HumanACGAN：人間の知覚を補助分類器に用いた条件付き敵対的生成ネットワークと音素知覚における評価*

☆上田 陽太（東大院・情報理工），藤井 一貴（徳山高専），齋藤 佑樹，
高道 慎之介（東大院・情報理工），馬場 雪乃（筑波大学），猿渡 洋（東大院・情報理工）

1 はじめに

敵対的生成ネットワーク (generative adversarial network: GAN) [1] は最も強力な深層生成モデルの一つであり，音声モデリングに利用されている [2]．GAN は生成器，識別器の deep neural network (DNN) のセットから構成される．生成器は識別器をだますように学習され，識別器は実データと生成データを識別するように学習される．この学習を繰り返すことで生成器は実データの分布を表現し，実データの分布に従うデータをランダムに生成できる．GAN は実データ分布の外側を表現できない他方で，人間は実データ分布の外側のデータであってもそれを自然なものとして許容することがある．音声の知覚では，音声が実データ分布 (つまり，実在する人間が発話した音声) に含まれない合成音声や加工音声であっても人間はその音声の自然性を許容できる．本研究では，この人間が自然性を許容できる分布を知覚分布 (human-acceptable distribution) と呼称する．HumanGAN [3] は GAN の識別器として人間を用いることで，この知覚分布の表現を可能にした．**図 1** の上段に GAN と HumanGAN の比較を示す．HumanGAN は人間を，生成データが与えられたとき事後確率 (自然性許容度) の差を出力するブラックボックスシステムとみなす．DNN で記述される生成器は，人間で記述される識別器を用いて誤差逆伝播法 (backpropagation) によって学習され，学習した生成器は知覚分布を表現できる．しかし，HumanGAN の生成器は条件付き知覚分布 (すなわち，ある条件が人間に与えられたときの知覚分布) を表現できないため，テキスト音声合成や音声変換などの実用的な生成器を学習することができない．我々は，GAN が条件付き GAN に拡張された [4, 5] ように，HumanGAN が条件つきモデルに拡張可能であると考える．

本論文では，HumanGAN の生成器を，所望のクラスラベルにより条件付きで学習させ，クラスごとの知覚分布を表現するため，人間の知覚評価を用いた DNN で記述された条件付き生成器を学習する HumanACGAN を提案する．この実現のため，DNN で記述される補助分類器を用いる auxiliary classifier GAN (ACGAN) [5] の考えを導入する．**図 1** において 4 つの GAN (GAN，ACGAN，HumanGAN，HumanACGAN) を比較する．ACGAN は GAN の識別器に加えて補助分類器を用いて DNN で記述された条件付き生成器を学習する．本論文では識別器と補助分類器を人間の知覚評価で置き換える．人間は識別器として自然性許容度を評価し，補助分類器としてクラス許容度を評価する．実験により音素の知覚に関して HumanACGAN を評価する．実験の結果，1) 音素による条件付きの人間の知覚分布が実データのものより広いこと，2) HumanACGAN は，生成器を条件付きでの人間の知覚分布を学習させることを報告する．

