Fed-StarGANv2-VC：連合学習を用いた多対多声質変換

平井 龍之介[1,a)]　齋藤 佑樹[1,b)]　猿渡 洋[1]

概要：本稿では，連合学習を用いたユーザ参加型の多対多声質変換モデル学習法を提案する．従来の多対多声質変換技術は，多数話者の音声を含むデータセットを用いて声質変換モデルを学習する．しかし，学習されたモデルが多種多様なユーザによる入力音声に対して高品質な声質変換を実現する保証はない．提案手法では，高品質な多対多声質変換を実現する StarGANv2-VC モデルを研究開発者とユーザが協同的に学習し，ユーザが所有する音声データのプライバシーを保護しながら，より多様な話者の音声を変換可能な深層学習モデルを構築する．実験的評価の結果より，提案手法が従来の非分散型学習法と同程度の話者類似性を達成しうることを示す．

Ryunosuke Hirai[1,a)]　Yuki Saito[1,b)]　Hiroshi Saruwatari[1]

1. はじめに

音声は人間がリアルタイムに情報交換を行うにあたって用いられる主要な手段の一つである．音声にはその発話内容にとどまらず，個人性や感情などの非言語・パラ言語的な部分にも多くの情報を持つという特性がある．音声の持つ情報の一つに声質と呼ばれる話者の個人性を表現する非言語情報がある．声質は発話の印象を大きく左右する重要な特徴の一つである一方で，その形成は年齢や身体的特徴などに大きく依存する．声質変換 (Voice Conversion: VC) [1] はある話者の音声データについて，その発話内容を保持したまま，声質を異なる話者のものに変換する技術である．声質変換技術はかねてより人間社会と深いかかわりを持ち，話者の持つ声質の匿名化による個人情報の保護 [2] や，音声バーチャルリアリティや歌声変換 [3] などのエンターテインメント応用がその一例として考えられる．

昨今の深層学習技術の発展と，研究開発者による大規模な多話者音声データセットの構築・整備に恩恵を受け，多様な話者の音声を高品質に変換可能な技術が数多く提案されている [4], [5], [6]．一方で，現状の声質変換技術は，研究開発者が構築した深層学習モデルに基づく声質変換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，多種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声質変換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（話者のみならず収録環境 [7]）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対多声質変換技術（**図 1**(b)）が実現できる．しかし，音声データは本質的にその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中央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．

本研究は，高品質な多対多声質変換が可能な StarGANv2-VC [8] に連合学習 (Federated Learning) を導入した Fed-StarGANv2-VC を提案する．提案手法では，研究開発者とユーザの間で声質変換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対多声質変換モデルを逐次的に学習する．実験的評価では，従来の非分散型の StarGANv2-VC モデル学習と提案手法を比較し，連合学習特有の困難性である分散学習データの非独立同一分布 (non-iid) 性と，連合学習の反復回数が声質変換性能に及ぼす影響を調査する．評価結果より，提案する Fed-StarGANv2-VC が従来型の学習法と同程度の話者類似性を達成しうることを示す．

2. 関連研究

2.1 深層学習に基づく声質変換

高品質な声質変換技術の実現には，入力音声からその発

[1] 東京大学
[a)] hirai-ryunosuke@g.ecc.u-tokyo.ac.jp
[b)] yuuki_saito@ipc.i.u-tokyo.ac.jp

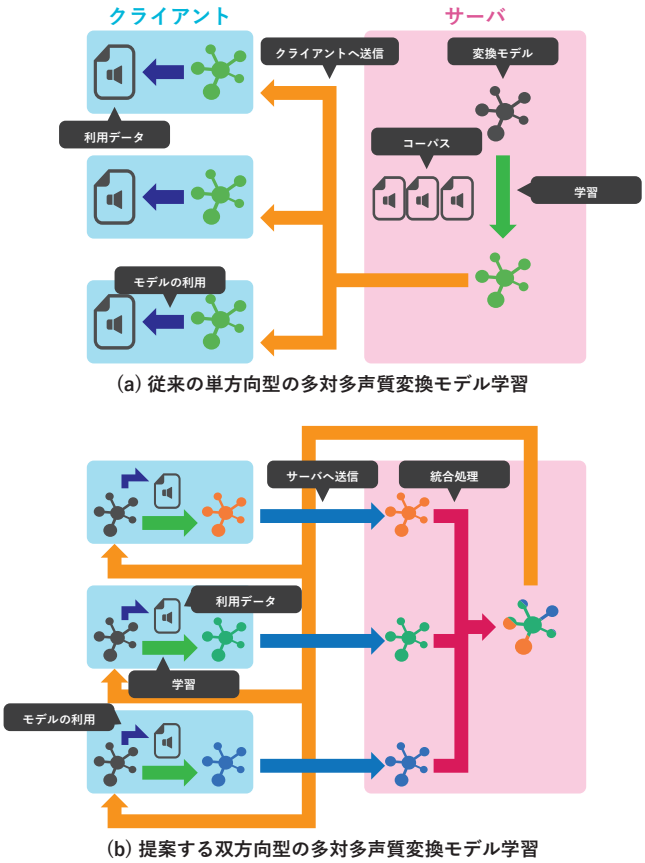

図 1 多対多声質変換における従来技術 (a) と提案技術 (b)

