音素事後確率を用いた多対一音声変換のための音声認識・生成モデルの同時敵対学習 *

◎齋藤 佑樹，△阿久澤 圭（ディー・エヌ・エー/東大），橘 健太郎（ディー・エヌ・エー）

1 はじめに

多対一音声変換とは，任意の話者の声質を特定の目的話者のものに変換する技術であり，音声バーチャルリアリティへの応用が期待されている．本稿で対象とする，Deep Neural Network (DNN) に基づく多対一音声変換において，目的話者の音声生成モデルを学習するための中間表現として，音声認識モデルにより予測される音素事後確率を用いる手法が提案されている [1]．この手法では，まず，多数話者を含む音声コーパスを用いて，入力された音声特徴量から発話内容を予測するように音声認識モデルを学習する．次に，目的話者のみを含む音声コーパスを用いて，音素事後確率から音声特徴量を予測するように音声生成モデルを学習する．最後に，音声認識・生成モデルを結合し，任意話者の音声特徴量から，目的話者の音声特徴量を予測する多対一音声変換モデルを構築する．これにより，任意話者の学習データを必要せず，目的話者の音声変換が可能となる．この手法では，パラレル音声コーパスを用いずに学習できるが，音声認識・生成モデルを個別に学習させているため，発話スタイルや録音環境といった話者間の違いに対処できない．Miyoshi ら [2] により，これらの話者間の違いが音素事後確率に影響を与え，変換音声の品質が劣化することが報告されている．また，音声生成モデルにより予測される音声特徴量の過剰な平滑化 [3, 4] により，変換音声の品質はさらに劣化する．

本稿では，音素事後確率を用いた多対一音声変換の高品質化を目的として，音声認識・生成モデルの同時敵対学習を提案する．提案法では，2 つの DNN 識別モデルを新たに導入する．一方は domain-adversarial training (DAT) [5] のドメイン識別モデルとして動作し，音声認識モデルから抽出される潜在変数を用いて，目的話者とそれ以外の話者を識別する．音声認識モデルは，音素認識誤差と，ドメイン識別器を誤識別させる損失の重み付き和を最小化するように学習されるため，音素事後確率が入力話者不変になることが期待できる．他方は話者認証モデルとして動作し，目的話者の音声特徴量と，音声生成モデルから予測される音声特徴量を識別する．音声生成モデルは，音声特徴量生成誤差と，話者認証モデルを騙す損失の重み付き和を最小化するように学習されるため，generative adversarial network (GAN) [6] に基づく音声合成 [7] と同様に，過剰な平滑化による品質劣化の緩和が期待できる．従来法と異なり，提案法では，音声認識・生成モデルの両方の損失関数を用いた同時学習により，入力音声特徴量から目的話者の音声特徴量を予測するように 2 つのモデルを最適化する．実験的評価により，従来法と比較して，提案法が変換音声の品質を有意に改善させることを示す．

2 従来の多対一音声変換

従来法 [1] では，個別に学習させた音声認識・生成モデルを結合し，任意話者の音声を目的話者の音声に変換する多対一音声変換モデルを構築する．

2.1 音声認識モデルの学習

音声認識モデル $R(\cdot)$ は，MFCC などの入力音声特徴量系列 $\boldsymbol{x}$ から，音素ラベル系列 $\boldsymbol{l}$ を予測するように学習される．入力音声特徴量と音素ラベルのペアは，多数話者を含む音声コーパス $\mathcal{D}^{(\mathrm{M})} = \{(\boldsymbol{x}_n^{(\mathrm{M})}, \boldsymbol{l}_n^{(\mathrm{M})})\}_{n=1}^{N^{(\mathrm{M})}}$ から抽出される．ここで，$N^{(\mathrm{M})}$ は音声認識モデルの学習データ数である．$R(\cdot)$ により予測される音素事後確率系列 $\hat{\boldsymbol{p}}^{(\mathrm{M})} = R(\boldsymbol{x}^{(\mathrm{M})})$ は，入力音声特徴量 $\boldsymbol{x}^{(\mathrm{M})}$ が与えられたもとでの音素ラベル $\boldsymbol{l}^{(\mathrm{M})}$ の事後確率を表す．$R(\cdot)$ は，音素ラベルと音素事後確率の softmax cross-entropy $L_{\mathrm{SCE}}(\boldsymbol{l}^{(\mathrm{M})}, \hat{\boldsymbol{p}}^{(\mathrm{M})})$ として定義される音素認識誤差を最小化するように学習される．