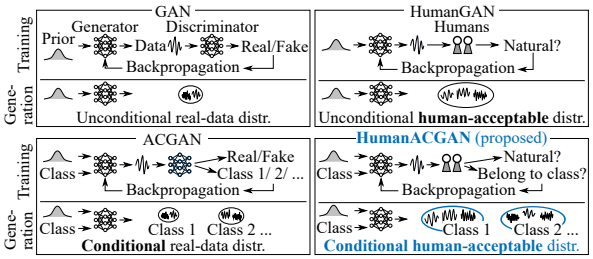

Fig. 1 GAN，ACGAN，HumanGAN，提案する HumanACGAN の比較

2 関連研究

2.1 ACGAN

ACGAN [5] はクラスのラベルによって条件付けられる実データの分布を表現する条件付き生成器を学習する．これを実現するために，ACGAN は GAN の DNN で記述された識別器に加えて補助分類器を用いる．生成器 $G(\cdot)$ は既知の事前分布（例えば一様分布）に従う乱数 $\boldsymbol{z} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n, \cdots, \boldsymbol{z}_N]$ をクラスラベル $\boldsymbol{c} = [c_1, \cdots, c_n, \cdots, c_N]$ に応じて $\hat{\boldsymbol{x}} = [\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_n, \cdots, \hat{\boldsymbol{x}}_N]$ に変換する．つまり，$\hat{\boldsymbol{x}}_n = G(\boldsymbol{z}_n, \boldsymbol{c}_n)$ である．N は実データの数を表す．ここで，実データを $\boldsymbol{x} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n, \cdots, \boldsymbol{x}_N]$ とする．実データ $\boldsymbol{x}$ も，クラスラベル $\boldsymbol{c}$ に対応する．生成器の学習には識別器 $D_{\mathrm{S}}(\cdot)$ と補助分類器 $D_{\mathrm{C}}(\cdot)$ が用いられる．$D_{\mathrm{S}}(\cdot)$ は $\boldsymbol{x}_n$ または $\hat{\boldsymbol{x}}_n$ を入力として受け取り，入力データが実データである事後確率を出力する．$D_{\mathrm{C}}(\cdot)$ は $\boldsymbol{x}_n$ または $\hat{\boldsymbol{x}}_n$ を入力として受け取り，入力データが属するクラスに関する事後確率を出力する．学習に用いられる目的関数は以下の様に記述される．

$$L_{\mathrm{S}} = \sum_{n=1}^{N} \log D_{\mathrm{S}}(\boldsymbol{x}_n) + \sum_{n=1}^{N} \log\left(1 - D_{\mathrm{S}}(\hat{\boldsymbol{x}}_n)\right) \quad (1)$$

$$L_{\mathrm{C}} = \sum_{n=1}^{N} \log D_{\mathrm{C}}(\boldsymbol{x}_n, \boldsymbol{c}_n) + \sum_{n=1}^{N} \log\left(D_{\mathrm{C}}(\hat{\boldsymbol{x}}_n, \boldsymbol{c}_n)\right) \quad (2)$$

*HumanACGAN: conditional generative adversarial network using human-based auxiliary classifier and its evaluation in representing conditional distribution of phoneme perception, by Yota Ueda, (The University of Tokyo), Kazuki Fujii (National Institute of Technology, Tokuyama College), Yuki Saito, Shinnosuke Takamichi (The University of Tokyo), Yukino Baba (University of Tsukuba), and Hiroshi Saruwatari (The University of Tokyo).

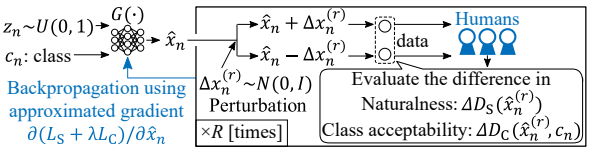

Fig. 2　HumanACGAN の学習

L_S は GAN の目的関数で，L_C は条件付き生成器の学習に用いられる．生成器は $-L_S + \lambda L_C$ を最大化するように学習する．ここで，λ はハイパーパラメータである．

ACGAN の生成器はクラスラベルに対応する実データの分布を表現する．しかし，人間の知覚分布は実データの分布より広いため [3]，ACGAN は知覚分布のすべての範囲を表現することはできない．

2.2　HumanGAN

HumanGAN [3] は実データの分布より広い人間の知覚分布を表現するために提案された．DNN で記述された条件付きなしの生成器は GAN の識別器の代わりに人間の知覚評価を用いて学習する．生成器 $G(\cdot)$ は $\boldsymbol{z}_n$ を $\hat{\boldsymbol{x}}_n$ に条件付けなしで変換する．人間による識別器 $D_S(\cdot)$ は人間の知覚評価を扱うものとして定義され，入力データが自然であるかどうか (自然性許容度) を評価する．すなわち $D_S(\cdot)$ は $\hat{\boldsymbol{x}}_n$ を入力として受け取り，入力の自然性が知覚的にどの程度許容できるかを 0 から 1 の値で事後確率として出力する．目的関数は以下のように記述される．

$$L_S = \sum_{n=1}^{N} D_S(\hat{\boldsymbol{x}}_n) \tag{3}$$

生成器は L_S を最大化するように学習する．$G(\cdot)$ のモデルパラメータ θ は $\theta \leftarrow \theta + \alpha \partial L_S/\partial\theta$ のように反復的に更新される．ここで α は学習係数である．また，$\partial L_S/\partial\theta = \partial L_S/\partial \hat{\boldsymbol{x}}_n \cdot \partial \hat{\boldsymbol{x}}_n/\partial\theta$ である．