話内容と話者性を高精度に分離する統計モデリング手法が要求される．変換元・変換先話者による同一発話内容の音声データから構成されるパラレルデータを用いた深層学習法 [9], [10], [11] では，音声特徴量間の精緻な変換を実現する一方で，パラレルデータの収録に要するコストから，変換可能な話者対に関するスケーラビリティが低いという欠点がある．一方で，近年では深層学習モデルのドメイン敵対学習 [12] による話者非依存な特徴表現学習や，学習済み自動音声認識 (Automatic Speech Recognition: ASR) モデルに由来する発話内容の潜在変数 [13], [14] の導入などにより，非パラレルデータを用いた高品質な声質変換モデルの学習が可能になりつつある．

2.1.1 一巡一貫性損失を用いた敵対的生成ネットワークに基づく非パラレル声質変換

Goodfellow らにより提案された敵対的生成ネットワーク (Generative Adversarial Network: GAN) [15] は，データの生成器 (generator) と真贋識別器 (discriminator) と呼ばれる 2 つの深層ニューラルネットワーク (Deep Neural Network: DNN) の交互最適化に基づく生成モデルであり，近年の高品質なメディア生成技術を支える基盤となっている．Kaneko らの CycleGAN-VC [4] では，CycleGAN [16] で提案された一巡一貫性 (cycle consistency) 損失を用いて変換元・変換先話者対ごとの GAN を学習させ，高品質な非パラレルデータ声質変換を実現した．Kameoka らはこ

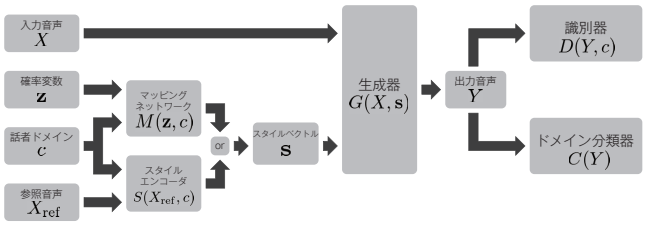

図 2 StarGANv2-VC のモデル構造

の枠組みを拡張し，入力音声特徴量からその話者を識別するドメイン分類器と，変換先話者を指定する離散的な属性情報を導入することで，単一の生成器で非パラレル多対多声質変換を実現する StarGAN-VC [5] を提案した．

2.1.2 StarGANv2-VC

Li らにより提案された StarGANv2-VC [8] は，StarGAN-VC に対して以下の改良を加え，多様な話者の高品質な非パラレル多対多声質変換を実現している．**図 2** に StarGANv2-VC のモデル構造を示す．

スタイルエンコーダによるスタイルベクトル抽出: StarGAN-VC では生成器を離散的な属性情報で条件付けするため，話者の多様な発話スタイルの再現精度に限界があった．StarGANv2-VC では，変換先の話者を指定する話者ドメイン c と当該話者の参照音声 $\boldsymbol{X}_{\mathrm{ref}}$ から，その話者の発話スタイルの連続値ベクトル表現であるスタイルベクトル $\boldsymbol{s}$ を生成するスタイルエンコーダ $S(\boldsymbol{X}_{\mathrm{ref}}, c)$ を導入し，再現可能な発話スタイルの多様性を向上させている．

マッピングネットワークによるスタイルベクトル生成: 話者の発話スタイルは確率的にゆらぐため，元来の GAN が有する，既知の確率分布からのランダムサンプリングが実現できることが望ましい．StarGANv2-VC では，前述のスタイルエンコーダに加え，Gauss 分布に従う確率変数 $\boldsymbol{z}$ から話者ドメイン c に属するスタイルベクトルを生成するマッピングネットワーク $M(\boldsymbol{z}, c)$ を同時に学習し，変換先話者の参照音声を用いることなく，スタイルベクトルのランダムサンプリングを実現する．

StarGANv2-VC で入力音声のメルスペクトログラム $\boldsymbol{X}$ を変換する生成器 $G(\cdot)$ は，以下の 3 つのサブモジュールから構成される．

(1) **エンコーダ:** $\boldsymbol{X}$ から，潜在変数 $\boldsymbol{h}_{\mathrm{x}}$ を抽出する．
(2) **F0 ネットワーク:** $\boldsymbol{X}$ から，韻律特徴量 $\boldsymbol{h}_{\mathrm{F0}}$ を抽出する．
(3) **デコーダ:** $\boldsymbol{h}_{\mathrm{x}}$ と $\boldsymbol{h}_{\mathrm{F0}}$ に加え，変換先話者のスタイルベクトル $\boldsymbol{s}$ を入力として受け取り，当該話者のメルスペクトログラム $\hat{\boldsymbol{X}}$ を復元する．

Li らの論文 [8] では，事前学習済みの Joint Detection and Classification (JDC) ネットワーク [17] を F0 ネットワークとして利用し，変換元話者の F_0 軌跡パターンが声質変換過程で不変であることを保証している．生成器 $G(\cdot)$ は，話者ドメイン c で条件付けされた真贋識別器 $D(\cdot)$ と，入

