DEIM2021 G33-3

差分プライバシーと有用性を両立する GANを用いた合成音声データセットの生成

河　明宗† 吉川　正俊† 曹　洋†

† 京都大学大学院 情報学研究科 社会情報学専攻 分散情報システム分野
〒 606-8501 京都府京都市左京区吉田本町 京都大学医学部構内先端化学研究棟 5 階
E-mail: †kawa.akimune.67x@st.kyoto-u.ac.jp, ††{yoshikawa,yang}@i.kyoto-u.ac.jp

あらまし　音声ベースのヒューマンコンピューターインタラクションが盛んになり, 音声データの収集や共有といった取り組みの重要度が高まる中で, プライバシー漏洩に関しての懸念が高まっている. 特に保護しなければならないのは, 生体認証にも用いられる声紋である. 声紋に関してのプライバシーを保護する手法はいくつか提案されているが, いずれも有用性や厳密性, 実用可能性に課題がある. 本研究では, そのような課題を解決する手法の一つとして, 声紋に関して匿名化を施した偽の合成音声データセットの生成を, GAN を応用して実現する手法の方針を示す.

キーワード　データ生成, 差分プライバシー, GAN, 音声データ

1 はじめに

近年, Amazon Alexa や Google アシスタント, Apple の Siri といった音声でのヒューマンコンピューターインタラクションを実現する機能やデバイスの開発が盛んになっている. そういったインタラクションの開発のためには, 実際の音声データの利用が必要であり, 音声データの収集や共有は重要な取り組みとして注目が高まっている.

しかし, 音声データの収集や共有を行うにあたってはプライバシーの漏洩が大きな懸念点となる. 音声データには様々な識別情報が含まれるが, その中でも生体識別子である声紋が漏洩することは大きな問題に発展する可能性がある. 声紋は個人を一意に識別可能な情報と考えられており, 攻撃者に知られることで特定に繋がる可能性がある. また, 様々なサービスにおいて声紋認証が実用化されている昨今において, 例え特定に繋がらなかったとしても, 情報漏洩が被害に繋がる可能性がある.

音声データ中の声紋に関するプライバシー保護に取り組んだ事例として, Han ら (2020) [6] の研究がある. Han らは, 声紋のデータに関して, プライバシーの厳密な定義がなされておらず, プライバシーと有用性のトレードオフも表現できていないという問題に着目した. そして, 差分プライバシーを拡張することで声紋データのプライバシーの厳密な指標であるボイス識別不能性を新しく定義した. また, ボイス識別不能性を満たすような音声データセットを生成する手法を提案した. ところが, この手法では元のデータに対して声紋データのみを入れ替えるという手法でデータセットを生成しており, これには以下の問題がある. 1) 声紋と個人を紐付けておくことに意味がある分析を行うことができないデータを生成してしまう. 例えば, 性別や年齢ごとに声紋にどのような違いがあるかを分析したい時に, 全ての個人と声紋の情報が入れ替えられた状況では分析を行えない. 2) 声紋自体が非常にセンシティブなデータであるため, 誰かの声紋がデータとして公開されているという状況そのものに危険性がある. 3) 計算量や学習のためのコストが大きい. Han らは声紋それぞれを別個に匿名化する手法と, あらかじめ匿名化しておいた声紋を学習させた合成モデルを用いて匿名化を行う手法を示したが, 前者はデータ数 n に対して計算量が $O(n^2)$ と大きく, 後者は匿名化された声紋を準備する必要があり手間が大きい.