この勾配のうち，$\partial \hat{\boldsymbol{x}}_n/\partial\theta$ は解析的に推定可能であるが，$D_S(\cdot)$ が人間による知覚評価を含むため，$\partial L_S/\partial \hat{\boldsymbol{x}}_n$ の解析的推定は困難である．このため HumanGAN は natural evolution strategy (NES) [6] を勾配を近似的に推定するために用いる．微小な摂動 $\Delta\boldsymbol{x}_n^{(r)}$ をガウス分布 $\mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ からランダムに生成し，生成データ $\hat{\boldsymbol{x}}_n$ を摂動する．r は摂動のインデックス $((1 \le r \le R))$ である．σ，$\boldsymbol{I}$ はそれぞれ定数の標準偏差，単位行列である．次に人間は摂動を与えられた 2 つのデータ $\{\hat{\boldsymbol{x}}_n + \Delta\boldsymbol{x}_n^{(r)}, \hat{\boldsymbol{x}}_n - \Delta\boldsymbol{x}_n^{(r)}\}$ を観測し，その事後確率の差

$$\Delta D_S(\hat{\boldsymbol{x}}_n^{(r)}) \equiv D_S(\hat{\boldsymbol{x}}_n + \Delta\boldsymbol{x}_n^{(r)}) - D_S(\hat{\boldsymbol{x}}_n - \Delta\boldsymbol{x}_n^{(r)}) \tag{4}$$

を評価する．$\Delta D_S(\hat{\boldsymbol{x}}_n^{(r)})$ は -1 から 1 までの値をとる．例えば，人間が $\hat{\boldsymbol{x}}_n + \Delta\boldsymbol{x}_n^{(r)}$ の方が $\hat{\boldsymbol{x}}_n - \Delta\boldsymbol{x}_n^{(r)}$ より自然性を許容できると感じた場合，人間は $\Delta D_S(\hat{\boldsymbol{x}}_n^{(r)}) = 1$ を返す．$\partial L_S/\partial\hat{\boldsymbol{x}}$ は次のように推定される [6]．

$$\frac{\partial L_S}{\partial \hat{\boldsymbol{x}}_n} = \frac{1}{2\sigma^2 R} \sum_{r=1}^{R} \Delta D_S(\hat{\boldsymbol{x}}_n^{(r)}) \cdot \Delta\boldsymbol{x}_n^{(r)} \tag{5}$$

3　HumanACGAN

3.1　生成器の学習

本節では HumanACGAN を提案する．**図 2** に示すように，HumanGAN は単一クラスの知覚分布を表現するに対し，HumanACGAN は条件付きの生成器を持つことで複数クラスの各知覚分布を表現できる．ACGAN と同じように，HumanACGAN は生成器，識別器，補助分類器からなる．DNN で記述された生成器 $G(\cdot)$ は ACGAN と同じもので，$\boldsymbol{z}_n$ からラベル $\boldsymbol{c}_n$ に応じてデータ $\hat{\boldsymbol{x}}_n$ を生成する．識別器と補助分類器は人間の知覚の評価を取り入れるために，以降の様に再定義される．人間による識別器 $D_S(\cdot)$ は HumanGAN の識別器と同様，入力データのクラスに依存しない自然性許容度を事後確率として出力する．加えて，人間による補助分類器 $D_C(\cdot)$ は入力データが当該クラス $\boldsymbol{c}_n$ へ属するかどうか（クラス許容度）を評価する．$D_C(\cdot)$ は生成データ $G(\cdot)$ とクラスラベル $\boldsymbol{c}_n$ を入力として受け取り，入力データが知覚的に当該クラスとして許容できる事後確率を出力する．目的関数は