力メルスペクトログラムが属するドメイン（即ち，話者）を分類するドメイン分類器 $C(\cdot)$ に敵対するように学習される．以降の説明では，話者ドメイン $c^{(\mathrm{s})}$ に属する音声のメルスペクトログラム $\boldsymbol{X}^{(\mathrm{s})}$ からドメインを $c^{(\mathrm{t})}$ に変換する処理を $\hat{\boldsymbol{X}}^{(\mathrm{s}\to\mathrm{t})} = G(\boldsymbol{X}^{(\mathrm{s})}, \boldsymbol{s}^{(\mathrm{t})})$ と表記する．ただし，$\boldsymbol{s}^{(\mathrm{t})} = S(\boldsymbol{X}^{(\mathrm{t})}, c^{(\mathrm{t})})$ は話者ドメイン $c^{(\mathrm{t})}$ に対応するスタイルベクトルである．また，元論文 [8] での定式化における期待値演算記号 $\mathbb{E}[\cdot]$ は表記の簡略化のため省略する．

真贋識別器の損失関数は，次式で定義される．

$$\mathcal{L}_{\mathrm{adv}} = -\log D\left(\boldsymbol{X}^{(\mathrm{s})}, c^{(\mathrm{s})}\right) - \log\left(1 - D\left(\hat{\boldsymbol{X}}^{(\mathrm{s}\to\mathrm{t})}, c^{(\mathrm{t})}\right)\right) \tag{1}$$

式 (1) の第一項と第二項は，それぞれ学習データに含まれる真の音声と，変換元話者の音声 $\boldsymbol{X}^{(\mathrm{s})}$ から $c^{(\mathrm{t})}$ に属する話者に変換された偽の音声を正しく識別するための損失である．ドメイン分類器の損失関数は，次式で定義される．

$$\mathcal{L}_{\mathrm{cls}} = \mathrm{CE}\left(C\left(\hat{\boldsymbol{X}}^{(\mathrm{s}\to\mathrm{t})}\right), c^{(\mathrm{s})}\right) \tag{2}$$

ここで，CE$(\cdot)$ はドメイン分類の Cross-Entropy 損失である．式 (2) に示す通り，ドメイン分類器は生成器 $G(\cdot)$ により変換された音声から，変換元の話者ドメイン $c^{(\mathrm{s})}$ を正しく識別するように学習される．このドメイン分類器に対し，生成器は次式の敵対分類損失の最小化により学習する．

$$\mathcal{L}_{\mathrm{advcls}} = \mathrm{CE}\left(C\left(\hat{\boldsymbol{X}}^{(\mathrm{s}\to\mathrm{t})}\right), c^{(\mathrm{t})}\right) \tag{3}$$

即ち，任意話者の音声を話者ドメイン $c^{(\mathrm{t})}$ に属する音声に変換した結果が，ドメイン識別器を詐称するようにモデルパラメータを更新する．加えて，生成器は以降に示す複数損失の重み付き和を最小化するマルチタスク学習により最適化される．

- **一巡一貫性損失:** 生成器により変換した音声が元のドメインに復元可能であることを保証する損失であり，$\mathcal{L}_{\mathrm{cyc}} = ||\boldsymbol{X}^{(\mathrm{s})} - \hat{\boldsymbol{X}}^{(\mathrm{s}\to\mathrm{t}\to\mathrm{s})}||_1$ として定義される．ここで，$\hat{\boldsymbol{X}}^{(\mathrm{s}\to\mathrm{t}\to\mathrm{s})} = G(\hat{\boldsymbol{X}}^{(\mathrm{s}\to\mathrm{t})}, \boldsymbol{s}^{(\mathrm{s})})$ である．
- **真贋敵対損失:** 真贋識別器を詐称する損失であり，$-\mathcal{L}_{\mathrm{adv}}$ として定義される．
- **スタイル復元損失:** 変換音声からスタイルエンコーダにより抽出されたスタイルベクトルと，変換先話者のスタイルベクトルの一貫性を保証する損失であり，$\mathcal{L}_{\mathrm{sty}} = ||\boldsymbol{s}^{(\mathrm{t})} - S(\hat{\boldsymbol{X}}^{(\mathrm{s}\to\mathrm{t})}, c^{(\mathrm{t})})||_1$ として定義される．
- **スタイル多様性損失:** 生成器が多様な発話スタイルの音声を変換可能にするための損失であり，$\mathcal{L}_{\mathrm{ds}} = ||\hat{\boldsymbol{X}}_1^{(\mathrm{s}\to\mathrm{t})} - \hat{\boldsymbol{X}}_2^{(\mathrm{s}\to\mathrm{t})}||_1 + ||\boldsymbol{h}_{\mathrm{F0},1}^{(\mathrm{s}\to\mathrm{t})} - \boldsymbol{h}_{\mathrm{F0},2}^{(\mathrm{s}\to\mathrm{t})}||_1$ として定義される．$\hat{\boldsymbol{X}}_1^{(\mathrm{s}\to\mathrm{t})}, \hat{\boldsymbol{X}}_2^{(\mathrm{s}\to\mathrm{t})}$ はそれぞれ同一話者ドメインに属する異なる 2 つのスタイルベクトル $\boldsymbol{s}_1^{(\mathrm{t})}, \boldsymbol{s}_1^{(\mathrm{t})}$ を用いて $\boldsymbol{X}^{(\mathrm{s})}$ から変換された音声である．また，$\boldsymbol{h}_{\mathrm{F0},1}^{(\mathrm{s}\to\mathrm{t})}, \boldsymbol{h}_{\mathrm{F0},2}^{(\mathrm{s}\to\mathrm{t})}$ はそれぞれ $\hat{\boldsymbol{X}}_1^{(\mathrm{s}\to\mathrm{t})}, \hat{\boldsymbol{X}}_2^{(\mathrm{s}\to\mathrm{t})}$ から F0 ネットワークにより抽出された韻律特徴量である．
- F_0 **一貫性損失:** 声質変換過程で韻律軌跡パターンが不変であることを保証する損失であり，$\mathcal{L}_{\mathrm{f0}} = ||\bar{F}(\boldsymbol{X}^{(\mathrm{s})}) - \bar{F}(\hat{\boldsymbol{X}}^{(\mathrm{s}\to\mathrm{t})})||_1$ として定義される．ここで，$\bar{F}(\boldsymbol{X})$ は音声から正規化された韻律軌跡パターンを抽出する処理であり，生成器に内包されている F0 ネットワーク $F(\cdot)$ の予測結果を用いて $\bar{F}(\boldsymbol{X}) = \frac{F(\boldsymbol{X})}{\|F(\boldsymbol{X})\|_1}$ として計算される．
- **発話内容一貫性損失:** 声質変換過程で発話内容が不変であることを保証する損失であり，$\mathcal{L}_{\mathrm{asr}} = ||\boldsymbol{h}_{\mathrm{asr}}(\boldsymbol{X}^{(\mathrm{s})}) - \boldsymbol{h}_{\mathrm{asr}}(\hat{\boldsymbol{X}}^{(\mathrm{s}\to\mathrm{t})})||_1$ として定義される．ここで，$\boldsymbol{h}_{\mathrm{asr}}(\cdot)$ は学習済み ASR モデルから抽出される音声の発話内容に関する中間表現である．
- **ノルム一貫性損失:** 声質変換過程でノルムが不変であることを保証する損失であり，$\mathcal{L}_{\mathrm{norm}} = \frac{1}{T}\sum_{t=1}^{T} \big|\, ||\boldsymbol{X}_t^{(\mathrm{s})}|| - ||\hat{\boldsymbol{X}}_t^{(\mathrm{s}\to\mathrm{t})}|| \,\big|$ として定義される．ここで，t, T はそれぞれメルスペクトログラムのフレームインデックス，総フレーム数を表す．