ところで, プライバシーを保護した上でデータの共有を行う際に用いる手法として, GAN を用いて性質の近いがプライバシーが保たれる程度に差分のある偽の合成データを生成するという手法が近年多く提案されており注目を集めている. Xie ら (2018) [19] は, GAN で学習された生成分布の密度が訓練データに集中してしまい, 訓練データが容易に予測できてしまうという問題に着目し, 学習過程で勾配にノイズを加えることで, 差分プライバシーを実現する DPGAN を提案した. Jordon ら (2019) [21] は, Private Aggregation of Teacher Ensembles (PATE) フレームワークの微分が不可能であるという問題を解決した上で GAN に適用し, 差分プライバシーを確保する合成データを作成する手法を提案した. そして, 様々なデータセットやアルゴリズムにおいて Xie ら (2018) のモデルを上回る性能を達成した. Xu ら (2019) [20] は, 条件付き GAN(CTGAN) という手法を提案し, ベイジアンネットワークを用いた手法と GAN を用いた手法のどちらのベースラインについても, それを上回る性能を発揮することを示した. また, 以上の研究は全てテーブルデータに対するものであるが, 顔画像データに対して GAN を用いた合成データセット生成を行っている研究 [17] もある.

本研究では, ボイス識別不能性を用いて差分プライバシーを根底としたプライバシー保証を実現しつつ, Han らのデータセット生成手法の問題点を解消するような生成モデルを, GAN を用いて実現することを目的とする.

本論文の流れは以下の通りである. 2 章では, 本研究の目的に

ついて詳しく述べて整理する．3 章では，本研究の前提となる声紋情報，差分プライバシー，敵対生成ネットワークについて説明する．4 章では，関連研究について説明する．5 章では，提案手法の方針について説明する．6 章では全体をまとめる．

2 目　　的

この章では本研究の目的について詳しく述べる．

2.1 プライバシー保護の対象

音声には生理学上の識別子と，発話内容などの行動に伴う識別子の二種類の識別子が含まれている．行動に伴う識別子は学習の際に重要なデータであるため，本研究ではプライバシー保護の対象ではないものとする．本研究においてプライバシー保護の対象とするのは，生理学上の識別子である声紋である．

声紋は不変性と自然さという 2 つの特徴をもつ．不変性とは，声紋が周りの環境や発話者の心理的状態によって変化しないという性質である．自然さとは，声紋は音声の自然な特性であるため，特定のシナリオに限定されることがないという性質である．

2.2 問 題 設 定

本研究の目的は，音声データを公開する際に，発話の内容などはそのままに，声紋に関してのプライバシーを保護することを目的としている．すなわち，本来の音声データから偽の音声データを生成し，その偽の音声データでは声紋から個人が特定不可能になっていることを目指す．図 1 では，本研究において実現する音声データの公開の全体の流れを示している．

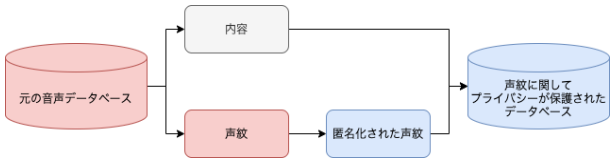

図 1　本研究において実現する音声データの公開の全体の流れ

3 基 本 事 項

この章では，本研究の前提となる基本的な事項について説明する．

3.1 差分プライバシー

差分プライバシーは，Dwork ら (2004) [3] によって提案された手法であり，統計量から個人のプライバシーを侵害しようとする攻撃者に対して，統計量に有用性を保つよう設計された適切なノイズを加えることでプライバシーを保護することができる．

定義 1 ((ϵ, δ)-差分プライバシー)**.** $\epsilon > 0, \delta \in [0, 1)$ とする．分布 ρ_D が (ϵ, δ)-差分プライバシーを満たすとは，任意の隣接するデータセット D, D'，任意の集合 $A \subset \Theta$ に対して，以下の式が成り立つことをいう．

$$\Pr_{\theta \sim \rho_D}[\theta \in A] \leqq \mathrm{e}^{\varepsilon} \Pr_{\theta \sim \rho_{D'}}[\theta \in A] + \delta$$