2.2 音声生成モデルの学習

音声認識モデル $R(\cdot)$ の話者非依存性を仮定し，目的話者の音声生成モデル $G(\cdot)$ は，音素事後確率 $\hat{\boldsymbol{p}} = R(\boldsymbol{x})$ からメルケプストラム係数などの目的音声特徴量系列 $\boldsymbol{y}$ を予測するように学習される．入力・目的音声特徴量のペアは，目的話者のみを含む音声コーパス $\mathcal{D}^{(\mathrm{O})} = \{(\boldsymbol{x}_n^{(\mathrm{O})}, \boldsymbol{y}_n^{(\mathrm{O})})\}_{n=1}^{N^{(\mathrm{O})}}$ から抽出される．ここで，$N^{(\mathrm{O})}$ は音声生成モデルの学習データ数である．音声特徴量の予測結果は音声認識・生成モデルを通じて $\hat{\boldsymbol{y}}^{(\mathrm{O})} = G(R(\boldsymbol{x}^{(\mathrm{O})}))$ として得られる．$G(\cdot)$ は，予測結果と目的音声特徴量の mean squared error $L_{\mathrm{MSE}}(\boldsymbol{y}^{(\mathrm{O})}, \hat{\boldsymbol{y}}^{(\mathrm{O})})$ として定義される音声特徴量生成誤差を最小化するように学習される．$R(\cdot)$ のモデルパラメータ（DNN の重みとバイアス）は，ここでは更新されず，事前学習後の状態で固定される．

2.3 従来法の問題点

従来法では，パラレル音声コーパスを用いずに多対一音声変換モデルを学習できる．しかし，音素認識誤差のみを最小化する学習は，音素事後確率の話者不変性までは保証しない．故に，Fig. 1(a) に示すように，音声生成モデルに入力される音素事後確率は，同一の発話内容でも話者ごとに大きく異なる．音声生成モデルの学習では，目的話者の音声コーパスだけを用いるため，変換元話者の音素事後確率の違いは，変換音声の品質劣化の一因となりうる．また，Fig. 2(b) に示すように，音声生成モデルから予測された音声特徴量の分布は過剰に平滑化され，変換音声の品質は著しく劣化する．

---

* Joint adversarial training algorithm of speech recognition and synthesis models for many-to-one voice conversion using phonetic posteriorgrams by SAITO, Yuki, AKUZAWA, Kei (DeNA Co., Ltd./The University of Tokyo), and TACHIBANA, Kentaro (DeNA Co., Ltd.).

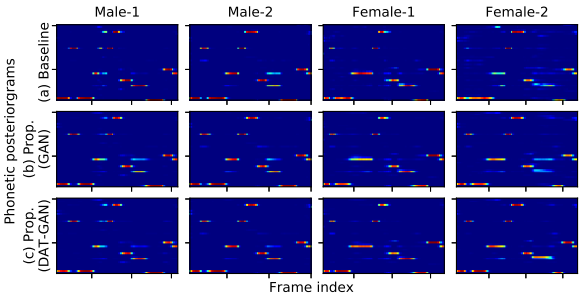

Fig. 1　異なる 4 話者の音声から予測された音素事後確率の例．この図は，4.2 節の主観評価で用いたものと同じ発話で作成した．図の色が赤に近いほど，事後確率がより高いことを意味する．音素継続長は話者ごとに異なるため，横軸を調整して表示している．