2.2 連合学習

連合学習 [18] 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．

2.2.1 FedAvg

McMahan らにより提案された FedAvg [19] は，連合学習における代表的なモデル統合法の一つである．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．

FedAvg を利用した学習のサーバ側，クライアント側のアルゴリズムをそれぞれ Algorithm 1 と Algorithm 2 に示す． FedAvg を用いて連合学習を行ったモデルは，一定の条件下でデータセットを集約させて学習したモデルと同等

Algorithm 1 FedAvg のサーバ側アルゴリズム

$\mathcal{C} = \{1, 2, \ldots, K\}$:= クライアントの添字集合
M := 各ラウンドでランダムに選択するクライアント数
$\boldsymbol{w}^k$:= k 番目のクライアントが所有するモデルのパラメータ
n^k := k 番目のクライアントが所有するデータ数
R := 連合学習の Round 数
for $r = 1, 2, \ldots, R$ **do**
  $\mathcal{C}_r \in \mathcal{C}$:= 乱択クライアントの添字集合 ($|\mathcal{C}_r| = M$)
  $N_r = \sum_{k \in \mathcal{C}_r} n^k$:= Round r で用いられる全データ数
  for all $k \in \mathcal{C}_r$ **do**
    モデルパラメータの配布: $\boldsymbol{w}_r^k \leftarrow \boldsymbol{w}_r$
    クライアントの更新: $\boldsymbol{w}_r^k \leftarrow \mathrm{Update\_Client}(k, \boldsymbol{w}_r^k)$
  end for
  モデルの統合: $\boldsymbol{w}_r \leftarrow \sum_{k \in \mathcal{C}_r} \frac{n^k}{N_r} \boldsymbol{w}_r^k$
end for

Algorithm 2 FedAvg のクライアント側アルゴリズム $\mathrm{Update\_Client}(k, \boldsymbol{w})$

$\mathcal{D}^k$:= k 番目のクライアントが所有するデータ
$\mathcal{B}^k$:= $\mathcal{D}^k$ から構成されるバッチの集合
E := Local epoch 数
η := 学習率
$h(b; \boldsymbol{w})$:= モデルパラメータ $\boldsymbol{w}$，バッチ b から計算される損失関数
for $i = 1, 2, \cdots, E$ **do**
  for $b \in \mathcal{B}^k$ **do**
    $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \boldsymbol{\nabla}_{\boldsymbol{w}} h(b; \boldsymbol{w})$
  end for
end for
モデルパラメータ $\boldsymbol{w}$ をサーバに送信

の性能を達成することが解析的に示されている [19].

