7N-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† 任福継‡ 康シン‡ 西出俊‡
徳島大学大学院創成科学研究科理工学専攻†
徳島大学大学院社会産業理工学研究部‡

1. はじめに

AI 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 AI 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ものである．

これら教師ロボットを採用した場合のメリットに人手不足の解消が挙げられる．近年，労働人口の減少によって様々な業界で人手不足が問題になっており，それは教育の分野においても同様である．そこでロボット教師を導入することで，人手不足を補いながら，一人ひとりの学生の情報を AI により分析し，より効率的な教育を施すことができるようになると考えられる．

2. 研究目的

本研究は，人間型ロボット「アクトロイド」に講義をさせるというプロジェクトの一環として行っている．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，現在の TTS 技術で生成される画一的な声質を持つ合成音声が加わることで人間的な外見と機械音声の間に違和感が生まれ，人間に近い外見を持っているアクトロイドの利点である親近感や興味を持ってもらえるなどの要素を打ち消してしまうことが考えられる．そこで TTS で生成される音声に声質変換を施すことにより話者の個人性を付与し，アクトロイドの外見に適した声質の音声を生成することで解決できるのではないかと考える．

以上のことから，本研究では対象話者の声質へと音声変換ができるシステムの構築を目指す．

3. 関連研究

従来の統計的声質変換の研究としては，GMM を用いた手法[1]がある．パラレルデータを用いて参照話者と目標話者のスペクトル特徴量の結合 GMM P(x,y) を学習し，入力 x が与えられたとき，P(y|x) が最大となるようにスペクトル特徴量を変換する．このように，あらかじめ話者間の関係をデータから学習しておくことで，未知の入力が来た時にも変換が可能になる．

また，近年では DNN を用いた変換手法が多く研究されている．GMM での学習と同じように，パラレルデータを用いてその基本周波数やスペクトル特徴量が目標話者の特徴量にできるだけ近くなるように変換関数の学習が行われる．

それらの手法の一つとして，音声認識モジュールと音声合成モジュールを組み合わせ，未知話者の入力に対しても声質変換が行えるようにする方法がある．sequence-to-sequence アルゴリズムを適用し，入力音声をコンテキスト事後確率などに変換した後，それを入力に音響特徴量を生成するようにしている．この手法では音素同士のアライメントも学習できるようになるため，モデルの学習と無関係に起きていたアライメントエラーを改善し，かつ，事前のアライメントも不要になるという利点も存在する．

ただし，ノイズに頑健で正確な変換ができる音声認識が必要になるため，音声認識の学習に大量の音声及びそのテキストデータがいるという問題がある．それに対し，元話者と目標話者の音声データのみで学習を進められる手法[2]が報告されている．田中らは言語特徴量を入力に使わずに，学習を安定させ，入力音声の言語情報を保存するため，補助的な働きをする 2 つの Decoder を導入した．結果，RNN ベースの音声合成に匹敵する品質の音声が生成できた．

4. 提案モデル

本研究で提案するモデル構成を図 1 に示す．本システムのベースは Text-To-Speech アルゴリズムの Tacotron2 であるが，Encoder への入力をテキスト情報からスペクトル特徴量に変更することで声質変換ができるようにしている．主な要素は，Encoder，Attention，AttentionRNN，DecoderRNN の 4 つである．Encoder は入力話者の特徴量を出力話者の特徴量との共有空間に埋め込む役割，Attention は入力をどこまで変換した

Sequence to sequence voice conversion for robot teachers
†Shogo Terao
‡Fuji Ren, Xin Kang, Shun Nishide
†Tokushima university
‡Graduate School of Technology, Industrial and Social Sciences

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

か示す役割，PreNet 及び AttentionRNN は Attention の Query を作る役割，DecoderRNN 及び LinearProjection は1step 分の出力を生成する役割を持つ．加えて，田中らの提案した 2 つの Aux Decoder で Encoder の機能を補助することで，スペクトル特徴量のみで声質変換ができるようになることを期待する．Aux Src Decoder は Encoder と合わせて AutoEncoder の形をとっており，入力 X をそのまま復元することで Encoder の適切な特徴抽出を補助する．Aux Tgt Decoder は，Encoder が入力 X を入力話者と目標話者の共有空間にエンコードする補助となるようにしている．

