［研究助成（A）］

連合学習に基づく多話者音声変換のユーザ参加型学習

Human-in-the-loop many-to-many voice conversion based on federated learning

2221013

研究代表者　東京大学
大学院情報理工学系研究科　助　教　齋　藤　佑　樹

［研 究 の 目 的］

音声は，人間同士の最も自然なコミュニケーション手段である。人間は，コミュニケーションにおける自らの意図や感情といった様々な情報を音声により相手に伝えられる。しかし，個人が発声可能な音声は，発声器官などの身体的制約により大きく制限され，更に，加齢による衰えや事故や病気による後天的要因により変化しうる。これらの制約を計算機で緩和できれば，音声アバターによる自己表現拡張や，発声器官に障害を有する話者の支援など，新たな娯楽の創出や医療福祉への応用が可能となる。本研究では，計算機を通じた音声コミュニケーション拡張を目指し，任意話者の音声の特徴を，別の任意話者のものに変換する多話者音声変換（voice conversion: VC）に取り組む。

［研究の内容，成果］

1. 背景

多様な話者の高品質な音声が学習データとして用意できる場合には，deep neural network（DNN）を用いたVC技術が高品質な変換を実現できる[1]。近年の深層生成モデリング技術の進展や，学術界・産業界による大規模音声言語データベース（コーパス）の整備に恩恵を受け，変換可能な話者に関する制約が緩和され，歌唱表現の拡張や動画コンテンツ制作など，ユーザの声を対象としたVCアプリケーションが実現されるようになりつつある。

一方で，現状のVC技術は，研究開発者が構築した深層学習モデルに基づくVCアプリケーションをユーザに提供する単方向型アプローチ（図1(a)）が主流であり，多様なユーザによる入力音声に対する変換精度を保証できない。もしユーザが所有するデータを用いたVCモデルの逐次更新が可能となれば，モデルの学習時と推論時における入力データのドメインの違いに対して柔軟に適応可能な双方向型アプローチのユーザ参加型多話者VC技術（図1(b)）が実

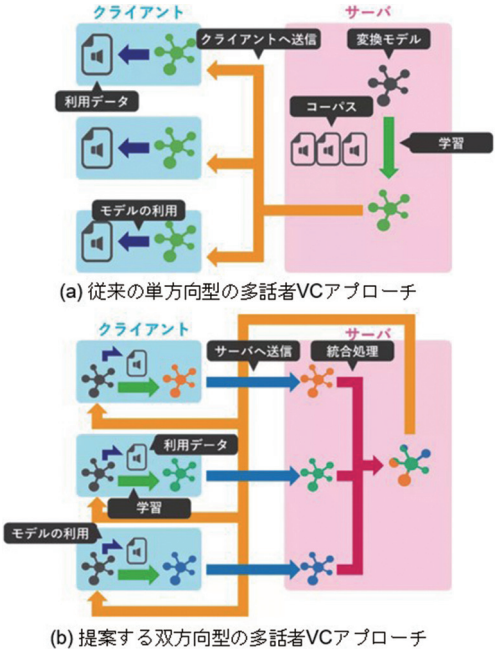

図1　多話者VCにおける（a）従来型，（b）提案型のアプローチ（研究成果1記載の図を一部修正）

現できる。しかし，音声データは本質的に発話内容と声質から発話者が特定される可能性を孕む個人情報であり，研究開発者が所有するサーバにユーザのデータを集約させる行為にはプライバシー漏洩のリスクが伴う。

本研究では，分散機械学習手法の一種である連合学習に基づく多話者 VC モデルの学習法を提案する。提案法では，研究開発者とユーザの間で VC モデルのパラメータのみを送受信することで，ユーザが所有する音声データのプライバシーを保持しつつ，より多様な話者の声質を再現可能な多話者 VC モデルを逐次的に学習する。本研究では，提案法の Proof of Concept を検証するために，VC モデルとして最先端の StarGANv2-VC[1] を用いた多話者 VC の連合学習を検討する。日本語の多話者コーパスを用いた実験的評価により，提案法が従来型のデータ集約型学習法と同程度の話者類似性（変換先話者にどの程度類似してるか）を達成しうることを示す。