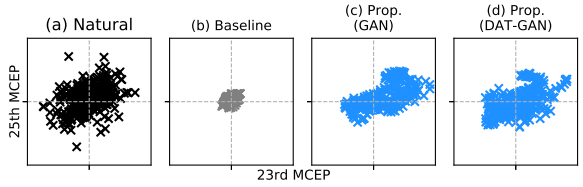

Fig. 2　多対一音声変換の目的話者のメルケプストラム係数 (MCEP) の散布図．この図は，音声生成モデルの学習に用いられていない 1 発話から生成した．

3　提案する多対一音声変換

3.1　音声認識・生成モデルの同時敵対学習

3.1.1　多対一音声変換のための音声認識モデルの domain-adversarial training

Domain-adversarial training (DAT) [5] は，ドメイン不変な潜在変数の学習により，DNN に基づく認識モデルの入力特徴量の変動に対する頑健性を向上させるための一般的な枠組みであり，話者認識 [8] などにも応用されている．DAT は認識モデルの精度向上を目的とした技術だが，autoencoder ベースの音声変換における有効性も報告されている [9]．この手法では，音声認識モデルは不要だが，autoencoder の潜在変数が入力音声の発話内容を表現する保証がない．本稿では，DAT に基づき，音声生成モデルに入力される音素事後確率の話者不変性を向上させる学習法を提案する．ここでは，音声認識・生成モデルの学習に用いた音声コーパスをそれぞれ (1) 多数話者ドメイン $\mathcal{D}^{(\mathrm{M})}$ と (2) 目的話者ドメイン $\mathcal{D}^{(\mathrm{O})}$ として定義し，DAT により，この 2 つのドメインの違いを緩和する．

以降，定式化のため，音声認識モデルを $R(\cdot) = R_{\mathrm{p}}(R_{\mathrm{f}}(\cdot))$ として 2 つの部分モデルに分解して表記する．$R_{\mathrm{f}}(\cdot)$ は特徴抽出モデルであり，$\hat{\boldsymbol{f}} = R_{\mathrm{f}}(\boldsymbol{x})$ として，入力音声特徴量 $\boldsymbol{x}$ から発話内容に関する潜在変数 $\hat{\boldsymbol{f}}$ を抽出する．$R_{\mathrm{p}}(\cdot)$ は音素予測モデルであり，$\hat{\boldsymbol{p}} = R_{\mathrm{p}}(\hat{\boldsymbol{f}}) = R_{\mathrm{p}}(R_{\mathrm{f}}(\boldsymbol{x}))$ として，潜在変数 $\hat{\boldsymbol{f}}$ から音素事後確率 $\hat{\boldsymbol{p}}$ を出力する．提案法では，ドメインの違いを捉えるために，ドメイン識別モデル $D_{\mathrm{dc}}(\cdot)$ を導入して音声認識モデルを学習する．$D_{\mathrm{dc}}(\cdot)$ の学習で最小化される損失関数は，次式で与えられる．

$$L_{\mathrm{dc}}\left(\hat{\boldsymbol{f}}^{(\mathrm{M})}, \hat{\boldsymbol{f}}^{(\mathrm{O})}\right) = -\log D_{\mathrm{dc}}\left(\hat{\boldsymbol{f}}^{(\mathrm{O})}\right) - \log\left(1 - D_{\mathrm{dc}}\left(\hat{\boldsymbol{f}}^{(\mathrm{M})}\right)\right), \quad (1)$$

ここで，$\hat{\boldsymbol{f}}^{(\mathrm{M})}$ と $\hat{\boldsymbol{f}}^{(\mathrm{O})}$ はそれぞれ $\boldsymbol{x}^{(\mathrm{M})}$ と $\boldsymbol{x}^{(\mathrm{O})}$ から抽出された潜在変数である．一方で，音声認識モデル学習で最小化される損失関数は，次式で与えられる．

$$L_{\mathrm{R}}\left(\boldsymbol{l}^{(\mathrm{M})}, \hat{\boldsymbol{p}}^{(\mathrm{M})}, \hat{\boldsymbol{f}}^{(\mathrm{M})}, \hat{\boldsymbol{f}}^{(\mathrm{O})}\right) = L_{\mathrm{SCE}}\left(\boldsymbol{l}^{(\mathrm{M})}, \hat{\boldsymbol{p}}^{(\mathrm{M})}\right) - \omega_{\mathrm{R}} L_{\mathrm{dc}}\left(\hat{\boldsymbol{f}}^{(\mathrm{M})}, \hat{\boldsymbol{f}}^{(\mathrm{O})}\right), \quad (2)$$

