CE-001

CycleGAN-VC 単一モデルによる音声匿名化変換の検討
A study on voice conversion based on single CycleGAN-VC model for anonymizing speaker

近藤 伊佐直[1)]　西村 竜一[2)]
Isanao Kondo　Ryuichi Nisimura

1　はじめに

近年，様々な教育・研究機関において，グループワークや問題解決型学習（PBL）の導入が進んでいる．グループディスカッションなどの複数人での対話を収録したデータを共有する際のプライバシーの保護や，グループワークにおける相互評価の公平性の担保などを目的に，収録データの話者匿名性の確保が必要となっている．また，2017 年個人情報保護法改正により，匿名加工情報を本人の同意を得ることなく利活用できるようになったことで，対話の音声に匿名加工の処理を適用するために必要な技術の需要は高まると考えられる [1].

音声の匿名化の方法として，音声認識によって発話をテキストに変換し，そのテキストを入力とする音声合成器によって音声を再合成する方法などが考えられる．本研究では，声質変換処理によって発話を直接的に別の音響特徴を持つ音声に変換する手法を採用することで，音声の匿名化処理を検討する．これは，グループワークの相互評価など，発話を分析・評価する際は，発話を文字に起こした言語情報だけでなく，発話に含まれる声量や速度，イントネーションといったパラ言語情報や非言語情報も重要であると考えるためである．また，個人情報を復元して特定の個人を再識別することができないように匿名化するには，単純な声質変換処理のみの適用では不十分である．

そこで，本研究では，変換前の話者を特定できないように音響的な変換をする声質変換処理システムの実現を目指す．開発システムでは，変換後の音声の発話内容が十分に聞き取れる品質であることが重要な要件となる．変換後の音声がそれぞれ別の話者の音声であることを識別可能な状態に保つこと，複数話者の音声を一括で変換できることも開発システムの実用化では重要である．

2　開発システム

開発システムである「マイクロホンアレイを入力とする CycleGAN-VC[2] を用いた声質変換処理システム」の概要と構成について述べ，その後，システムの内部処理を概説する．

2.1　システム概要

開発システムは，グループワークなどの対話の場面で音声を収録し，収録された複数人の音声を，別の特徴を持つ音声にそれぞれ変換する声質変換処理システムである．

本システムは，マイクロホンアレイシステムを用いて対話の音声を収録し，収録された各音声に対して声質変換処理を適用した音声を出力する．声質変換処理では，音声の特徴量を抽出したのち，声質に相当する特徴量を非線形変換し，変換後の特徴量を使用して音声を再合成する．変換後音声に含まれる発話の内容（言語情報）・時間，発話タイミングのパラメータは操作しないため，元の音声と同じ特徴を有する．

本システムは主に，グループワークなどの複数人で机を囲って行う対話の環境での使用を想定している．このため，机の中央にマイクロホンアレイシステムを設置し，効率よく話者の音声を収録する．

2.2　システム構成

開発システムは，以下の 2 つの処理部に大別される．

- 音声収録部：マイクロホンアレイシステムを用いて音声の収録を行う．
- 声質変換部：収録した音声に対して声質変換処理を行う．

開発システムを使用する前の準備段階と使用する段階の処理手順を以下に示す．なお，処理 1〜処理 3 が準備段階である．声質変換で使用するデータセットを変更した際には，改めて処理 1〜処理 3 を実行し，ネットワークを再構築する．一方，処理 4 以降が，使用時の手順である．

処理 1： 対話を行う話者の音声でデータセットを構築する．
処理 2： データセットから，声質変換ネットワークの学習に用いる特徴量を抽出する．
処理 3： 抽出した特徴量を用いて声質変換ネットワークを学習する．
処理 4： マイクロホンアレイシステムと収録処理のコントローラとなる RaspberryPi を机の中央に設置する．
処理 5： 机の周囲を話者で囲み，対話を収録する．
処理 6： 収録された音声を学習済みのネットワークを含む声質変換部に入力する．
処理 7： 変換され，別の特徴を持った音声が出力される．