ここで，ϵ はプライバシーレベルを表す．ϵ が小さいほど，D, D' の分布の差が大きくなり，加えられるノイズが大きいため，より強いプライバシーを保証していると言える．また，δ はどの程度の確率で $O(\epsilon)$-差分プライバシーを満たさないかを決定している．そのため，δ はデータセットのサイズの多項式の逆数より小さくすることが推奨されている．

3.2 ボイス識別不能性

Han らは，2 つの音声データの声紋が識別不可能であることを，以下のように定式化した．

定義 2 (ボイス識別不能性)**.** 機構 K は，任意の 2 つの声紋 $x, x' \in \mathcal{X}$ について以下が成り立つとき，ϵ-ボイス識別不能性を満たす．

$$\frac{\Pr(\tilde{x} \mid x)}{\Pr(\tilde{x} \mid x')} \leqq e^{\epsilon d_{\mathcal{X}}(x, x')}$$

$$d_{\mathcal{X}} = \frac{\arccos(\cos\ similarity\ < x, x' >)}{\pi}$$

ここで，$\mathcal{X}$ は全ての声紋の集合，$d_{\mathcal{X}}$ は角距離，cos $similarity$ はコサイン類似度を表す．

3.3 Generative Adversarial Networks (GAN)

GAN は Goodfellow ら (2014) [5] によって提案された生成モデルであり，データから特徴を学習することで，実在しないデータを生成したり，存在するデータの特徴に沿って変換をおこなったりすることができる．GAN のアーキテクチャはノイズ入力から偽のデータを出力するモデル G と，G の出力もしくは真のデータを入力に受け取り，データの真偽を判定するモデル D からなる．G は D を欺くため，D はより正確に真偽を判定するためという相反する目的のもとに学習を行うことで，まるで本物かのようなデータ生成を実現可能であるのが特徴である．$p_{\mathbf{z}}(\mathbf{z})$ を G への入力ノイズの分布，$p_{\mathrm{data}}(\mathbf{x})$ を真のデータの分布とすると，GAN は次のような価値関数 $V(G, D)$ におけるミニマックスゲームとして定式化される．

$$\min_G \max_D V(G, D) = E_{\mathbf{x} \sim p_{\mathrm{data}}(\mathbf{x})}[\log(D(\mathbf{x}))] + E_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]. \quad (1)$$

4 関 連 研 究

この章では，音声データのプライバシー保護に関する先行研究や，GAN を用いて合成データセットを生成することに取り組んだ先行研究を紹介する．

4.1 音声データのプライバシー保護の先行研究

音声データに含まれる声紋は，指紋と同様に個人を識別可能な生体情報である [2]．そして，GDPR の推進と拡大し続けるプライバシー問題への不安から，音声データを共有することは非常に困難な問題に直面している [12]．声紋は様々な認証システムに用いられているので [2]，スプーフィングなどの攻撃を受けてしまうこともある [18]．

この問題を解決するために, いくつかの研究により発話者の匿名化が試みられている [4], [7], [8], [15]. 声紋をデータセットから完全に除去して, 機械音声に置き換えて音声データを構成するという手法 [7] は, 有用性の面で大きな問題がある. 音声変換によってより高い有用性を達成した研究 [8], [15] においても, 変換時のパラメータを隠蔽することでプライバシーを保護する手法が用いられており, 隠蔽したパラメータを知る攻撃者は元の音声データに逆変換することが可能になってしまう. k-匿名性を拡張する形で匿名性を保証した研究 [4] もあるが, k-匿名性は攻撃者の背景知識によっては必ずしも安全ではないということが知られている [9], [11].

そのような背景のもと, Han ら (2020) は, 差分プライバシーを拡張することで声紋データのプライバシーの厳密なプライバシー指標であるボイス識別不能性を新しく定義し (定義 2), ボイス識別不能性に基づくデータを合成するメカニズムを提案した.

