5X-06

深層学習を用いた多対多声質変換による発話者匿名化

安井　慎一郎† 伊達　伸之輔‡ 岩本　健嗣†

富山県立大学　工学部　電子・情報工学科†

富山県立大学　工学研究科　情報システム工学専攻‡

1 はじめに

個人情報保護法では，人の声は個人が識別可能である場合は個人情報として扱うべきであるとされている[1]．そのため報道番組のような個人情報を扱う際は，音声による個人の特定を防ぐために，変声機や声質変換を利用することで匿名化を行なっている．変声機を利用する場合，元音声のピッチとスペクトルの変換を行う手法であるため，抑揚や声の強さといった個人を特定できる発話者の特徴が残り，聞き手が不自然と感じる音声に変換される．また，声質変換を利用する場合，発話者の発話内容を保持しつつ，音声の声質を別人の声質に変換するが，変換先の音声は実在する話者であり，個人情報保護の観点から，発話者の匿名化に適していないと言える．また，二話者間で同一発話内容の音声(パラレルデータ)を用意して学習する必要があり，パラレルデータの発話位置を時間軸方向にアライメントする必要がある．以上より，既存の手法を利用した場合，発話者の完全な匿名化は困難である．

2 提案手法

本研究では，完全な匿名化をする上で，深層学習により，新しい音声を生成する手法に着目する．既存手法として，1 対 1 の声質変換手法である CycleGAN-VC2[2]がある．この手法は，多対多の画像生成ネットワークである CycleGAN[3]を発展させたものである． CycleGAN の画像変換手法は，一つの種類の画像の集合から別の種類の画像(例えば、男性から女性)の集合に変換するといったドメイン変換が可能である．加えて，CycleGAN は，パラレルデータが不要という特徴がある．そのため，パラレルデータの用意の難易度が高い声質変換に適応しやすい．また，多対多になることによってドメインは変換されたとしても，生成された画像は対象となるドメイン内の特定の誰かの画像でもないため，CycleGAN-VC2 においても多対多を適応することでドメイン内のどの話者でもない音声を生成することができると考えられる．

Deep Learning based Speaker anonymization by many-to-many voice conversion

Shinichiro Yasui†, Shinnosuke Date‡ Takeshi Iwamoto†

†Department of Electrical and Computer Engineering, Faculty of Engineering, Toyama Prefectural University

‡Department of Information System Engineering, Graduate School of Engineering, Toyama Prefectural University

3 目的

本来 1 対 1 の声質変換を目指したモデルである CycleGAN-VC2 を用いて，本研究では，CycleGAN のドメイン変換が行えるという特性を活かし，多対多声質変換を行うことで発話者の匿名化を目指す．

4 実装

CycleGAN-VC2 を用いて 1 対多の学習を行うにあたり，変換するモデルの損失が高くなることが予想される．そのため，全体の損失の影響を考慮し，ドメイン毎に非対称に学習率の調整を行った．また，Identity-Mapping Loss という損失が言語情報の保持に起因しているため，発話内容を保持しやすいように， Identity-Mapping Loss の重みを学習が進む世代ごとに小さくした．

5 実験

本実験では，多対多声質変換を行う前に，JVS コーパス[4]を用い，1 対 1 の声質変換と 1 対多の声質変換の学習を行った．JVS コーパスとは，100 人の声優，俳優などの話者が日本語テキストをパラレルデータ 100 センテンス，ノンパラレルデータ 30 センテンス，ささやき声 10 センテンス，裏声 10 センテンスを読み上げたものを収録したものである．

1 対 1 の声質変換では， 100 センテンスのパラレルデータを用いて学習を行った．1 対多の声質変換では，男性話者 1 人 30 センテンスと女性ドメイン 5 人各 30 センテンスの 150 センテンスを用いて学習を行った．