処理 1 で使用するデータセットが，処理 6 で変換に使用する話者の音声データを含んでおらず，対話を行う話者とは異なる第三者の音声を入力とする場合でも同様に声質変換が行われる．この場合，第三者の音声データセットを用いて事前にネットワークを学習しておき，本システムを使用する際には，その既存のネットワークによる声質変換処理を適用することになる．使用のたびにネットワークを学習する仕様とはしていない．

音声収録部では，到来音声に対する 360 度任意の方向検知（マイクロホンアレイに任意の方向から入力された音声に対し，その音源のある方向を検知する処理）と，収録音声へのビームフォーミング（特定の方向からの音声を強調し，その他の方向からの音声を抑圧する処理）の適用ができるマイクロホンアレイモジュール

1）　和歌山大学大学院システム工学研究科
2）　和歌山大学データ・インテリジェンス教育研究部門

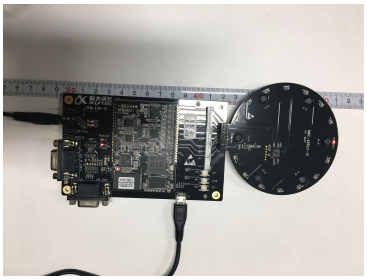

図 1　マイクロホンアレイモジュール XFM10621

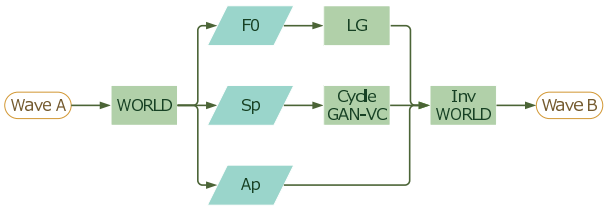

図 2　声質変換部の構成

XFM10621（iFLYTEK 製）を用いた（図 1）．

2.3　声質変換部の内部処理

声質変換部では，CycleGAN-VC[2] に基づき，変換システムを構築した．CycleGAN-VC は，2 話者間の声質変換をノンパラレルデータで行うことができる深層学習手法の一種である．[2] の手法では，音声分析合成システムの WORLD[3](D4C edition[4]) によって推定された基本周波数とスペクトル包絡の特徴量を変換する．

今回，基本周波数の推定において，WORLD(D4C edition) で標準的に採用されている推定法の DIO[6][7] ではなく，Harvest[5] を採用した．これは，DIO が，高 SNR の音声からの基本周波数推定を実時間で実行できる計算速度重視の手法であるのに対し，Harvest は，高い耐雑音性と推定精度を重視した手法であるためである [5]．なお，開発システムは深層学習を用いており，そのネットワークの学習に基本周波数推定と比べて膨大な計算時間（約 1 日程度）を必要とするため，DIO の特徴である実時間性を有効に利用することができないことも採用しなかった理由である．

CycleGAN-VC を用いた声質変換部の構成を図 2 に図解する．図中の **WaveA** が入力される音声，**F0**，**Sp**，**Ap** がそれぞれ推定された基本周波数，スペクトル包絡，非周期性指標，**WaveB** が出力される音声に対応している．また，図中の処理 **WORLD**，**Inv WORLD** がそれぞれ WORLD(D4C edition) での分析，再合成を表し，**LG** が基本周波数の変換，**CycleGAN-VC** がスペクトル包絡の変換を表している．

基本周波数の変換手法には，[2] と同様に logarithm Gaussian normalized transformation(LG) を用いた．

入力音声の話者を，声質変換ネットワークの学習話者と同一とすることができる場合（ネットワークの事前学習が可能な場合），[2] と同様に下記の I〜IV に示す方法で基本周波数 $f0$ の変換を行う．なお，入力となる話者の音声の基本周波数 $f0_{in}$ はデータセットに含まれる変換前（ソース）話者のものである．

I. ソース話者の基本周波数 $f0_s$ の対数 $\log(f0_s)$ を算出．
II. データセットに含まれる変換目標（ターゲット）話者の基本周波数 $f0_t$ の対数 $\log(f0_t)$ を算出．
III. $\log(f0_s)$，$\log(f0_t)$ の平均 $E(\log(f0_s))$，$E(\log(f0_t))$，標準偏差 $\sigma(\log(f0_s))$，$\sigma(\log(f0_t))$ をそれぞれ算出．
IV. 式 (1) によって $f0_{in}$ を $f0_{out}$ に変換．

$$f0_{out} = \exp\left(\frac{\log(f0_{in}) - E(\log(f0_s))}{\sigma(\log(f0_s))} \times \sigma(\log(f0_t)) + E(\log(f0_t))\right) \quad (1)$$