4.2 GAN を用いた合成データセット生成の先行研究

GAN を用いた合成データセット生成の最も先駆的な研究は Xie ら (2018) によって行われた. GAN 共通の問題として, ディープニューラルネットワークの複雑性が高いために, 学習された生成分布の密度が学習データ点に集中してしまい, 学習サンプルを容易に記憶してしまうというものがある. これにより, カルテなどのセンシティブな個人情報を含むデータを GAN で扱う際に, 情報漏洩や特定などのリスクが生じてしまう. Xie らはこの問題に着目し, GAN の学習過程で勾配にノイズを加えることで差分プライバシーを達成する DPGAN というモデルを提案した. なお, ディープニューラルネットワークにおいて差分プライバシーを達成する研究は Abadi ら (2016) [1] によってすでに行われていたが, Xie らのモデルは勾配へのノイズに制限を加え, より効率的に差分プライバシーを満たせるようになっているという点で先進的だった. 彼らは, 提案モデルが (ϵ, δ)-差分プライバシーを満たした上で訓練データを保護可能であるということを証明した. また, 実験により, 一定のノイズと制限の中で質の良いデータが生成されることを示した.

Jordon ら (2019) は, Private Aggregation of Teacher Ensembles (PATE) フレームワークを用いることで, Xie らのモデルよりも質の良いデータ生成を行うことに成功した. PATE とは, Papernot ら (2018) [13] で提案され, 後に Papernot ら (2019) [14] で改善されたモデルである. PATE は差分プライバシーを保ちながらクラス分類をするためのモデルであり, 複数のモデルの出力それぞれにラプラスノイズを加えるモデルのアンサンブルで表現される. PATE モデルを D に使用する際には, これが微分不可能であるという問題があるが, Jordon らは PATE の出力を学習するモデルをさらに追加することでこれを解決した.

Xu ら (2019) は, 条件付き GAN(CTGAN) という手法を提案し, ベイジアンネットワークを用いた手法と GAN を用いた手法のどちらのベースラインについても, それを上回る性能を発揮することを示した.

5 提案手法の方針

この章では, 本研究で用いようと考えている手法について, その方針を述べる.

5.1 全 体 像

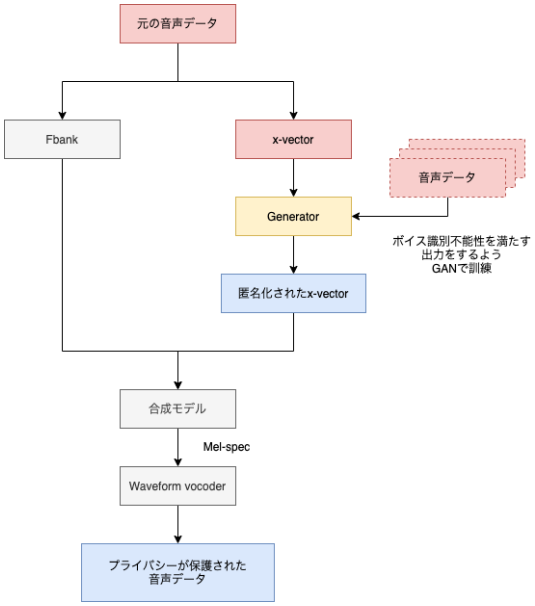

図 2 匿名化された音声データ生成の全体像

図 2 は, 本論文で提案する手法の全体像を示したものである. 与えられた音声データは, filter-bank (Fbank) という特徴と x-vector という特徴に分解される. filter-bank とは, 音声識別のためによく用いられる特徴である. x-vector は, 声紋のディープニューラルネットをベースにした埋め込み表現である [16]. x-vector を用いることで, 可変長の音声データを固定次元に埋め込むことが可能になる. 分解の後は, GAN によって訓練された Generator を用いて匿名化された x-vector を生成する. ただし, 元の x-vector と生成される x-vector はボイス識別不能性を満たすように Generator を訓練してあるものとする. この訓練の方針については, 次節で詳しく述べる. その後, Fbank と匿名化された x-vector で音声合成の標準的な入力として用いられる Mel-spectgram (Mel-spec) を生成する.