$$L_S = \sum_{n=1}^{N} D_S(\hat{\boldsymbol{x}}_n) \tag{6}$$

$$L_C = \sum_{n=1}^{N} D_C(\hat{\boldsymbol{x}}_n, \boldsymbol{c}_n) \tag{7}$$

$G(\cdot)$ のモデルパラメータ θ は $L_S + \lambda L_C$ を最大化するように学習され，以下の様に反復的に更新される．

$$\theta \leftarrow \theta + \alpha \frac{\partial(L_S + \lambda L_C)}{\partial\theta} \tag{8}$$

$$\frac{\partial(L_S + \lambda L_C)}{\partial\theta} = \left(\frac{\partial L_S}{\partial \hat{\boldsymbol{x}}_n} + \lambda \frac{\partial L_C}{\partial \hat{\boldsymbol{x}}_n}\right) \cdot \frac{\partial \hat{\boldsymbol{x}}_n}{\partial\theta} \tag{9}$$

HumanGAN と同じように，$\partial \hat{\boldsymbol{x}}_n/\partial\theta$ は推定可能であるが，$\partial L_S/\partial \hat{\boldsymbol{x}}_n$ と $\partial L_C/\partial \hat{\boldsymbol{x}}_n$ は推定不可能である．そこで NES を用いて両方の勾配を近似する手法を提案する．人間は摂動を与えられた 2 つのデータ $\{\hat{\boldsymbol{x}}_n + \Delta\boldsymbol{x}_n^{(r)}, \hat{\boldsymbol{x}}_n - \Delta\boldsymbol{x}_n^{(r)}\}$ を観測し，2 種類の事後確率の差を評価する．1 つ目は HumanGAN と同様に，人間は “どの程度自然性許容度の観点で差があるか” を評価し，事後確率の差 $\Delta D_S(\hat{\boldsymbol{x}}_n^{(r)})$ を回答して式 (4) により $\partial L_S/\partial \hat{\boldsymbol{x}}_n$ を近似する．2 つ目はクラス固有の質問で，“どの程度クラス許容度の観点で差があるか” を問う．$\Delta D_C(\hat{\boldsymbol{x}}_n^{(r)}, \boldsymbol{c}_n)$ は以下の様に定義される．

$$\Delta D_C(\hat{\boldsymbol{x}}_n^{(r)}, \boldsymbol{c}_n) \equiv D_C\left(\hat{\boldsymbol{x}}_n + \Delta\boldsymbol{x}_n^{(r)}, \boldsymbol{c}_n\right) - D_C\left(\hat{\boldsymbol{x}}_n - \Delta\boldsymbol{x}_n^{(r)}, \boldsymbol{c}_n\right) \tag{10}$$

HumanGAN と同様に $\partial L_S/\partial \hat{\boldsymbol{x}}_n$ は $\Delta D_S(\hat{\boldsymbol{x}}_n^{(r)})$ を用いて推定される．$\Delta D_S(\cdot)$ とは異なり，$\Delta D_C(\cdot)$ はクラス固有の事後確率差である．つまり $\hat{\boldsymbol{x}}_n + \Delta\boldsymbol{x}_n^{(r)}$ が $\hat{\boldsymbol{x}}_n + \Delta\boldsymbol{x}_n^{(r)}$ よりも提示されたクラスとして許容できる場合 $\Delta D_C(\hat{\boldsymbol{x}}_n^{(r)}, \boldsymbol{c}_n) = 1$ となる．$\Delta D_C(\hat{\boldsymbol{x}}_n^{(r)}, \boldsymbol{c}_n)$ は -1 から 1 までの範囲をとり，$\partial L_C/\partial\hat{\boldsymbol{x}}$ は以下の式で推定される．

$$\frac{\partial L_C}{\partial \hat{\boldsymbol{x}}_n} = \frac{1}{2\sigma^2 R} \sum_{r=1}^{R} \Delta D_C\left(\hat{\boldsymbol{x}}_n^{(r)}, \boldsymbol{c}_n\right) \cdot \Delta\boldsymbol{x}_n^{(r)} \tag{11}$$