2.2.2 FedProx

連合学習における既知の課題の一つが，各クライアントのデータ分布の非独立同一分布 (non-iid) 性が学習に与える悪影響である．一般に，各クライアントが所有するデータの分布はクライアント毎に異なると想定されるため，それぞれのデータセットに過適合したモデルの統合により，学習収束速度の低下やモデルの性能低下が生じる．Li らにより提案された FedProx [20] では，クライアントのモデルパラメータ $\boldsymbol{w}$ がサーバのモデルパラメータ $\boldsymbol{w}_0$ から極端に乖離することを防ぐために，$\mu||\boldsymbol{w} - \boldsymbol{w}_0||^2$ で表される正則化項を導入する．ここで，$\mu \in [0, 1]$ は正則化項の影響を調節するハイパーパラメータである．

3. 提案手法: Fed-StarGANv2-VC

本研究では，ユーザが所有するデータのプライバシーを保護しながらユーザ参加型多対多声質変換モデル学習の実現を目的とし，StarGANv2-VC に連合学習の枠組みを適用した Fed-StarGANv2-VC を提案する．

3.1 問題設定

多対多声質変換技術のアプリケーションでは，多様な声質のユーザからの入力音声を想定する必要がある．故に，2.2.2 節で述べたように，連合学習で多対多声質変換モデルを学習する場合においても，参加するクライアントが所有するデータセットに iid 性を仮定することは非現実的である．一方で，近年では研究開発者により多数の話者を含む大規模な音声データセット（例えば，日本語の JVS コーパス [21] や英語の LibriTTS コーパス [22]）の構築・整備が進められており，これらを活用することで non-iid 性への対処が容易になると考えられる．そこで，本研究では連合学習に用いるデータセットを以下の 2 つに分類する．

- **Anchor データセット** $\mathcal{D}_{\mathrm{Anc}}$**:** すべてのクライアントが共通でアクセス可能なデータセット
- **Client データセット** $\mathcal{D}_{\mathrm{Cli}}$**:** 各クライアントのみアクセス可能なデータセット

即ち，Algorithm 2 に示すクライアント側のモデルパラメータ更新では，k 番目のクライアントは $\mathcal{D} = \mathcal{D}_{\mathrm{Anc}} \cup \mathcal{D}_{\mathrm{Cli}}^k$ を用いて損失関数と勾配を計算できると仮定する．

3.2 Fed-StarGANv2-VC のアルゴリズム

提案手法では，StarGANv2-VC を構成する各サブモジュールを連合学習により最適化する．提案手法における FedAvg のサーバ側・クライアント側のアルゴリズムはそれぞれ Algorithm 1 と Algorithm 2 と同じであるが，クライアント側の更新において学習時のバッチを構成する処理（Algorithm 2 2 行目）が $\mathcal{B}^k := \mathcal{D} = \mathcal{D}_{\mathrm{Anc}} \cup \mathcal{D}_{\mathrm{Cli}}^k$ となるという点で異なる．

4. 実験的評価

4.1 実験条件

本稿では，StarGANv2-VC モデル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手法と，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手法を比較した．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，本稿では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 [8] の著者により公開されているオープンソース実装[*1] 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，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 [23] を利用した．

データセットとして，JVS コーパス [21] の parallel100 サブセット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

[*1] https://github.com/yl4579/StarGANv2-VC

表 1　評価する変換対

ラベル名	入力話者	出力話者
Anc → Anc	Anchor 話者	Anchor 話者
Anc → Cli	Anchor 話者	Client 話者
Cli → Anc	Client 話者	Anchor 話者
Cli → Cli	Client 話者	Client 話者

タの数はそれぞれ 3284, 411, 411 であった．本研究では，クライアントごとに異なる話者の音声データを持っている状況を再現するために，40 名の話者を 10 名の Anchor 話者 ($\mathcal{D}_{\mathrm{Anc}}$) と 30 名の Client 話者 ($\mathcal{D}^k_{\mathrm{Cli}}$ $(k = 1, \ldots, 30)$) に分類した．Anchor 話者は全てのクライアントが学習に利用できる話者，Client 話者はそれぞれ一つのクライアントのみにデータが存在する話者とした．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 [24] 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．

以降の各評価は，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．評価における変換対のラベルを**表 1** に示す．

4.2　客観評価

本稿では，変換音声の話者類似性の客観評価指標として x-vector [25] のコサイン類似度 (x-vector cossim) を，自然性の客観評価指標として UTMOS [26] による自然性 Mean Opinion Score (MOS) 評価の予測値 (pMOS) を用いた．x-vector の抽出には JTubeSpeech コーパス [27] を用いて学習された事前学習済みモデル[*2]を利用した．

4.2.1　連合学習の実験条件に関する調査

はじめに，提案手法である Fed-StarGANv2-VC 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．提案手法における x-vector cossim，pMOS の評価結果をそれぞれ**表 2**，**表 3** に示す．

モデル統合時のクライアント数の影響: Round 数を 400 で固定した場合，x-vector cossim の評価結果（**表 2**）については FedProx 正則化項の有無・変換話者対のパターンの違いに関わらず，概ね僅かに改善した．一方で，pMOS の評価結果（**表 3**）については，FedProx 正則化項の有無に関わらず，クライアント数の増加により，Anchor 話者を変換先に指定した場合では pMOS 値が劣化し，その逆の場合では pMOS 値が改善した．本稿の実験設定では，Anchor 話者の数 (10 名) と比較して Client 話者の数 (30 名) が多いため，複数クライアントからの学習結果のモデルを統合する処理により，全体の割合としては少数派にあたる Anchor 話者への変換性能が自然性の観点で劣化したと考えられる．