ここで，ω_{R} は式 (2) の第二項の影響を調整するハイパーパラメータである．この損失関数は，従来の音素認識誤差と，潜在変数の修正により $D_{\mathrm{dc}}(\cdot)$ を誤認識させる損失の重み付き和として解釈でき，音素事後確率に含まれる話者間の違いの緩和が期待できる．

3.1.2　Generative adversarial network に基づく音声生成モデルの学習

Generative adversarial network (GAN) に基づく音声合成 [7] は，自然音声特徴量 $\boldsymbol{y}$ と合成音声特徴量 $\hat{\boldsymbol{y}}$ を識別する話者認証モデル $D_{\mathrm{sv}}(\cdot)$ を導入して音声生成モデルを学習させる手法である．GAN の目的関数は，真のデータと生成データの分布間の距離規範最小化であるため，$\hat{\boldsymbol{y}}$ の分布を $\boldsymbol{y}$ のものに近づける学習により，過剰な平滑化を緩和する．本稿では，この手法を音素事後確率を用いた多対一音声変換に導入し，変換音声の品質改善を試みる．

先行研究 [7] の結果に基づき，本稿では，$D_{\mathrm{sv}}(\cdot)$ に Wasserstein GAN [10] ベースの識別器を用いる．$D_{\mathrm{sv}}(\cdot)$ の学習で最小化される損失関数は，次式で与えられる．

$$L_{\mathrm{sv}}\left(\boldsymbol{y}^{(\mathrm{O})}, \hat{\boldsymbol{y}}^{(\mathrm{O})}\right) = -D_{\mathrm{sv}}\left(\boldsymbol{y}^{(\mathrm{O})}\right) + D_{\mathrm{sv}}\left(\hat{\boldsymbol{y}}^{(\mathrm{O})}\right). \quad (3)$$

ここで，$D_{\mathrm{sv}}(\cdot)$ の K-Lipschits 性を満たすために，$D_{\mathrm{sv}}(\cdot)$ の更新後に重みパラメータの値を $[-0.01, 0.01]$ のような一定区間内に収める．式 (3) の最小化により，$\boldsymbol{y}^{(\mathrm{O})}$ と $\hat{\boldsymbol{y}}^{(\mathrm{O})}$ の分布間の Earth-Mover 距離が近似される．一方で，$G(\cdot)$ は，この近似された Earth-Mover 距離を最小化する制約を考慮して学習される．$G(\cdot)$ の学習で最小化される損失関数は，次式で与えられる．

$$L_{\mathrm{G}}\left(\boldsymbol{y}^{(\mathrm{O})}, \hat{\boldsymbol{y}}^{(\mathrm{O})}\right) = L_{\mathrm{MSE}}\left(\boldsymbol{y}^{(\mathrm{O})}, \hat{\boldsymbol{y}}^{(\mathrm{O})}\right) + \omega_{\mathrm{G}} L_{\mathrm{adv}}\left(\hat{\boldsymbol{y}}^{(\mathrm{O})}\right), \quad (4)$$

ここで，$L_{\mathrm{adv}}(\hat{\boldsymbol{y}}^{(\mathrm{O})}) = -D_{\mathrm{sv}}(\hat{\boldsymbol{y}}^{(\mathrm{O})})$ は，$D_{\mathrm{sv}}(\cdot)$ を騙すための損失であり，ω_{G} はこの損失の影響を調整するハイパーパラメータである．式 (4) を最小化する提案法の学習により，多対一音声変換における音声特徴量の過剰な平滑化の緩和が期待できる．

