ランダム直交行列を用いた秘密鍵による音声のプライバシー保護法*

☆丹羽祥子, 塩田さやか, 貴家仁志 (都立大)

1　はじめに

近年，クラウドサービス上で機械学習モデルを使用することが増えている．その背景として，計算機の初期投資やメンテナンスコストを削減できるという理由が挙げられる．しかし，クラウドサービスは外部のプロバイダによって管理されているため，悪意のあるインサイダーや外部からの攻撃によるデータ漏洩などの様々な脅威が懸念されている [1]．クラウドサービス上で機械学習モデルを使用するときには学習済みのモデルとクエリとなるデータをクラウドサービス上にアップロードする必要があるため，アップロードされた学習済みのモデルやクエリは，窃取されたり，悪用されたりするなどのリスクがある．このようなリスクを防ぐため，データをクラウドサービス上にアップロードする前に，それらのプライバシーを保護することが重要である．

音声データは年齢や性別，言語，発話内容などの個人情報を多く含んでいる．そのため，近年では音声のプライバシーを保護するための技術が注目され始めており，音声の匿名化や仮名化などの音声のプライバシー保護手法が提案されている [2,3]．しかし，音声の匿名化や仮名化手法は話者性の秘匿に焦点を当てた手法であり，発話内容の秘匿は目的としていない．そこで筆者らは話者性だけでなく発話内容の秘匿を目的とした研究として、秘密鍵を用いた暗号化手法を提案した [4]．しかし，この手法では秘密鍵の鍵空間が小さく，第三者に予測されてしまいやすいという問題点があった．そこで本論文では，秘密鍵の鍵空間を大きくするために，ランダム直交行列を用いて音声を暗号化する手法を提案する．ランダム直交行列は実数値の要素から構成されるため，秘密鍵の鍵空間を大きくすることができ，鍵の安全性が高められる．実験では話者照合タスクを用いて提案手法のプライバシー保護性能を評価した．実験結果から，正しい秘密鍵を用いて音声を暗号化した場合は，暗号化を適用する前と同じ精度でモデルを使用することが可能であり，正しくない秘密鍵を用いた場合はモデルの精度が大幅に下がることが確認された．さらに，提案手法はランダム直交行列を用いることで，暗号化のブロックサイズが小さい場合でも，大きな鍵空間を保持しながら音声の発話内容の秘匿を行うことができることが示された．

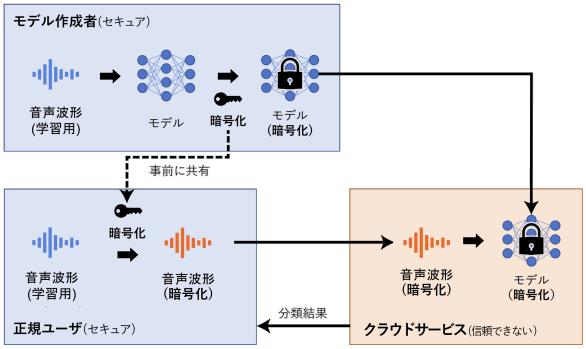


Fig. 1: 想定するプライバシー保護シナリオ

2　プライバシー保護シナリオ

本稿で想定しているプライバシー保護のシナリオを Fig. 1 に示す．まず，モデル作成者はセキュアな環境下で暗号化を施していない状態の音声データを入力として扱うモデルを作成し，作成した学習済みのモデルを秘密鍵を用いて暗号化する．次にモデル作成者は，暗号化したモデルをクラウドサービスなどの外部プロバイダにアップロードし，モデルの暗号化に用いた秘密鍵を正規ユーザに提供する．このとき，外部プロバイダは第三者によって管理されているため，セキュアな環境ではないと想定している．正規ユーザが外部プロバイダにアップロードされたモデルを使用するときは，クエリとなる音声データに対してモデル作成者から受け取った秘密鍵を用いて暗号化を施し，暗号化された音声データを外部プロバイダに送信する．そして，外部プロバイダでは暗号化されたモデルに暗号化されたクエリを入力し，その結果を正規ユーザに返す．このとき，外部プロバイダには暗号化されたモデルとクエリのみが送信されるため，正しい鍵を知っている正規ユーザのみがモデル作成者の意図した通りにモデルを利用することができる．また正しい鍵を知らない第三者が外部プロバイダからクエリを窃取しても，暗号化されたクエリから情報を得ることはできないという状況を想定している．