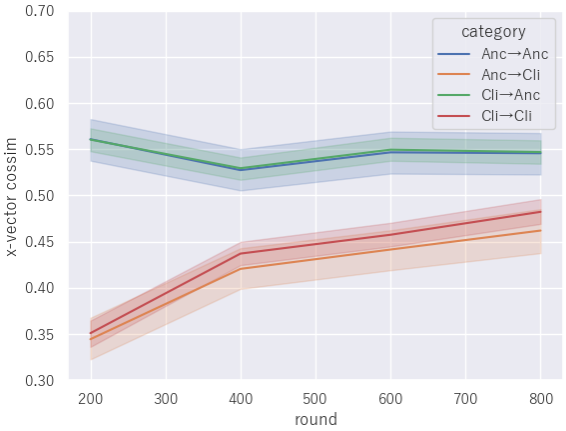

図 3　連合学習の Round 数に対する x-vector cossim 評価値の遷移

連合学習の Round 数: 表 2 及び**表 3** より，どちらの評価指標においても，Round 数を増加させることで概ね改善する傾向にあった．注目すべき点として，Anchor 話者を変換先に指定した場合ではその改善傾向が弱いのに対し，その逆の場合では顕著な改善傾向が示された．この傾向は，統合クライアント数 3，FedProx 正則化項ありの場合について横軸を Round 数，縦軸を x-vector cossim として作成した**図 3** に顕著に示されている．これらの結果は，連合学習によって多対多声質変換モデルが表現可能な話者の多様性を向上させるためには，十分な Round 数が要求されることを示唆している．

FedProx 正則化項の有無: 表 3 より，達成可能な変換音声の自然性に関しては，FedProx 正則化項の有無は大きく寄与せず，学習の Round 数による影響が支配的であった．一方で，**表 2** より，Anchor 話者を変換先に指定した場合では正則化の影響は小さく，その逆の場合では顕著な改善が見られた．これらの結果は，多対多声質変換モデルの連合学習における学習データの non-iid 性への対処は，モデルが変換先として表現可能な話者の多様性を増加させるために重要である可能性を示唆している．

以降の評価では，Client 話者を対象とした変換時の話者類似性が最も高い {# Cli = 3, # Round = 800, FedProx 正則化あり } で学習した提案手法を比較対象として採用した．

[*2] https://github.com/sarulab-speech/xvector_jtubespeech

表 2 提案手法の x-vector cossim 評価結果（平均 ± 標準偏差）

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**

表 3 提案手法の pMOS 評価結果（平均 ± 標準偏差）

Model			Anc→Anc	Anc→Cli	Cli→Anc	Cli→Cli
# Cli	# Round	FedProx?				
1	200	No	2.03 ± 0.34	1.93 ± 0.30	2.10 ± 0.32	2.05 ± 0.28
1	400	No	**2.25 ± 0.3**3	2.28 ± 0.30	2.33 ± 0.32	2.38 ± 0.28
3	200	No	2.13 ± 0.35	2.30 ± 0.31	2.19 ± 0.34	2.43 ± 0.25
3	400	No	2.16 ± 0.32	2.44 ± 0.29	2.28 ± 0.34	2.59 ± 0.27
1	200	Yes	2.17 ± 0.27	2.23 ± 0.27	2.21 ± 0.29	2.30 ± 0.28
1	400	Yes	**2.25 ± 0.33**	2.25 ± 0.31	**2.35 ± 0.32**	2.39 ± 0.30
3	200	Yes	2.19 ± 0.39	2.32 ± 0.29	2.23 ± 0.37	2.42 ± 0.26
3	400	Yes	2.18 ± 0.34	**2.45 ± 0.29**	2.27 ± 0.34	2.56 ± 0.25
3	600	Yes	2.13 ± 0.34	2.41 ± 0.25	2.22 ± 0.34	**2.65 ± 0.26**
3	800	Yes	2.09 ± 0.33	2.37 ± 0.29	2.20 ± 0.34	2.63 ± 0.28

4.2.2 従来の非分散型学習法に関する調査

次に，本実験のベースラインである，連合学習を用いずに学習を行う StarGANv2-VC のモデルを比較するために，異なる Epoch 数での客観評価指標を計算した．**表 4** と**表 5** にそれぞれ従来手法で学習された多対多声質変換モデルの x-vector cossim と pMOS 値の評価結果を示す．なお，変換話者対のパターンは提案手法と同様の 4 種を示しているが，従来手法の学習では {Anc, Cli} を区別していない．これらの表から，Epoch の増加によって x-vector cossim（**表 4**）は改善傾向にあるが，pMOS 値（**表 5**）の改善傾向は弱いことが確認できる．また，600 Epoch と 700 Epoch 終了時点での客観指標に大きな差は見られないが，800 Epoch 終了時点にて x-vector cossim と pMOS の評価結果に僅かな劣化が見られたため，以降の評価では，700 Epoch 学習後の従来手法を比較対象として採用した．