また，Attention の学習を加速，安定させるために，guided attention loss[3]を実装する．音声合成や声質変換では機械翻訳と違い，Attention が単調になるという特徴が存在する．これを利用し，Attention の損失関数に，対角線上から離れるほど大きいペナルティを課すという制約を付ける．

5. 実験

5.1 実験条件

今回の実験は，Aux Decoder 及び guided attention loss を除いたモデルで行っている．

データセットは CMU Arctic database[4]より，bdl を入力話者，rms を目標話者として用いた．評価実験では従来手法として sprocket[1] の DIFFVC 手法を用いる．学習には588 文の音声データを使用し，従来手法での学習では事前にアライメントを行った．提案手法の特徴量抽出及び Vocoder には音声分析変換合成システム World[5]を利用した．

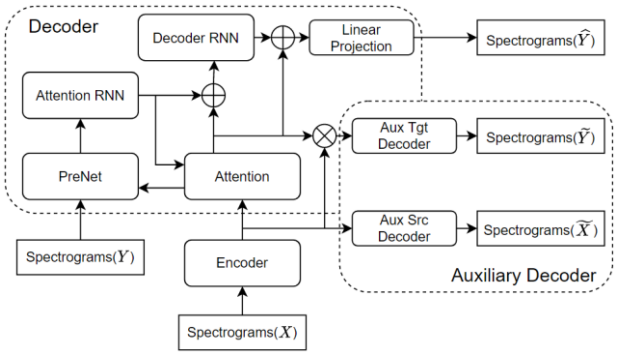
5.2 実験結果

図 2 は学習に使用していない音声を使って推論を実行したときの Attention のグラフ，表 1 はスペクトル変換精度の客観的な評価尺度として用いられているメルケプストラム歪みの比較を示している．この指標は小さければ小さいほど声質が似ていることを表すのだが，従来手法は 8.7，提案手法は 22.32 となっており，大幅に精度が低くなってしまっている．図 2 から分かる通り，Attention が正常に学習できていないことが原因ではないかと考える．guided attention loss を実装することで改善されることを期待する．

6. おわりに

本研究ではロボット教師の音声に個人性を付与することを目的に，sequence-to-sequence アルゴリズムを用いた声質変換のシステムを構築した．現在実装時点のモデルでの実験では従来手法より精度が低いという結果になった．今後新たな機能を追加することで精度の向上を目指す．

図 1 提案モデル



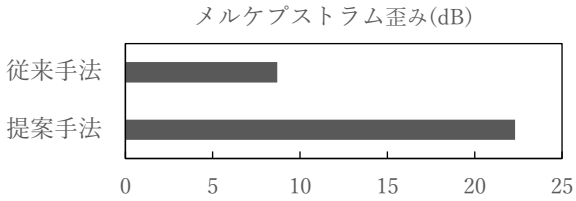


表 1 客観評価結果

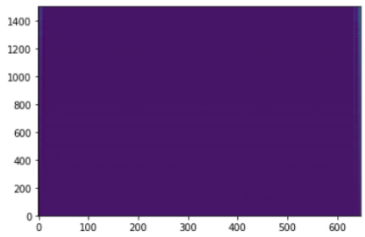


図 2 推論時の Attention

参考文献

[1] K. Kobayashi *et al.*, “sprocket: Open-source voice conversion software,” in Proc. Odyssey 2018 The Speaker and Language Recognition Workshop, 2018, pp. 203-210.
[2] K. Tanaka *et al.*, “AttS2S-VC: Sequence-to-sequence voice conversion with attention and context reservation mechanisms,” in Proc. ICASSP, 2019.
[3] H. Tachibana *et al.*, “Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention,” in 2018 IEEE ICASSP. IEEE, 2018, pp. 4784-4788.
[4] J. Kominek *et al.*, “The CMU Arctic speech databases,” in Fifth ISCA workshop on speech synthesis, 2004.
[5] M. Morise *et al.*, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE TRANSACTIONS on Information and Systems, vol. 99, no. 7, pp. 1877-1884, 2016.

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.