2. 手法

2.1 ベースラインモデル：StarGANv2-VC

StarGANv2-VC[1] は，音声特徴量から話者非依存な潜在変数を抽出する Encoder，音高の潜在変数を抽出する F0 Network，話者の潜在変数を抽出する Style Encoder，話者の潜在変数の確率分布をモデル化する Matching Network，これら 3 つの潜在変数から変換音声の特徴量を生成する Decoder，入力音声の話者識別と，入力音声が人間の肉声か変換音声かの真偽を判定する Discriminator から構成される。本研究では，これらのモデルのうち，F0 Network のモデルパラメータ（DNN の結合重み・バイアス）は，高品質多話者コーパスを用いた事前学習により最適化されると仮定し，それ以外のモデルパラメータを複数のユーザが各々所有するデータにより更新し，その結果を統合する連合学習を想定する。学習アルゴリズムや DNN アーキテクチャの詳細は，文献[1] を参照されたい。

2.2 連合学習

連合学習[2] はクライアント・サーバ方式に基づく分散機械学習フレームワークの一種であり，学習に参加するクライアントが所有するデータのプライバシーを保護しつつ，単一の機械学習モデルを逐次的に学習する。連合学習は，以下に示す Round と呼ばれる一連の処理の反復として表現される。

(1) サーバはランダムに選択した複数のクライアントに自身が所有するモデルのパラメータを送信する。
(2) クライアントは各自が所有するデータを利用し，サーバから受信したモデルを学習する。
(3) クライアントで一定の期間（local epoch）だけ学習したモデルをサーバ側に送信する。
(4) サーバに集約させたモデルを特定の処理に基づいて一つのモデルへと統合し，サーバの所有するモデルを更新する。この Round の繰り返しによって，個々のクライアントの所有するデータを用いた学習の結果を取り入れ，各クライアントのデータに対応可能なモデルを共同的に学習する。

本研究では，最も広く使われているモデル統合方である FedAvg[3] を用いた。FedAvg では，各クライアントが学習を終了した後のモデルパラメータに対してデータ数で重み付けされた平均を計算することで，統合後のモデルパラメータを計算する。

2.3 提案法の問題設定

1 節で説明したように，深層学習に基づく従来の多話者 VC 技術では，まず，研究開発者が高品質な多話者コーパスを用いて単一の多話者 VC モデルを構築し，そのモデルを用いた VC アプリケーションをユーザに提供する。ユーザはこの VC アプリケーションを利用するにあたり，必要に応じて自身が所有する音声データを用いて学習済み VC モデルを fine-tuning することも可能である。このような単方向型の VC

アプローチでは，ユーザによる fine-tuning の結果を研究開発者側にフィードバックし，より多様な話者の声を変換できるようにすることは困難である。

この問題に対し，ユーザ参加型の多話者 VC モデル学習を実現するために，本研究では連合学習に用いるデータセットを以下の 2 つに分割する。

(1) Anchor データ D_{Anc}：連合学習に参加するすべてのユーザがアクセス可能なデータセット

(2) Client データ $D_{\mathrm{Cli}}^{(k)}$：k番目のユーザだけがアクセス可能なデータセット

D_{Anc}と$D_{\mathrm{Cli}}^{(k)}$は，それぞれオンラインで誰でも入手可能な公共の多話者音声コーパスとプライバシー保護が必要なユーザ所有データに対応する。すなわち，提案法ではk番目のユーザはこれらの和集合$D_{\mathrm{Anc}} \cup D_{\mathrm{Cli}}^{(k)}$を用いて自身の計算機環境で VC モデルを学習する。これらの 2 つのデータセットを用いることで，VC モデルは全てのユーザで共通の Anchor 話者特徴と，各ユーザが所有するデータの話者特徴を分離してモデル化することができる。