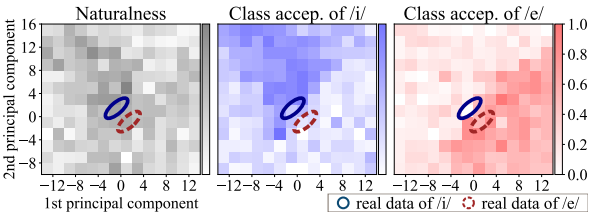

Fig. 3　自然性許容度と音素/i/，/e/のクラス許容度 (Class accep.) の事後確率

ここで，HumanACGAN の補助分類器は明示的に分類問題を解いているわけではないことに注意する必要がある．言い換えれば，多クラスの確率関数 (ACGAN での softmax 関数) ではなく，各クラスへの許容度を推定するだけでよい．

3.2　技術的な制限

HumanGAN [3] は mode collapse [7]，勾配消失，データサイズや次元のスケーラビリティの問題がある．これらの問題は HumanACGAN でも同様に発生する．そこで，本実験は HumanGAN の論文 [3] の実験における，生成器パラメータと生成データの初期化方法を参考にする．初めに事後確率のヒストグラムを推定し，ヒストグラムを参照して θ の初期値を設定した．加えて，主成分分析を施してデータの次元を削減した．

4　実験的評価

本節では日本語の音素をクラスとして HumanACGAN の有効性を評価する．

4.1　実験の準備

日本語の 2 つの音素/i/と/e/を用いた．基本的に HumanGAN の論文 [3] に記載されている実験の準備に従った．データには 199 人の女性話者の発話からなる JVPD [8] コーパスを用いた．音声特徴量を抽出する前に，音量の正規化を行い，16 kHz にダウンサンプリングした．WORLD ボコーダー [9, 10] を用いて 5 ms ごとに音声波形から 513 次元の対数スペクトル包絡，基本周波数，非周期性成分を抽出した．Julius [11] による音素アライメントを行って音素/i/と/e/の音響特徴量を用いた．対数スペクトル包絡に主成分分析を施し，第 1 主成分，第 2 主成分を使用した．この 2 次元の主成分を平均 0，分散 1 となるように正規化した．人間によって評価される音声波形は以下の方法で合成した．まず，第 1 主成分と第 2 主成分が生成器から出力され，これを逆正規化した．その他の特徴量，つまり基本周波数と非周期性成分は全話者の平均を用いた．これらの音響特徴量は 1 フレーム (5 ms) に対応する．次に，知覚評価を容易にするためにこの特徴量を 200 フレーム分コピーして WORLD ボコーダーを用いて 1 秒の音声波形を合成した．

4.2　知覚分布と実データの分布の違い

初めに人間の知覚分布が実データの分布より広いことを確認した．自然性許容度の評価，クラス許容度の評価の 2 つのテストを行った．この実験では，クラス許容度はデータが提示された音素に聞こえるか