入力音声の話者が，声質変換ネットワークの学習データセットにない場合（ネットワークの事前学習ができない場合）の基本周波数の変換手法には，手法 A 及び手法 B の 2 つを比較，検討している．手法 A では，式 (1) の入力 $f0_{in}$ に，ソース話者とは異なる話者のものを入力し，ソース話者のデータセットから算出した $f0_s$ の平均と標準偏差を用いた変換を行っている．手法 B では，下記の i〜iv に示す方法を用いた．なお，入力 $f0_{in}$ はソース話者とは異なる話者のものである．

i. 2.2 節の処理 6 における入力音声の基本周波数 $f0_{in}$ の対数 $\log(f0_{in})$ を算出．
ii. データセットのターゲット話者の基本周波数 $f0_t$ の対数 $\log(f0_t)$ を算出．
iii. $\log(f0_{in})$, $\log(f0_t)$ の平均 $E(\log(f0_{in}))$, $E(\log(f0_t))$, 標準偏差 $\sigma(\log(f0_{in}))$, $\sigma(\log(f0_t))$ をそれぞれ算出．
iv. 式 (2) によって $f0_{in}$ を $f0_{out}$ に変換．

$$f0_{out} = \exp\left(\frac{\log(f0_{in}) - E(\log(f0_{in}))}{\sigma(\log(f0_{in}))} \times \sigma(\log(f0_t)) + E(\log(f0_t))\right) \quad (2)$$

なお，現状では手法 A を採用しているが，手法 A は，基本周波数がソース話者と大幅に異なる話者の音声を入力した際に不自然な変換結果が得られる問題がある．その解決策として手法 B の導入を進めている．両手法の比較評価は，今後の研究課題とする．

3　評価

本節では，開発システムを構成する各処理部の性能評価について述べる．

3.1　評価目的

開発システムは，収録した音声の匿名化処理での利用を想定しているため，本研究において検討すべき要件として，以下の 2 項目が挙げられる．

1. 変換処理後の音声が発話内容を十分に聞き取れる品質であること
2. 聴取した際に変換処理前の話者を特定することができない変換音声となっていること

また，開発システムが複数人対話の場面での実用化を目指したシステムであるため，以下の 2 項目も検討課題である．

1. 変換後音声がそれぞれ別の話者による音声であると識別できる状態であること
2. 複数話者の音声を一括で変換可能である利便性を有すること

その上で，評価実験を設計する際には，以下の事項を考慮する必要がある．

- CycleGAN-VC ではデータセットの音声をオンラインで入力した場合など，高 SNR データの入力に対する変換品質はある程度保証されていると考えることができる．しかしながら，複数人対話の環境下で収録された音声など，クリーンなデータ以外の音声を入力した場合の品質については保証されていない．
- CycleGAN-VC はパラレルデータを必要としない声質変換手法として提案されたものであり，音声の匿名化処理を想定したものではない．そのため，先行研究では，変換後音声の匿名性は検証されていない．
- CycleGAN-VC はデータセットで与えられた 2 話者間のマッピングを行うアルゴリズムであり，データセットに含まない話者の音声（未知話者音声）を入力した場合の挙動は保証されていない．

そこで，本研究の評価実験では，以下の 2 つの項目を調査することとした．

- グループワークにおける複数人対話の環境など，本研究で想定している収録環境下での音声収録部の性能
- 匿名化を目的として開発された本システムにおける声質変換部の匿名化性能

また，未知話者音声を入力とした場合の変換性能について，現状のシステムにおける品質を明らかにすることも，評価目的の一つとした．

3.2 実験条件

評価実験は，実験 1 から実験 3 で構成される．各実験条件を以下に示す．

実験 1： 変換音声の文章了解性の評価（変換音声の発話内容を聴き取ることができるか）
実験 2： 変換音声の話者認識可能性の評価（変換音声から変換前の話者を認識することができるか）
実験 3： 変換ネットワークの学習に使用していない話者の発話を入力した場合の文章了解性の評価（未知の音声を入力した場合にも変換は正しく動作するか）