提案法の問題設定では，各ユーザが所有するデータ分布の独立同分布（independent and identically distributed: iid）性はもはや仮定できない。なぜなら，図 2 に示すように，各$D_{\mathrm{Cli}}^{(k)}$の分布は互いに異なると想定されるためである。このような non-iid データを用いた連合学習は，ある特定のユーザが所有するデータセットへの過学習といった問題の原因となりうるため，本研究では，連合学習の各 Round でサーバ上のモデルからの乖離に対する正則化項を導入する FedProx[4]を用いる。FedProx の正則化項は，あるハイパーパラメータ$\mu \in [0, 1]$を用いて$\mu \|\boldsymbol{w}_r^k - \boldsymbol{w}_r\|^2$と表される。ここで，$\boldsymbol{w}_r, \boldsymbol{w}_r^k$はそれぞれ Round r でのサーバ上のモデルパラメータとk番目のユーザが所有するモデルパラメータである。

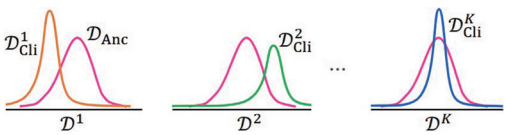

図 2　提案法における学習データの non-iid 性

3. 実験的評価

3.1 実験条件

本研究では，StarGANv2-VC の学習を単一の中央サーバに全データを集約させて行う従来法と，連合学習によってデータを各クライアントに分散させた状態で行う提案法を比較した。実際の連合学習ではサーバ・クライアントは異なる計算機やエッジ端末であることを想定するが，実験の簡略化のため，同一の計算機内に仮想的なサーバ・クライアントを用意して連合学習を実施した。

StarGANv2-VC の実装は，当該論文[1]の著者により公開されているオープンソース実装を用いた。DNN のアーキテクチャはすべてこの実装と同じであり，音声特徴量（メルスペクトログラム）から音声波形データを合成するボコーダには事前学習済みの Parallel WaveGAN [5]を利用した。

データセットとして，日本語の JVS コーパス[6]に含まれる男性話者と女性話者をそれぞれ 20 名ずつ選択した。これらの各話者の音声を 100[ms]を基準に無音区間を削除し，5 秒ごとに分割したものを一つのデータ単位とした。訓練データ，検証データ，テストデータの数はそれぞれ 3284, 411, 411 であった。Anchor と Client の話者数はそれぞれ 10, 30 とした。Anchor 話者の音声データと Client 話者の音声データが均等になるように各クライアントにデータを割り当てた。

連合学習のクライアント更新時の local epoch 数は 10，バッチサイズは 10 とした。学習時の optimizer として AdamW を用いた。ただし，連合学習に基づく提案手法では，optimizer は各クライアントごとに異なる内部パラメータが用いられるものとし，各 Round

におけるクライアントへのモデル配布と同時に optimizer の内部パラメータを初期化するものとした。FedProx を用いる場合のハイパーパラメータ μ は 1 とした。

以降の評価は，VC の入出力話者が Anchor 話者，Client 話者である場合の計 4 通りの変換対についてそれぞれ分けて実施した。各変換対の学習の難易度は異なり，特に，Client 話者同士の変換は連合学習で直接的に学習不可能であるため，最も困難であると考えられる。

3.2 客観評価

変換音声の話者類似性の客観評価指標として，音声由来の話者埋め込みベクトルの一種である x-vector [8] のコサイン類似度（x-vector cossim）を用いた。評価結果を図 3 に示す。提案法の性能を（1）連合学習で統合処理に用いるクライアント数，（2）連合学習の Round 数，（3）non-iid 性に対処するための FedProx 正則化項の有無について調査した結果から，以下の傾向が明らかになった。

- 複数クライアントによるモデルパラメータ統合により，客観評価指標は一貫して改善した。
- Round 数を増やすことで客観評価指標は改善し，特に，Client 話者を変換先話者とした VC においてその改善が顕著であった。この傾向は，図 4 にも示されている。
- FedProx を導入することによる改善は，Client 話者を変換先とした VC において顕著であった。