3.1.3　音声認識・生成モデルの同時学習

音声認識・生成モデルを多対一音声変換に最適化するために，本稿では，2 つのモデルを統一的な枠組みで同時に学習する．まず，式 (1) と式 (3) を最小化するように，$D_{\mathrm{dc}}(\cdot)$ と $D_{\mathrm{sv}}(\cdot)$ をそれぞれ学習する．その後，式 (2) と式 (4) の和を最小化するように，$R(\cdot)$ と $G(\cdot)$ を同時に学習する．最終的な多対一音声変換モデルは，このような識別モデルと音声認識・生成モデルの交互最適化により構築される．$G(\cdot)$ の損失関数が $R(\cdot)$ の学習時にも考慮されるため，学習される音素事後確率が入力話者不変で，かつ，高品質な目的音声特徴量を予測できるようになると期待できる．Figure 3 に提案法の損失関数の計算手順を示す．

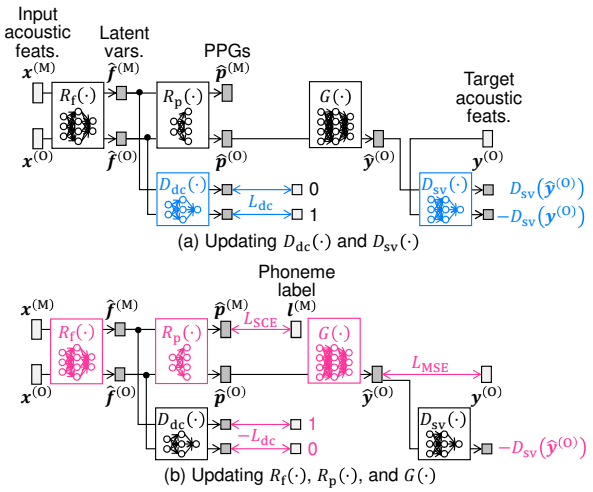

Fig. 3　提案法の損失関数の計算手順．

3.2　考察

Figure 2(c) より，GAN に基づく学習により，予測された音声特徴量の過剰な平滑化の緩和が確認できる．しかし，Fig. 1(b) より，GAN を適用するだけでは，音素事後確率に含まれる話者間の違いは緩和できない．一方で，Fig. 1(c) 及び 2(d) より，DAT と GAN の両方に基づく学習は，過剰な平滑化だけでなく，音素事後確率の違いまでも緩和しており，多対一音声変換における変換音声の品質改善が期待できる．

GAN に基づく学習では，目的話者の音声特徴量 $\boldsymbol{y}^{(\mathrm{O})}$ と，それ以外の話者から予測された音声特徴量 $\hat{\boldsymbol{y}}^{(\mathrm{M})} = G(R(\boldsymbol{x}^{(\mathrm{M})}))$ の分布間距離規範も最小化できる．しかし，予備実験により，このような学習が変換音声の品質を著しく劣化させることを確認した．この一因として，音声特徴量空間での話者間の違いは，音声認識モデルから抽出される潜在変数空間での違いよりも大きいため，前者を GAN で最小化する学習がより困難であったことが推測される．

4　実験的評価

4.1　実験条件

本稿では，3 つの音声コーパスを用いて実験的評価を行う．1 つ目は CSJ (Corpus of Spontaneous Japanese) コーパス [11] であり，男性話者 947 名・女性話者 470 名によるモノローグ，対話，朗読といった様々な発話様式の音声を含む．このうちの約 99%の発話を音声認識モデル学習用の多数話者コーパス $\mathcal{D}^{(\mathrm{M})}$ とする．2 つ目は NICT 声優対話コーパス [12] であり，1 名の女性声優による模擬対話音声発話を含む．予備検討において，一般話者と声優では，音素事後確率の分布に顕著な違いが確認されたため，本稿では，このコーパスに含まれる 5,174 発話を音声生成モデル学習用の目的話者コーパス $\mathcal{D}^{(\mathrm{O})}$ とする．3 つ目は ATR デジタル音声データベースのセット C [13] であり，男性話者 148 名・女性話者 143 名による読み上げ発話様式の音声を含む．このうちの男性 10 名・女性 10 名を多対一音声変換の入力話者とし，変換音声の品質評価を容易にするために，各話者のパラレル音声 1 発話（音素バランス文 A01）を用いる．音声データのサンプリング周波数は 16kHz である．音声特徴量は，WORLD ボコーダ [14] を用いて抽出された対数 F0，39 次のメルケプストラム係数，そして非周期性指標 [15] である．多対一音声変換では，1 次から 39 次までのメルケプストラム係数を DNN によって変換し，対数 F0 は線形変換する．非周期性指標と 0 次のメルケプストラム係数は，変換元話者のものを用いる．

