4Q-08

機械学習を用いた声質変換手法

吉田　天哉‡　田村　仁‡
日本工業大学‡

1. 研究背景

近年, 動画投稿サイトやコミュニケーションツール上で, キャラクターに成りきるという楽しみ方が注目されている. しかし, 姿は自由に変えられても声は本人の声のまま, ということが多い. そこで, テキスト読み上げソフトに使われている声に自分の声を変換することが出来ればその音声を提供できる場が広がると期待できる.

そこで, 機械学習を用いた声質変換で音声を異性や特定人物の物に変換しようと考えた. 声質変換とはある音声に含まれる声質を他の声質に変換する技術である.

既存の声質変換については[1]にまとめられている. 最近ではそれに加えて機械学習を用いた試みも多く, 例えば[2]などがある. これらの手法では複数の話者が同じ内容を話しているペア音声が大量に必要になる. しかし, 音声データを大量に用意するには時間がかかり, 二人分の音声データセットを集めたとしてもそのデータセットは本人の声質変換にしか使えない. 気軽に別人の声になることはできず, 準備に時間がかかってしまうというのは大きな問題だと考える. テキスト読み上げソフトで音声データを集め, 多対一声質変換に利用できれば, 使用者の音声データを集める必要なく, 使用者の音声の抑揚や感情をテキスト読み上げソフトの声で再現できるのではないかと考えた.

2. 提案手法

本研究では, 人間の音声をテキスト読み上げソフトに使用されている声に変換することを目的とし, データセットの作成に時間のかかる人間の音声も読み上げソフトで代用する方法を提案する.

複数の機械音声をターゲット音声に変換する学習を行い（図 1）, それによって得られた学習モデルを使用して話者本人の音声をターゲット音声に変換する（図 2）.

Voice conversion method using machine learning
†Yoshida　Takaya
†Tamura　Hitoshi
‡Nippn Institute of Technology

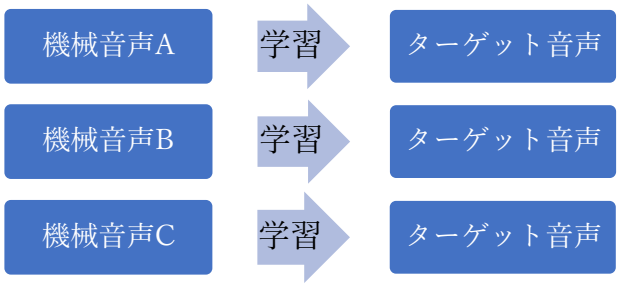

図 1　学習のイメージ

図 2　実際に行う変換

今回は日本声優統計学会が公開している音素バランス文[3]からテキスト 100 文を用意した. それをテキスト読み上げソフトで音声データにすることで学習を実行し, そこで作られた学習モデルを使い声質変換を行った.

3. 評価実験

3.1 実験方法

テキスト読み上げソフトは女性音声が男性音声に比べて手に入りやすかったため, 複数の女性音声で学習データを作成し, その学習データを用いて音声を変換した場合の変換結果を評価する.

今回, 声質変換を行うための学習用音声を集めるために三つのテキスト読み上げソフトを用意した. 一つを変換目標となるターゲット音声とし, パラレルデータを集める音声としてターゲット音声とそれ以外の二つを用い, ピッチを下げることでデータの量を増やし音声バランス文 100 文五つ分, 計 500 文の音声データを作成し学習モデルを作成した.

作成したモデルを用いて, 男性音声（評価音声 1）, 女性音声（評価音声 2）, 学習に使用していない読み上げソフト（評価音声 3）, 学習に使用した読み上げソフト（訓練音声）の四つの音声で「こんにちは」と発話し音声を変換する.

変換方法として, 音声を基本周波数, スペクトル包絡, 非周期性指標の 3 つに分解して, 調整する

ことで,音声変換を行う.

3.2 実験結果

変換したデータとターゲット音声データの類似度を定量的に評価するためにメルケプストラム歪み(Mel-cepstral distortion:MelCD) [dB]という指標を用いる.