これらの結果は，（1）複数クライアントによる連合学習は，VC モデルが再現可能な話者の多様性を増加させること，（2）十分な Round の反復により，直接学習が不可能な Client 話者

Model			Anc→Anc	Anc→Cli	Cli→Anc	Cli→Cli
# Cli	# Round	FedProx?				
1	200	No	0.48 ± 0.19	0.23 ± 0.39	0.48 ± 0.18	0.23 ± 0.39
1	400	No	0.51 ± 0.17	0.37 ± 0.40	0.51 ± 0.17	0.38 ± 0.40
3	200	No	0.53 ± 0.19	0.26 ± 0.41	0.54 ± 0.20	0.27 ± 0.41
3	400	No	0.53 ± 0.16	0.39 ± 0.37	0.53 ± 0.17	0.41 ± 0.36
1	200	Yes	0.53 ± 0.20	0.31 ± 0.38	0.53 ± 0.21	0.32 ± 0.38
1	400	Yes	0.52 ± 0.17	0.41 ± 0.36	0.52 ± 0.18	0.43 ± 0.35
3	200	Yes	**0.56 ± 0.19**	0.34 ± 0.36	**0.56 ± 0.19**	0.35 ± 0.36
3	400	Yes	0.53 ± 0.18	0.42 ± 0.35	0.53 ± 0.18	0.43 ± 0.34
3	600	Yes	0.55 ± 0.18	0.44 ± 0.35	0.55 ± 0.19	0.46 ± 0.34
3	800	Yes	0.55 ± 0.19	**0.46 ± 0.35**	0.55 ± 0.19	**0.48 ± 0.34**

図 3　x-vector cossim の評価結果

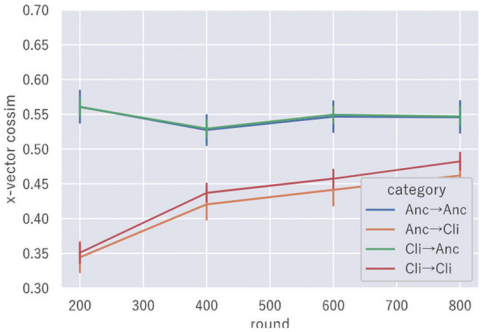

図 4　Round 数に対する客観評価値の遷移

同士の VC も実現可能となること，（3）多話者 VC の連合学習においても，学習データの non-iid 性に対処する正則化は有効に機能することを示した。以降の評価では，Client 話者への変換性能が最も高い {#Cli＝3, #round＝800, FedProx あり} で学習された VC モデルを用いた。

3.3 主観評価

変換先話者との類似性に関するプリファレンス XAB テストと変換音声の自然性に関するプリファレンス AB テストを実施した。これらの評価では，まず各 Client 話者のテストデータから無作為に 3 つを取得し，自身を除く他の全ての話者へと変換することで，ユーザからの入力音声を扱う VC の設定を模擬した。評価セットの数は 2（XAB or AB）×2（Cli-to-{Anc, Cli}）×4（{male, female}-to-{male, female}）＝16 であり，各評価につき 50 名ずつの評価者を Lancers によるクラウドソーシングで募集した。すなわち，合計の評価者数は 800 名である。XAB テストでは，評価者はまず変換先話者の音声を参照し，その後に再生される従来法（連合学習なしで 700 エポック学習された StarGANv2-VC）と提案法による変換音声サンプル対のうち，どちらのサンプルがより変換先話者に類似しているかどうかを回答した。この際，評価者に提示する音声サンプル対の順番はランダムシャッフルした。AB テストでは，参照音声は提示せず，提示された従来法・提案法の音声サンプルのうち，どちらがより人間らしく自然に発話しているかどうかを回答した。

(a) プリファレンスXABテストの結果

VC setting	Conventional	Proposed
Cli(F) → Anc(F)	0.448	**0.552**
Cli(F) → Anc(M)	0.484	0.516
Cli(M) → Anc(F)	0.432	**0.568**
Cli(M) → Anc(M)	0.490	0.510
Cli(F) → Cli(F)	0.520	0.480
Cli(F) → Cli(M)	0.472	0.528
Cli(M) → Cli(F)	0.510	0.490
Cli(M) → Cli(M)	0.480	0.520