Table 1　実験的評価で用いた DNN のアーキテクチャ

音声認識モデル $R(\cdot) = R_{\mathrm{p}}(R_{\mathrm{f}}(\cdot))$	音声生成モデル $G(\cdot)$
特徴抽出モデル $R_{\mathrm{f}}(\cdot)$	Conv1D(256, 15, 1)
Conv1D(256, 15, 1)	Conv1D(512, 5, 2)
Conv1D(512, 5, 2)	Conv1D(1024, 5, 2)
Conv1D(1024, 5, 2)	Deconv1D(512, 5, 2)
Deconv1D(512, 5, 2)	Deconv1D(256, 5, 2)
Deconv1D(256, 5, 2)	Conv1D(39, 15, 1)
音素予測モデル $R_{\mathrm{p}}(\cdot)$	
Conv1D(43, 15, 1)	
ドメイン識別モデル $D_{\mathrm{dc}}(\cdot)$	**話者認証モデル $D_{\mathrm{sv}}(\cdot)$**
Conv1D(512, 1, 1)	Conv1D(512, 1, 1)
Conv1D(512, 5, 1)	Conv1D(512, 5, 1)
Conv1D(512, 5, 1)	Conv1D(512, 5, 1)
Conv1D(1, 1, 1)	Conv1D(1, 1, 1)

本稿で用いる全ての DNN のアーキテクチャは，時間方向の 1D convolutional neural networks [16] であり，学習時の入力系列長を 128 フレームとする．特徴抽出モデル $R_{\mathrm{f}}(\cdot)$ は，13 次の MFCC とその動的特徴量から，256 次元の潜在変数を抽出する．音素予測モデル $R_{\mathrm{p}}(\cdot)$ は，潜在変数を用いて，43 次元の日本語音素事後確率を予測する．音声生成モデル $G(\cdot)$ は，音素事後確率から目的話者の 39 次元のメルケプストラム係数を予測する．DNN 学習時には，特徴量を次元ごとに平均 0，分散 1 に正規化する．ドメイン識別モデル $D_{\mathrm{dc}}(\cdot)$ は，潜在変数を用いて，2 つのドメイン $\mathcal{D}^{(\mathrm{O})}$ と $\mathcal{D}^{(\mathrm{M})}$ を識別する．話者認証モデル $D_{\mathrm{sv}}(\cdot)$ は，目的話者のメルケプストラム係数と $G(\cdot)$ の予測結果を識別する．全ての DNN の隠れ層の活性化関数は，leaky rectified linear (LReLU) [17] とする．出力層の活性化関数は，音声認識モデル $R(\cdot) = R_{\mathrm{p}}(R_{\mathrm{f}}(\cdot))$ では softmax 関数，$D_{\mathrm{dc}}(\cdot)$ では sigmoid 関数とする．過学習を防ぐために，全ての DNN の隠れ層に対して，ユニット脱落率を 0.5 とした dropout [18] を適用する．学習の収束高速化のために，$G(\cdot)$ の第 1 層を除く全ての隠れ層に batch normalization [19] を適用する．Table 1 に DNN アーキテクチャの詳細を示す．ここで，表中の “Conv1D(C_{out}, k, s)” と “Deconv1D(C_{out}, k, s)” はそれぞれ 1D convolution 及び 1D deconvolution 層を表し，C_{out}，k，s はそれぞれ出力チャネル数，convolution 演算のカーネルサイズとストライド幅である．