実験 1，実験 2，実験 3 の実験試料には，JVS (Japanese versatile speech) corpus[8] に含まれる音声データを用いた．ネットワークの学習に用いたデータは，2 話者分（話者 1，話者 2），計 75 発話のノンパラレルデータのデータセットである．ささやき声，裏声は，今回の実験から除外した．なお，話者 1 は男性，話者 2 は女性である．

3.3 実験 1（文章了解性）

実験 1 では，音声収録部の評価を目的に，変換音声の文章了解性について主観評価を行った．下記の条件 (a)〜条件 (e) の 5 種類の条件で収録した音声を用意し，開発システムで変換を行った．5 条件で収録した音声それぞれに対し，事前に話者 1→ 話者 2 のマッピングを学習したネットワークによる変換処理を適用した．条件 (a)〜条件 (c) に関しては，前処理として同一の環境下で話者 1 の音声をスピーカーから再生し，3 種類の条件で再収録をした音声データを声質変換部の入力としている．

(a) ビームフォーミングを音源方向に正しく設定して収録した音声
(b) ビームフォーミングを音源と 90 度異なる方向に設定して収録した音声
(c) 単独の MEMS (Micro Electro Mechanical System) マイクを用いて収録した音声
(d) スピーカーで再生することなくオンラインで入力した話者 1 の音声
(e) 条件 (a) と同一の方法で収録した，データセットに含まれない話者 3 が発話した音声

出力音声の文章了解性を協力者 9 名の主観評価で調査した．ここで，協力者は，変換後音声を聴取したのち，以下の選択肢から該当する一つを選択して回答するようにした．

A： 全ての単語を聞き取れた
B： 8 割程度の単語を聞き取れた
C： 半分程度の単語を聞き取れた
D： 2 割程度の単語を聞き取れた
E： 全ての単語を聞き取れなかった

表 1 に，条件 (a)〜条件 (e) の変換後音声に対して，協力者 9 名が回答した A〜E の集計結果を示す．条件 (a) と条件 (c) を比較すると，条件 (c) では D が最頻となっており，最もよくて B であったのに対し，条件 (a) ではほぼ全員がほとんどの単語を聞き取れていることが分かる．この結果は，開発システムの音声収録部のマイクロホンアレイが有効に機能していることを示す．次に，条件 (b) の結果を見ると，E が最も多く C 以上が存在しない．これは，ビームフォーミングで設定された方向以外からの音響信号をマイクロホンアレイが抑圧したため，収録音声の品質が劣化したことが原因だと考えられる．最後に，条件 (e) の結果を見ると，条件 (a) や条件 (d) には劣るものの，すべての協力者がほとんどの単語を聞き取れていることが分かる．この結果は，事前に話者 1→ 話者 2 のマッピングを学習したネットワークを用いた声質変換であるにもかかわらず，ネットワークの学習に用いたデータセットに含まれない第三者である話者 3 の入力音声に対しても高い品質で変換音声を出力できたことを示している．

表 1　文章了解性実験の集計結果

	(a)	(b)	(c)	(d)	(e)
回答 A	8	0	0	7	3
回答 B	1	0	1	2	6
回答 C	0	0	2	0	0
回答 D	0	3	4	0	0
回答 E	0	6	2	0	0

3.4 実験 2（話者認識可能性）

実験 2 では，声質変換部の評価を目的に，下記の異なる手法で変換した各音声に対し，話者認識可能性を協力者 9 名の主観評価で調査した．

- 開発システムの声質変換ネットワークで変換した

図 3　音声聴取の様子

表 2　実験 3（未知話者音声の文章了解性）の集計結果

回答	個数
A	26
B	3
C	1
D	0
E	0

音声

- WORLD(D4C edition) で抽出した特徴量を線形変換して再合成した音声

図 3 は，実験 2 の聴取の様子である．画像中央のスピーカーより音声が再生される．

各手法による変換後の音声を，変換前話者をブラインド状態にして再生し，続いて変換前話者を含む 5 人の異なる話者の発話を順に再生した．協力者には，最初の音声の話者が，後続の音声 1〜5 のどの話者と同一かを回答するように求めた．なお，「判らない」を選択することも可能にした．