音声は何を話しているかを表す音韻情報と声のピッチやパワー,リズムを表す韻律情報に分けられる. メルケプストラム歪みとは,音韻情報を要素数の少ない音響特徴量として表現する際に使用されるメルケプストラム[4]という音響特徴量がどのくらい離れているかを表す数値である.

$$MelCD = \left(\frac{10}{log10}\right)\sqrt{2\sum_{d}^{24}(mc_d^{conv} - me_d^{tar})^2} \quad (1)$$

ここで, mc_d^{conv}, me_d^{tar}は変換メルケプストラム, 目標メルケプストラムにおけるd次元目の特徴量を表す. MelCD の値が小さいほど,近い声であるといえる.メルケプストラム係数を比較するにあたって今回は,24 次メルケプストラム係数を使用した.

表 1 に各評価音声の変換前 MelCD と変換後 MelCD を纏めた.

表 1 各評価音声の MelCD

変換音声	平均 MelCD[dB]	
	変換前	変換後
訓練音声	19.1	16.0
評価音声 1	45.0	31.5
評価音声 2	45.9	30.8
評価音声 3	14.9	20.1

4. 考察

データセットの作成に関して,音声データを集めるにあたり, 今回使用した音素バランス文 100 文を用意する場合,誤読せずに読み上げるだけで 11 分ほどかかってしまう.さらに,録音機材の操作や誤読,環境音の混入,録音データの確認があり実際の録音にはかなりの時間がかかってしまう.テキスト読み上げソフトを使用することで自分の声を録音することに比べ大幅に時間短縮することができた.表 1 より評価音声 3 以外は MelCD が低くなりターゲット音声との差が小さくなっていることがわかる.評価音声 3 について平均 MelCD は増加しており正しく変換できていない.しかし,図 4 の評価音声 3 の MelCD を時間ごとに比較してみると一部を除き MelCD が変換前より下がっていることがわかる. 以上の結果から本手法で学習データにない音声でもターゲット音声に変換することは可能だと考えられる. しかし,学習用データが少ないため精度が上がらず,誰の声でも変換できるという目標は達成できなかった.

主観的評価としては,変換後音声は言葉の発音などは問題なく聞き取れるが,変換後音声とターゲット音声を聞き比べた場合,ターゲット音声の声質に近づいてはいるが同じ声質とは感じず,変換の精度が足りないと考えられる.

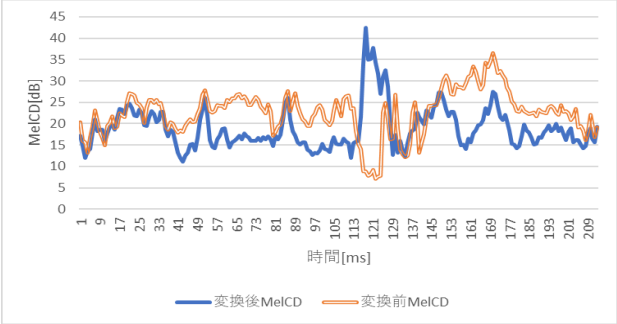

図 4 MelCD 時間変化

5. まとめ

本研究では声質変換を用いて気軽に別人の声に変換することを目的とし, 音声データを録音する代わりにテキスト読み上げソフトで音声データを集め,多対一声質変換を行う手法を提案した.

今後の課題として学習データの充実と精度の向上,男性用の音声データの収集が考えられる.

参考文献

[1] 戸田智基・小林和弘 (2018) 「統計的声質変換ソフトウェア入門」 (「アイサイ研究者のための音声情報処理ソフトウェア入門」特集号),< https://www.jstage.jst.go.jp/article/isciesci/62/2/62_69/_pdf/-char/ja>2019 年 7 月 24 日アクセス

[2] 「ディープラーニングの力で結月ゆかりの声になってみた」,< https://blog.hiroshiba.jp/became-yuduki-yukari-with-deep-learning-power/>2019 年 7 月 23 日アクセス
Copyright (c) 2018 Kazuyuki Hiroshiba.

[3] voice-statistics/voice-statistics.github.com, < https://github.com/voice-statistics/voice-statistics.github.com/blob/master/assets/doc/balance_sentences.txt >
2020 年 1 月 8 日アクセス

[4] 日本音響学会 (2017) 『音響学入門ペディア:日本音響学会編』コロナ社.