実験の初期設定として，多数話者コーパス $\mathcal{D}^{(\mathrm{M})}$ に含まれる全発話を用いた 1 エポックの事前学習により $R(\cdot)$ を構築する．事前学習の最適化アルゴリズムは，学習率を 0.01 とした AdaGrad [20] である．事前学習後，CSJ コーパスの学習データ以外の発話に対するフレームごとの音素認識率は，80.4%であった．本稿では，この音声認識モデルを用いた以下の 3 つの学習法を比較する．

Baseline: 固定された $R(\cdot)$ を用いた $G(\cdot)$ の学習 [1]

Prop. (GAN): $(\omega_{\mathrm{R}}, \omega_{\mathrm{G}}) = (0.0, 0.5)$ と設定した提案法による $R(\cdot)$ と $G(\cdot)$ の同時学習

Prop. (DAT-GAN): $(\omega_{\mathrm{R}}, \omega_{\mathrm{G}}) = (0.25, 0.5)$ と設定した提案法による $R(\cdot)$ と $G(\cdot)$ の同時学習

多対一音声変換モデルは，$\mathcal{D}^{(\mathrm{O})}$ に含まれる全発話を用いた 5 エポックの学習により構築する．提案法の学

Table 2　変換音声の自然性に関する MOS スコアと 95%信頼区間

	F2F	M2F
Baseline	2.703 ± 0.124	2.510 ± 0.113
Prop. (GAN)	**2.997 ± 0.131**	2.553 ± 0.116
Prop. (DAT-GAN)	**2.953 ± 0.125**	**2.747 ± 0.119**

Table 3　変換音声の話者類似性に関するプリファレンススコア

(a) 同性間の音声変換 (F2F) の結果

Method A	Score	Method B
Baseline	0.317 vs. **0.683**	Prop. (DAT-GAN)
Prop. (GAN)	0.387 vs. **0.613**	Prop. (DAT-GAN)

(b) 異性間の音声変換 (M2F) の結果

Method A	Score	Method B
Baseline	0.283 vs. **0.717**	Prop. (DAT-GAN)
Prop. (GAN)	0.373 vs. **0.627**	Prop. (DAT-GAN)

習では，目的話者以外の音素ラベル付き学習データ $(\boldsymbol{x}^{(\mathrm{M})}, \boldsymbol{l}^{(\mathrm{M})})$ が $\mathcal{D}^{(\mathrm{M})}$ からランダムに抽出される．全ての DNN に対する最適化アルゴリズムは，学習率を 0.01 とした AdaGrad である．

4.2　主観評価

クラウドソーシングによる評価システムを用いて，変換音声の自然性に関する 5 段階の mean opinion score (MOS) テストと，話者類似性に関するプリファレンス XAB テストを実施する．被験者の数は，それぞれ 30 名ずつである．

4.2.1　変換音声の自然性に関する評価結果

ここでは，各被験者は，ランダムに提示された 60 サンプルの変換音声（10 名の変換元話者，女性話者から女性話者 (F2F) と男性話者から女性話者 (M2F) の 2 通りの変換，3 手法) の自然性を 5 段階で評価する．

Table 2 に評価結果を示す．まず，同性間の音声変換 (F2F) では，“Prop. (GAN)” と “Prop. (DAT-GAN)” の両方が “Baseline” よりも有意に高いスコアを獲得している．しかしながら，異性間の音声変換 (M2F) では，“Baseline” と “Prop. (GAN)” のスコアの間に有意差はみられない．故に，話者間の違いが特に顕著になる異性間の変換において，GAN に基づく学習を適用するだけでは，多対一音声変換の変換音声の自然性が改善しないことが示唆された．一方で，“Prop. (DAT-GAN)” は，他の 2 手法よりも有意に高いスコアを獲得しており，DAT と GAN の両方に基づく提案法が，変換元話者の性別に依らずに変換音声の自然性を改善させることが示された．