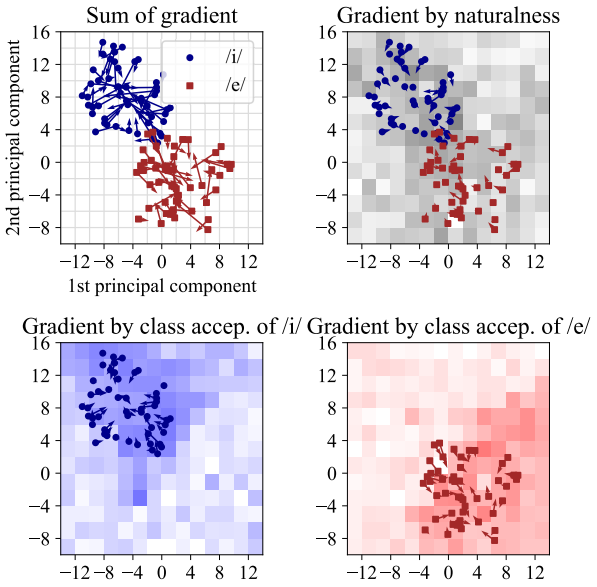

Fig. 4　生成データ (点) と推定された勾配 (矢印)．“Class accep.” はクラス許容度を表す．

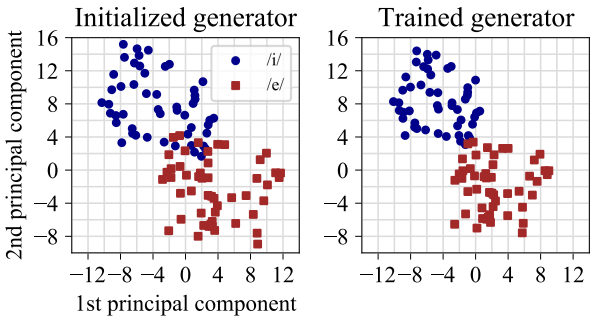

Fig. 5　初期化された生成器と学習済みの生成器からの生成データ．データ点の色は**図 3** のクラス許容度の色に対応している．**図 3** の事後確率を参照することで学習によってデータが事後確率の高い方向へ動いていることがわかる．

どうかを示す．2 次元空間をグリッドに区切り，それぞれのグリッドについて音声波形を合成した．そして人間に音声波形を提示し，自然性許容度を 1 (悪い) から 5 (良い) までの 5 段階で評価させた．次に，音声波形と音素ラベル (/i/ または /e/) を提示し，クラス許容度を同じように 5 段階で評価させた．得られた 1 から 5 までの評価は事後確率 ($D_{\mathrm{S}}(\hat{\boldsymbol{x}}_n)$ または $D_{\mathrm{C}}(\hat{\boldsymbol{x}}_n, \boldsymbol{c}_n)$) の 0.00, 0.25, 0.50, 0.75, 1.00 にそれぞれ対応させた．クラウドソーシングサービスである Lancers を利用して評価を行った．それぞれのグリッドは少なくとも 5 人以上に評価されており，評価で得られた事後確率を平均している．評価者の総人数は 105 であった．

図 3 に結果を示す．**4.1 節**で述べられているように，実データは平均 0，分散 1 となるように正規化されている．しかし，**図 3** が示すように，人間の知覚分布，すなわち色の濃い部分は自然性許容度もクラス許容度も実データの分布より広いことがわかる．つまり，ACGAN が十分に表現できないこの分布を表現するために提案する HumanACGAN が必要であることが示される．

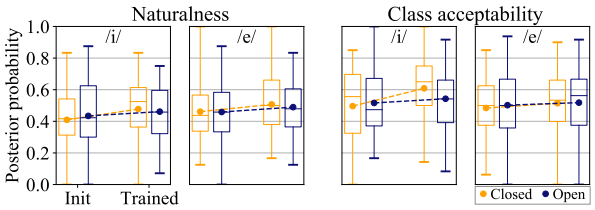

Fig. 6　初期化された生成器 (“Init”) および学習済みの生成器 (“Trained”) からの生成データの事後確率

4.3　学習中の生成データの遷移

次に，HumanACGAN を学習させ，生成データと推定された勾配を定性的に評価した．学習の前に 2 次元の一様分布 $U(0,1)$ から乱数を生成し，学習中はこの乱数を固定して使用した． 生成データのうち半分にクラスラベル/i/を与え，残りの半分にクラスラベル/e/を与えた．生成器は 2 ユニットの入力層，2×4 ユニットの sigmoid 隠れ層，2 ユニットの線形出力層からなる feed-forward neural network とした．出力層は 2 次元の one-hot のクラスラベルベクトルで条件付けした．以下 2 つの条件を満たすまで生成器のランダムな初期化を繰り返した．条件は，初期化された生成器からの生成データが自然性許容度とそれぞれのクラス許容度の事後確率の高い範囲を覆っていること，分類の性能を確かめるためそれぞれのクラスの出力が互いに重なるようになることである．実験による経験からハイパーパラメータを $\lambda=2$，学習率を $\alpha=0.0005$ に設定した．実装には Chainer を使用した．生成データの総数 N，摂動の回数 R，学習の反復回数，NES の標準偏差 σ はそれぞれ 50，5，4，2.0 に設定した．HumanGAN の学習において 2 種類の知覚評価を行った．初めに人間に摂動を与えた 2 つのデータを与え，どちらが自然かを 1 : 1 番目，3 : 同等，5 : 2 番目の 5 段階で評価させた．次に，摂動を与えた 2 つのデータと音素のラベル (/i/ or /e/) を与え，どちらが提示された音素に聞こえるかを 1 : 1 番目，3 : 同等，5 : 2 番目の 5 段階で評価させた．得られた 1 から 5 までの評価は事後確率 ($\Delta D_{\mathrm{S}}(\hat{\boldsymbol{x}}_n)$ or $\Delta D_{\mathrm{C}}(\hat{\boldsymbol{x}}_n, \boldsymbol{c}_n)$) の 0.00, 0.25, 0.50, 0.75, 1.00 にそれぞれ対応させる．