4.3 主観評価

節での客観評価で最良な結果を示した従来手法 (Conventional) と提案手法 (Proposed) により変換された音声の品質に関する主観評価を実施した．評価手法は，変換音声の話者類似性に関するプリファレンス XAB テストと，自然性に関するプリファレンス AB テストとした．評価ではテストデータから全 40 話者のそれぞれの音声データ 3 つをランダムに選択し，従来手法・提案手法で学習した多対多声質変換モデルを用いて自身以外の話者に変換した音声を用いた．また，本評価では学習後のモデルがユーザにより提供される音声を高精度に変換可能かどうかに着目するために，変換元は Client 話者のデータで固定した．主観評価の被験者は Lancers[*3]でのクラウドソーシングにより，変換話者対の種別ごとに独立して各 50 名ずつ，計 800 名を募集した．被験者は従来手法と提案手法で変換された同一発話内容の変換音声対を聴取し，AB テストの場合はどちらの音声がより自然かを回答した．XAB テストの場合は，同時に提示された変換先話者の参照音声データと比較し，どちらの音声が当該話者に類似しているかを回答させた．

表 6 と**表 7** にそれぞれ話者類似性と自然性に関する主観評価結果を示す．なお，入出力話者の性別は男性を (M)，女性を (F) で表現した．ここで，太字は対照手法のスコアを $p < 0.05$ の有意差で上回っていることを示す．**表 6** より，提案手法はすべての声質変換設定において従来手法と同程度・もしくは有意に上回る話者類似性を達成した．故に，提案手法はユーザ参加型の多対多声質変換モデル学習により，変換可能な話者の多様性を改善できる可能性が示唆された．この結果に対する考察として，従来手法では，40 話者間の変換を単一のモデルで一度に学習する一方で，

[*3] https://www.lancers.jp/

表 4 従来手法の x-vector コサイン類似度評価結果（平均 ± 標準偏差）

# Epoch	Anc→Anc	Anc→Cli	Cli→Anc	Cli→Cli
200	0.49 ± 0.21	0.50 ± 0.34	0.49 ± 0.22	0.51 ± 0.34
300	0.54 ± 0.17	0.52 ± 0.33	0.55 ± 0.17	0.53 ± 0.32
400	0.51 ± 0.20	0.52 ± 0.35	0.53 ± 0.20	0.53 ± 0.34
500	0.52 ± 0.18	0.50 ± 0.35	0.52 ± 0.17	0.50 ± 0.34
600	**0.55 ± 0.19**	0.55 ± 0.32	0.55 ± 0.18	**0.56 ± 0.32**
700	0.54 ± 0.18	**0.55 ± 0.33**	**0.56 ± 0.18**	0.55 ± 0.33
800	0.53 ± 0.19	0.53 ± 0.33	0.54 ± 0.18	0.53 ± 0.33

表 5 従来手法の pMOS 評価結果（平均 ± 標準偏差）

# Epoch	Anc→Anc	Anc→Cli	Cli→Anc	Cli→Cli
200	2.13 ± 0.31	2.21 ± 0.29	2.24 ± 0.30	2.28 ± 0.27
300	2.18 ± 0.31	**2.27 ± 0.28**	2.27 ± 0.31	**2.37 ± 0.28**
400	2.22 ± 0.34	2.26 ± 0.28	2.28 ± 0.32	2.33 ± 0.29
500	**2.27 ± 0.32**	2.29 ± 0.26	2.32 ± 0.33	2.35 ± 0.28
600	2.17 ± 0.31	2.21 ± 0.28	2.26 ± 0.32	2.28 ± 0.28
700	2.23 ± 0.38	2.25 ± 0.29	**2.28 ± 0.35**	2.32 ± 0.29
800	2.19 ± 0.31	2.23 ± 0.26	2.25 ± 0.34	2.31 ± 0.28

表 6 話者類似性に関するプリファレンス XAB スコア

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

表 7 自然性に関するプリファレンス AB スコア

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**

提案手法では，各クライアントのデータセットに存在する話者数は Anchor 話者 10 名 + Client 話者 1 人であるため，より少ない話者間の変換を学習することになる．連合学習では先述の少ない話者間の変換を学習したものを統合するため，結果として，従来手法よりも話者性の再現精度が悪い局所解に陥ることなく，より良いモデルパラメータに収束したと考えられる．

一方で，**表 7** より，変換音声の自然性に関しては，変換話者対の設定で大きく異なる傾向を示した．つまり，提案手法により変換された音声の自然性は，従来手法と比較して，Anchor 話者を変換先とした場合は有意に低く，Client 話者を変換先とした場合は有意に高い結果となった．この傾向は客観評価における pMOS 指標の大小関係と概ね等しい．即ち，4.2.1 節で述べたように，変換音声の自然性については，Anchor/Client 話者のデータ量の不均衡さが連合学習における統合結果に悪影響を及ぼす可能性が客観評価結果と主観評価結果の両方において示唆された．

5. おわりに

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声質変換モデルを学習する手法として Fed-StarGANv2-VC 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 同士の話者変換についても，データを集約させて学習を行う従来手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，仮想的にではなく，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．

謝辞 本研究は，公益財団法人 立石科学技術振興財団 2022 年度研究助成 (A)，JSPS 科研費 21K21305 による支援を受けたものです．

参考文献

[1] Toda, T., Black, A. W. and Tokuda, K.: Voice conversion based on maximum likelihood estimation of spectral parameter trajectory, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2222–2235 (2007).
[2] Fang, F., Wang, X., Yamagishi, J., Echizen, I., Todisco, M., Evans, N. and Bonastre, J.-F.: Speaker anonymization using x-vector and neural waveform models, *Proc. SSW*, Vienna, Austria, pp. 155–160 (2019).
[3] Doi, H., Toda, T., Nakano, T., Goto, M. and Nakamura,

S.: Singing voice conversion method based on many-to-many Eigenvoice conversion and training data generation using a singing-to-singing synthesis system, *Proc. APSIPA ASC*, Hollywood, U.S.A., pp. 1–6 (2012).