4.2.2　変換音声の話者類似性に関する評価

ここでは，Table 2 の評価結果に基づき，(1) “Baseline” と “Prop. (DAT-GAN)，” そして (2) “Prop. (GAN)” と “Prop. (DAT-GAN)” を比較し，“Prop. (DAT-GAN)” の有効性を検証する．各被験者は，ランダムに提示された 40 サンプルの変換音声（10 名の変換元話者，F2F と M2F の 2 通りの変換，2 通りの比較) の話者類似性を評価する．話者類似性を評価するためのリファレンス音声として，学習データに含まれない目的話者の 1 発話を用いる．

Table 3 に評価結果を示す．この表より，“Prop. (DAT-GAN)” は，“Baseline” と “Prop. (GAN)” の両方に対して有意に高いスコアを獲得しており，この提案法の話者類似性改善に関する有効性が示された．

5　おわりに

本稿では，音素事後確率を用いた多対一音声変換の品質改善のための音声認識・生成モデルの同時敵対学習を提案し，実験的評価によりその有効性を示した．今後は，提案法のハイパーパラメータの影響の調査や，系列変換モデリング [21] の導入を検討する．

謝辞: 本研究は，JSPS 科研費 18J22090 の支援を受けた．

参考文献

[1] L. Sun *et al.*, *Proc. ICME,* Seattle, U.S.A., Jul. 2016.

[2] H. Miyoshi *et al.*, *Proc. INTERSPEECH,* pp. 1268–1272, Stockholm, Sweden, Aug. 2017.

[3] H. Zen *et al.*, *Speech Communication,* vol. 51, no. 11, pp.1039–1064, Nov. 2009.

[4] T. Toda *et al.*, *IEEE Trans. on ASLP,* vol. 15, no. 8, pp.2222–2235, Nov. 2007.

[5] Y. Ganin *et al.*, *Journal of Machine Learning Research,* vol. 17, no. 59, pp. 1-35, Apr. 2016.

[6] I. Goodfellow *et al.*, *Proc. NIPS,* pp. 2672– 2680 Montreal, Canada, Dec. 2014.

[7] Y. Saito *et al.*, *IEEE/ACM Trans. on ASLP,* vol. 26, no. 1, pp. 84–96, Jan. 2018.

[8] Q. Wang *et al.*, *Proc. ICASSP,* pp. 4889–4893, Alberta, Canada, Apr. 2018.

[9] J.-C. Chou *et al.*, *Proc. INTERSPEECH,* pp. 501–505, Hyderabad, India, Sep. 2018.

[10] M. Arjovsky *et al.*, *arXiv:1701.07875,* 2017.

[11] K. Maekawa *et al.*, *Proc. LREC,* pp. 947–952, Athens, Greece, May 2000.

[12] K. Sugiura *et al.*, *Advanced Robotics,* vol. 29, no. 7, pp. 449–456, Mar. 2015.

[13] A. Kurematsu *et al.*, *Speech Communication,* vol. 9, no. 4, pp. 357–363, Aug. 1990.

[14] M. Morise *et al.*, *IEICE Trans. on Information and Systems,* vol. E99-D, no. 7, pp. 1877–1884, Jul. 2016.

[15] M. Morise, *Speech Communication,* vol. 84, pp. 57–65, Nov. 2016.

[16] O. Abdel-Hamid *et al.*, *IEEE/ACM Trans. on ASLP,* vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[17] A. L. Maas *et al.*, *Proc. ICML,* Atlanta, U.S.A., Jun. 2013.

[18] N. Srivastava *et al.*, *The Journal of Machine Learning Research,* vol. 15, no. 1, pp. 1929–1958, Apr. 2014.

[19] S. Ioffe *et al.*, *Proc. ICML,* Lille, France, Jul. 2015.

[20] J. Duchi *et al.*, *Journal of Machine Learning Research,* vol. 12, pp. 2121–2159, Jul. 2011.

[21] J.-X. Zhang *et al.*, *IEEE/ACM Trans. on ASLP,* vol. 27, no. 3, pp. 631–644, Jan. 2019.