図 4 はそれぞれのデータでの勾配を示している．**図 3** の事後確率が参考のために描画されているが，学習には用いられていない．自然性許容度，クラス許容度の両方の勾配が色の濃い方へ向いていることが確認できる．これにより式 (5)，(11) に示されている勾配が正しく推定されていることがわかる．**図 5** は初期化された生成器，学習した生成器から出力されたデータを示す．**図 3** に示された事後確率より，学習によってデータが移動し，自然性許容度，クラス許容度両方の事後確率が高い方へ移動していることが言える．これは目的関数式 (6)，(7) が生成器を条件付きの知覚分布を表現させることができることを示している．

4.4　学習による事後確率の上昇

最後に，学習によって自然性許容度の事後確率 $D_{\mathrm{S}}(\cdot)$ とクラス許容度の事後確率 $D_{\mathrm{C}}(\cdot)$ が上昇したことを定量的に確認する．Close データと open データの 2 種類のデータを用意した．Closed データは学習に用いられた乱数から生成されたデータで，open データは学習に用いられていない乱数から生成されたデータである．Closed，open それぞれのデータの事後確率を **4.2 節**と同じように評価した．評価者の総人数は 160 であった．

図 6 は事後確率の箱ひげ図を示す．学習によって自然性許容度，クラス許容度の事後確率がともに closed，open データ両方で上昇していることがわかる．よって，HumanACGAN により学習された生成器は条件付きの知覚分布を表現できることが定量的に示される．

5　まとめ

本論文では，人間の知覚分布を条件付きで表現する HumanACGAN を提案した．HumanACGAN では，従来の ACGAN の識別器と補助分類器を人間の知覚評価で置き換え，DNN で記述された条件付き生成器を自然性許容度，クラス許容度の 2 つの観点で学習する．定量的．定性的な評価を行い，HumanACGAN が条件つきの知覚分布を表現できることを示した．今後は，学習に用いるデータ量や次元に対するスケーラビリティを拡張する．

謝辞：本研究開発は総務省 SCOPE(受付番号 182103104) の委託を受けて実施した．

参考文献

[1] I. Goodfellow et al., “Generative adversarial nets,” in *Proc. NIPS*, Montreal, Canada, Dec. 2014, pp. 2672–2680.

[2] Y. Saito et al., “Statistical parametric speech synthesis incorporating generative adversarial networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, Jan. 2018.

[3] K. Fujii et al., “HumanGAN: generative adversarial network with human-based discriminator and its evaluation in speech perception modeling,” in *Proc. ICASSP*, Barcelona, Spain, May 2020, pp. 6239–6243.

[4] M. Mirza, S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014.

[5] A. Odena et al., “Conditional image synthesis with auxiliary classifier GANs,” in *Proc. ICLR*, Vancouver, Canada, Apr. 2018.

[6] A. Ilyas et al., “Black-box adversarial attacks with limited queries and information,” in *Proc. ICML*, vol. 2, Stockholm, Sweden, Jul. 2018, pp. 2137–2146.

[7] I. Goodfellow, “NIPS 2016 tutorial: Generative adversarial networks,” in *Proc. NIPS*, Barcelona, Spain, Dec. 2016. [Online]. Available: https://arxiv.org/abs/1701.00160

[8] “Vowel database: Five Japanese vowels of males, females, and children along with relevant physical data (JVPD),” http://research.nii.ac.jp/src/en/JVPD.html.

[9] M. Morise et al., “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, Jul. 2016.

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

[11] A. Lee et al., “Julius — an open source real-time large vocabulary recognition engine,” in *Proc. EUROSPEECH*, Aalborg, Denmark, Sep. 2001, pp. 1691–1694.