5.2 匿名化された声紋を生成する GAN

もっとも単純な GAN はデータセットの生成分布をモデル化するため, 声紋から声紋への変換には適していない. 声紋から声紋への変換を行うためには, conditional GAN (cGAN) [10] を用いる. cGAN は, 条件付きのデータの分布を扱えるように拡張された GAN のことである. すなわち, 次の変換を学習することを目指す. $G : \{x, z\} \to x'$ なお, z はランダムなノイズである.

しかし, 単純な cGAN ではまだプライバシーを保証をするに至らない. Yifan ら (2018) [17] はプライバシー保護と有用性のトレードオフのバランスを保ちながら顔画像のデータセット合

成を行えるモデルを cGAN を拡張することで実現した．Yifan らは cGAN にプライバシー保護を保証する verificator, 顔画像の構造情報を保持して有用性を保つための regulator という 2 つのモジュールを新たに追加し, cGAN, verificator, regulator の 3 つのモデルの損失関数の線形結合を全体のモデルの価値関数とした．本研究で提案する手法の一つとして, Yifan らの手法を参考に, cGAN にボイス識別不能性を保証する verificator を追加して 2 つのモデルの損失関数の線形結合を全体のモデルの価値関数とする手法が考えられる．verificator にどのようにしてボイス識別不能性を保証させるかについては今後の論点である．

6 まとめ

音声ベースのヒューマンコンピューターインタラクションが盛んになり, 音声データ利用が活発になる中で, プライバシー漏洩に関しての懸念が高まっている．本研究では, 声紋に関して匿名化を施した偽の音声データを合成することで特定などを防止する手法を, GAN を用いて実現する手法の方針を示した．また, そのプライバシー保護の尺度には, Han らの提案したボイス識別不能性を用いることを述べた．今後は手法についてさらに考察を深め, 実験を行うことが方針である．

文　献

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2016.

[2] A. Boles and P. Rad. Voice biometrics: Deep learning-based voiceprint authentication system. In *2017 12th System of Systems Engineering Conference (SoSE)*, pp. 1–6, 2017.

[3] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, pp. 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[4] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-Francois Bonastre. Speaker anonymization using x-vector and neural waveform models, 2019.

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[6] Y. Han, S. Li, Y. Cao, Q. Ma, and M. Yoshikawa. Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2020.

[7] T. Justin, V. Štruc, S. Dobrišek, B. Vesnicer, I. Ipšić, and F. Mihelič. Speaker de-identification using diphone recognition and speech synthesis. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 04, pp. 1–7, 2015.

[8] B. M. Lal Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent. Evaluating voice conversion-based privacy protection against informed attackers. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2802–2806, 2020.

[9] Martin M. Merener. Theoretical results on de-anonymization via linkage attacks. *Trans. Data Privacy*, Vol. 5, No. 2, p. 377–402, August 2012.

[10] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, Vol. abs/1411.1784, , 2014.

[11] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125, 2008.

[12] Andreas Nautsch, Catherine Jasserand, Els Kindt, Massimiliano Todisco, Isabel Trancoso, and Nicholas Evans. The GDPR Speech Data: Reflections of Legal and Technology Communities, First Steps Towards a Common Understanding. In *Proc. Interspeech 2019*, pp. 3695–3699, 2019.

[13] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data, 2017.

[14] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate, 2018.

[15] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, and Xiang-Yang Li. Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, SenSys '18, p. 82–94, New York, NY, USA, 2018. Association for Computing Machinery.

[16] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[17] Yifan Wu, Fan Yang, and Haibin Ling. Privacy-protective-gan for face de-identification, 2018.

[18] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, Vol. 66, pp. 130 – 153, 2015.

[19] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *CoRR*, Vol. abs/1802.06739, , 2018.

[20] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. *CoRR*, Vol. abs/1907.00503, , 2019.

[21] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2019.