(b) プリファレンスABテストの結果

VC setting	Conventional	Proposed
Cli(F) → Anc(F)	0.526	0.474
Cli(F) → Anc(M)	**0.574**	0.426
Cli(M) → Anc(F)	**0.584**	0.416
Cli(M) → Anc(M)	**0.604**	0.396
Cli(F) → Cli(F)	0.366	**0.634**
Cli(F) → Cli(M)	0.368	**0.632**
Cli(M) → Cli(F)	0.326	**0.674**
Cli(M) → Cli(M)	0.408	**0.592**

図5　主観評価結果

各評価で10個の音声サンプル対を評価させた。

XABテスト，ABテストの評価結果をそれぞれ図5(a)，(b)に示す。図中の（M）と（F）はそれぞれVC変換元話者の性別が男性，女性であることを意味する。太字のスコアは比較対象よりも有意（$p<0.05$）に高いことを示している。

図5(a)より，全てのVC設定において，提案法は従来法よりも同程度か，有意に高い話者類似性を達成した。すなわち，多話者VCモデルの連合学習でユーザが所有するデータも活用することにより，モデルが再現可能な話者の多様性を改善できることが主観的にも確認された。この改善の要因の一つとして，従来法では40名の話者間のVCを単一のモデルで一度に学習するが，提案法では10名のAnchor話者＋1名のClient話者という複数の小規模VCタスクに分散させることで，学習が容易になったということが推測される。

一方で，図5(b)より，変換音声の自然性については，VC設定によって異なる傾向が示された。具体的には，提案法による変換音声の自然性は，Client-to-Anchor VCの設定では従来法より劣り，Client-to-Client VCの設定では従来法を上回る結果となった。この要因の一つとして，Clientデータと Anchorデータの不均衡さが挙げられる。すなわち，Client話者数（30）はAnchor話者数（10）よりも多く，少数派のAnchor話者の変換は多話者VCモデルが十分に学習できなかったということが推測される。

4. 結論と今後の課題

本研究では，プライバシーを保持しつつユーザが所有する音声データを活用して多対多VCモデルを学習を手法を提案した。実験的評価の結果から，データを分散させた状態では直接的に学習不可能であるClient同士のVCについても，データを集約させて学習を行う従来法と同程度の変換音声の話者類似性を達成できることを示した。

［今後の研究の方向，課題］

今後はAnchor/Clientデータセットの不均衡さが連合学習に及ぼす影響や，実際に異なる端末間での連合学習を実施した場合の提案手法の振る舞いなどを詳細に調査する。また，悪意のあるユーザが連合学習に参加し，完全なノイズなどの学習に悪影響を与えるケースを想定した学習も検討する。

［参考文献］

［1］ Y. Li et al., “StarGANv2-VC: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion,” in Proc. INTERSPEECH, 2021.

［2］ J. Konečný et al., “Federated learning: strategies for improving communication efficiency,” in NIPS Workshop on Private Multi-Party Machine Learning, 2016.

［3］ H. McMahan et al., “Communication-efficient learning of deep networks from decentralized data,” in Proc. AISTATS, 2017.

［4］ T. Li et al., “Federated optimization in heterogeneous networks,” in Proc. AMTL, 2019.

［5］ R. Yamamoto et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. ICASSP, 2020.

［6］ S. Takamichi et al., “JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research,” Acoustical Science and Technology, vol. 41, no. 5, 2020.

［成果の発表，論文など］

1） 平井龍之介，齋藤佑樹，猿渡洋，“Fed-Star GANv2-VC：連合学習を用いた多対多声質変換，”情報処理学会研究報告，2023年.

2） Ryunosuke Hirai, Yuki Saito, and Hiroshi Saruwatari: “Federated Learning for Human-in-The-Loop Many-to-Many Voice Conversion,” in Proc. 12th ISCA Speech Synthesis Workshop, Aug. 2023.（Accepted）