6 結果・考察

一般に男性は低音域から中音域にかけて，女性は中音域から高音域にかけて音声が分布する．男性と女性を比較した時には低音域が，女性同

士を比較した時には中音域以上に声の特徴に差が出る．また男女間を比較した時には低音域に差が分布する．本実験では周波数を帯域ごとに見ることができる MFCC(Mel Frequency Cepstral Coefficient)を用いて比較を行うことで，声の特徴の差を確認する．その差分を可視化したものを図 1〜5 に示す．これらの図では人の声の周波数帯域である 0-2000Hz 程度の結果を示しており，白に近いほど差が大きいことを示している．縦軸は周波数[Hz]，横軸は時間軸[s]を示している．Source は変換前の音声，Target は変換先の音声，Result は深層学習によって変換された音声である．Result が Target にどれだけ近くなるかを確認することで，提案手法の評価を行う．1 対 1 の声質変換の実験では，Source は男性の音声，Target は女性の音声とした．図 1 は 1 対 1 で変換した時の Source と Target の比較，図 2 は Result と Target の比較の図である．図 1 では，低音域に多く差が見られるが，図 2 では発話の終了時以外では，図 1 より低音域に差は見られない．そのため，Result は Target に近づいたと考えられる．

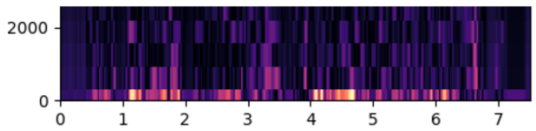

図 1 Source と Target の比較

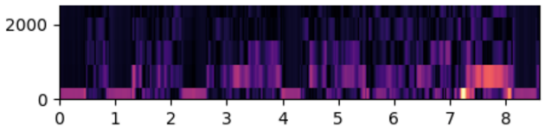

図 2 Result と Target の比較

1 対多の声質変換の実験では，Source は男性の音声，Target は 5 人の女性の音声とした．この実験では，Target に含まれる女性話者 A，B との比較を行った．図 3 は 1 対多で変換した時の Result と女性話者 A の比較の図である．女性話者 A は 1 対 1 の変換で用いた女性の音声である．図 1 に比べると低音域に差が確認できず，男性の特徴は出ていない．加えて，図 3 は中音域に差が確認できる．そのため，Result は Source である男性の特徴を持たず，女性話者 A とも異なる特徴を持つ音声であると考えられる．

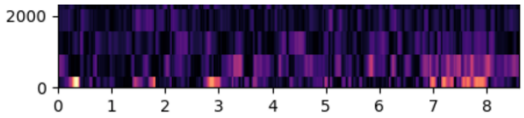

図 3 Result と女性話者 A の比較

図 4 は 1 対多で変換した時の Source と女性話者 B の比較，図 5 は Result と女性話者 B の比較の図である．図 4 では，低音域と中音域に差が見られる．図 5 では，図 4 に比べて大きな差は見られないが中音域に差が見られる．そのため，女性話者 A との比較と同様に，Result は Source である男性の特徴を持たず，女性話者 B とも異なる特徴を持つ音声であると考えられる．また，Target に含まれる他の 3 話者の女性においても同様に女性の声ではあるが，各 Target とも異なる特徴を持っていることがわかった．

以上の結果より，提案手法による声質変換によって，Source である男性の音声から Target である女性ドメインの音声への 1 対多のドメイン変換を行うことができた．つまり，1 対多声質変換によって 1 人の男性の声を特定の誰かではない女性というドメインの声に変換が可能であることがわかった．今後，これを多対多に拡張することで，匿名性の高い声質変換を目指す．

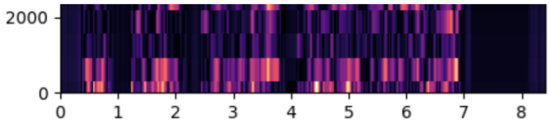

図 4 Source と女性話者 B の比較

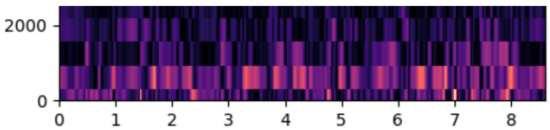

図 5 Result と女性話者 B の比較

参考文献

[1] 総務省｜行政機関・独立行政法人等における個人情報の保護｜＜3　個人情報の該当性＞ https://www.soumu.go.jp/main_sosiki/gyoukan/kanri/question03.htmlion03.html.

[2] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo: CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion, In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2019.

[3] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros: Unpaired image-to-image translation using cycle-consistent adversarial networks, in Proc. ICCV, pp. 2223–2232, 2017.

[4] Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari: “JVS corpus: free Japanese multi-speaker voice corpus,” arXiv preprint, 1908.06248, Aug. 2019.