[4] Kaneko, T. and Kameoka, H.: CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks, *Proc. EUSIPCO*, Rome, Italy, pp. 2114–2118 (2018).

[5] Kameoka, H., Kaneko, T., Tanaka, K. and Hojo, N.: StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks, *Proc. SLT*, Greece, Athens, pp. 266–273 (2018).

[6] Qian, K., Zhang, Y., Chang, S., Yang, X. and Hasegawa-Johnson, M.: AutoVC: Zero-shot voice style transfer with only autoencoder loss, *Proc. ICML*, Long Beach, U.S.A., pp. 5210–5219 (2019).

[7] Xie, C., Wu, Y.-C., Tobing, P. L., Huang, W.-C. and Toda, T.: Direct noisy speech modeling for noisy-to-noisy voice conversion, *Proc. ICASSP*, Singapore, pp. 6787–6791 (2022).

[8] Li, Y. A., Zare, A. and Mesgarani, N.: StarGANv2-VC: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion, *Proc. INTERSPEECH*, Brno, Czech Republic, pp. 266–273 (2021).

[9] Desai, S., Raghavendra, E. V., Yegnanarayana, B., Black, A. W. and Prahallad, K.: Voice conversion using Artificial Neural Networks, *Proc. ICASSP*, Taipei, Taiwan, pp. 3893–3896 (2009).

[10] Tanaka, K., Kameoka, H., Kaneko, T. and Hojo, N.: AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms, *Proc. ICASSP*, Brighton, U.K., pp. 6805–6809 (2019).

[11] Hayashi, T., Huang, W.-C., Kobayashi, K. and Toda, T.: Non-autoregressive sequence-to-sequence voice conversion, *Proc. ICASSP*, Montreal, Canada, pp. 7068–7072 (2021).

[12] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M. and Lempitsky, V.: Domain-adversarial training of neural networks, *Journal of Machine Learning Research*, Vol. 17, No. 59, pp. 1–35 (2016).

[13] Sun, L., Li, K., Wang, H., Kang, S. and Meng, H.: Phonetic posteriorgrams for many-to-one voice conversion without parallel data training, *Proc. ICME*, Seattle, U.S.A. (2016).

[14] Zhang, J.-X., Ling, Z.-H. and Dai, L.-R.: Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, No. 1, pp. 540–552 (2020).

[15] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: generative adversarial nets, *Proc. NIPS*, Montreal, Canada, pp. 2672–2680 (2014).

[16] Zhu, J.-Y., Park, T., Isola, P. and Efros, A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks, *Proc. ICCV*, Venice, Italy, pp. 2223–2232 (2017).

[17] Kum, S. and Nam, J.: Joint detection and classification of singing voice melody using convolutional recurrent neural networks, *Applied Sciences*, Vol. 9, No. 7 (2019).

[18] Konečný, J., McMahan, H. B., Yu, F. X., Richtarik, P., Suresh, A. T. and Bacon, D.: Federated learning: strategies for improving communication efficiency, *NIPS Workshop on Private Multi-Party Machine Learning* (2016).

[19] McMahan, H. B., Moore, E., Ramage, D., Hampson, S. and y Arcas, B. A.: Communication-efficient learning of deep networks from decentralized data, *Proc. AISTATS*, Fort Lauderdale, U.S.A.

[20] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A. and Smith, V.: Federated optimization in heterogeneous networks, *Proc. AMTL*, Long Beach, U.S.A.

[21] Takamichi, S., Sonobe, R., Mitsui, K., Saito, Y., Koriyama, T., Tanji, N. and Saruwatari, H.: JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research, *Acoustical Science and Technology*, Vol. 41, No. 5, pp. 761–768 (2020).

[22] Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z. and Wu, Y.: LibriTTS: A corpus derived from LibriSpeech for text-to-speech, *Proc. INTERSPEECH*, Graz, Austria, pp. 1526–1530 (2019).

[23] Yamamoto, R., Song, E. and Kim, J.: Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram, *Proc. ICASSP*, Barcelona, Spain, pp. 6199–6203 (2020).

[24] Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization, *Proc. ICLR*, New Orleans, U.S.A. (2019).

[25] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S.: X-Vectors: Robust DNN embeddings for speaker recognition, *Proc. ICASSP*, Alberta, Canada, pp. 5329–5333 (2018).

[26] Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S. and Saruwatari, H.: UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022, *Proc. INTERSPEECH*, Incheon, South Korea, pp. 4521–4525 (2022).

[27] Takamichi, S., Kürzinger, L., Saeki, T., Shiota, S. and Watanabe, S.: JTubeSpeech: Corpus of Japanese speech collected from YouTube, *arXiv*, Vol. abs/2112.09323 (2021).