしており，アライメントの学習がよりうまくできている事がわかる．実際に生成された音声を聞いてみても，図 4 右図と同じモデルで生成された音声の方が一音一音をはっきりと発音できているように感じた．また，図 4 の 3 つは全て同じ文を入力しているが右図だけ横軸 500 程度で変換が終了している．ファインチューニングに使用する音声データが 150 文を超えたあたりから，入力文を読み終えた適切なタイミングで変換が終了するようになったためである．ただ図 5 に示されている，CSS10 でファインチューニングした際の使用データ数による MCD 変化のグラフを見ると，一概にデータ数を多くすれば教師データと出力データとの MCD が小さくなるわけではないことが判明した．

5. おわりに

特定人物の音声を再現する方法として声質変換を用いた手法と音声合成を用いた手法を提案した．声質変換については，従来手法との比較実験により提案手法は有効とは言えないことが示された．音声合成手法については，事前学習を用いた手法が有効であることが示された．今後の研究として学習データの増強や評価指標の吟味に取り組む．

参考文献

[1] K. Kobayashi and T. Toda, “sprocket: Open-Source Voice Conversion Software”, Proc. Odyssey, 2018, pp. 203-210.
[2] J. Shen, R. Pang, et al. “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions”, ICASSP, 2018,