変換された音声を聴取し，変換前話者を識別できた協力者はどちらの手法もともに 2 名で，正答率は 22%となった．この結果より，声質変換部の処理によって変換された音声から，聴取による変換前話者の識別が難しいことを確認した．

3.5　実験 3（未知話者音声の文章了解性）

実験 3 では，データセットに含まれない第三者の音声（未知話者音声）を入力とし，その出力音声の文章了解性を主観評価で調査した．話者 1→ 話者 2 のマッピングを学習したネットワークに対して，ネットワークの学習に使用していない 30 人の話者による発話を発話内容が被らないように選定して入力した．出力を聴取して実験 1 と同一の選択肢により主観で評価した集計結果を表 2 に示す．話者 1 から話者 2 への声質変換を行うネットワークに対して，ネットワークの学習に含まれない未知話者の 30 発話のうち 26 発話において全ての単語を聴き取りできていることが分かる．

これは，実験 1 の表 1 に示した条件 (e) に対する文章了解性が高かったという結果が，偶然によるものでないことを示唆している．これによって，開発システムの声質変換部は，第三者の音声データセットで学習したネットワークを用いることで，対話を行う話者による事前の音声収録とその音声を用いたネットワークの学習を行わなくとも，高品質に声質変換を行うことが可能であることが確認できた．

4　おわりに

本研究では，グループワークなどの対話の場面において対話音声を収録し，匿名化処理のために声質変換をするシステムを開発し，実験によって評価を行った．声質変換部の処理に，マイクロホンアレイを備えた音声収録部での収録音声を入力とすることで，MEMS マイク単体で録音した音声を変換するよりも高品質な変換音声を出力できることを主観評価で確認した．また，声質変換部によって変換された音声は，変換前の話者の認識可能性が低いことを示した．さらに，未知話者音声を入力した場合においても，高品質な声質変換が可能であることを確認した．これは，事前に参加者の音声を収録してネットワークを学習することなく声質変換が可能となることを示唆しており，開発システムにおいて共通の学習済みネットワークを用いた複数話者の一括声質変換を可能とする利便性に繋がる結果である．

今後の課題として，単一の学習済みネットワークを用いた未知話者音声の変換における話者認識可能性の詳細な評価や，対話参加者の音声それぞれに対する変換後音声がそれぞれ別の話者の音声であることを識別可能である状態に保つこと（変換前後の話者弁別性の保持）についての分析が必要である．また，実験 3 においては，未知話者音声の入力に対する変換後音声の文章了解性が高いという結論に至ったが，実験協力者の人数が少ないため（1 名），協力者を追加した実験が必要である．その上で，2.3 節で述べた 2 つの基本周波数の変換手法の性能評価比較を行う予定である．

謝辞

本研究は，JSPS 科研費 JP18K02862 の助成を受けて実施したものである．

参考文献

[1] 第 1 部第 3 章第 2 節スマートフォンの普及と ICT 利活用, 情報通信白書, 平成 28 年版, p.181, 2016. https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/pdf/n3200000.pdf (2020 年 6 月 19 日アクセス確認)

[2] Takuhiro Kaneko and Hirokazu Kameoka, CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks, Proc. EUSIPCO, pp. 2114–2118, 2018.

[3] M. Morise, F. Yokomori, and K. Ozawa, WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE transactions on information and systems, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[4] M. Morise, D4C, a band-aperiodicity estimator for high-quality speech synthesis, Speech Communication, vol. 84, pp. 57–65, 2016.

[5] 森勢将雅, 高い雑音耐性と推定精度を両立する基本周波数推定法の提案と評価, 電子情報通信学会技術研究報告, vol. 116, no. 378, pp. 107–112, 2016.

[6] 森勢将雅, 河原英紀, 西浦敬信, 基本波検出に基づく高 SNR の音声を対象とした高速な F0 推定法, 電子情報通信学会論文誌, vol. J93-D, no. 2, pp. 109–117, 2010.

[7] M. Morise, H. Kawahara and H. Katayose, Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech, Proc. AES 35th International Conference, CD-ROM, 2009.

[8] 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.