3　提案手法

本章では，提案手法におけるクエリの暗号化，モデルの暗号化，そしてランダム直交行列を用いた暗号化について説明する．

---

*Speech privacy-preserving method with secret key using random orthogonal matrix. by NIWA Shoko, SHIOTA Sayaka, KIYA Hitoshi (Tokyo Metropolitan University)

3.1　クエリの暗号化

クエリとなる音声データに対する秘密鍵を用いた暗号化手法の手順を示す．

1. 音声データ $\boldsymbol{X}$ をブロックサイズ M となるように分割する．分割された 1 ブロック $\boldsymbol{X}_{\mathrm{b}}$ は式 (1) のように表される．

$$\boldsymbol{X}_{\mathrm{b}} = \begin{bmatrix} x_{11} & \dots & x_{1T} \\ \vdots & \ddots & \vdots \\ x_{F1} & \dots & x_{FT} \end{bmatrix} \tag{1}$$

ただし，T，F はそれぞれ 1 ブロックの時間方向および周波数方向の総要素数を表し，$\boldsymbol{X}$ が音声波形などの 1 次元データの場合は $T = M$，$F = 1$，スペクトログラムなどの 2 次元データの場合は $T = M$，$F = M$ である．

2. 暗号化に用いる秘密鍵 $\boldsymbol{K}_{\mathrm{r}}$ を生成する．ここで，秘密鍵 $\boldsymbol{K}_{\mathrm{r}}$ は式 (2) のように表される．

$$\boldsymbol{K}_{\mathrm{r}} = \begin{bmatrix} k_{11} & \dots & k_{1N} \\ \vdots & \ddots & \vdots \\ k_{N1} & \dots & k_{NN} \end{bmatrix} \tag{2}$$

ただし，N は 1 ブロック $\boldsymbol{X}_{\mathrm{b}}$ の総要素数を表しており，$N = T \times F$ となる．

3. 各ブロック $\boldsymbol{X}_{\mathrm{b}}$ に対して，共通する秘密鍵 $\boldsymbol{K}_{\mathrm{r}}$ を用いた暗号化を施すために，$\boldsymbol{X}_{\mathrm{b}}$ を 1 次元のベクトルに平坦化して $\hat{\boldsymbol{X}}_{\mathrm{b}}$ を得る．式 (3) のように $\hat{\boldsymbol{X}}_{\mathrm{b}}$ と秘密鍵 $\boldsymbol{K}_{\mathrm{r}}$ との積をとり，暗号化した $\hat{\boldsymbol{X}}_{\mathrm{b}}^{(\boldsymbol{K}_{\mathrm{r}})}$ を得る．

$$\begin{aligned} \hat{\boldsymbol{X}}_{\mathrm{b}}^{(\boldsymbol{K}_{\mathrm{r}})} &= \hat{\boldsymbol{X}}_{\mathrm{b}} \boldsymbol{K}_{\mathrm{r}} \\ &= \begin{bmatrix} x_1 & \dots & x_N \end{bmatrix} \begin{bmatrix} k_{11} & \dots & k_{1N} \\ \vdots & \ddots & \vdots \\ k_{N1} & \dots & k_{NN} \end{bmatrix} \end{aligned} \tag{3}$$

4. 1 次元ベクトルである $\hat{\boldsymbol{X}}_{\mathrm{b}}^{(\boldsymbol{K}_{\mathrm{r}})}$ を暗号化前のブロック $\boldsymbol{X}_{\mathrm{b}}$ と等しい形の行列となるように戻し，暗号化されたブロック $\boldsymbol{X}_{\mathrm{b}}^{(\boldsymbol{K}_{\mathrm{r}})}$ を得る．

手順 1 から 4 を音声データ $\boldsymbol{X}$ の全ブロックに対して行い，暗号化された音声データを $\boldsymbol{X}^{(\boldsymbol{K}_{\mathrm{r}})}$ とする．

3.2　モデルの暗号化

3.1 節の手順で暗号化された音声データ $\boldsymbol{X}^{(\boldsymbol{K}_{\mathrm{r}})}$ を復号することなく直接モデルに入力するために，モデルの一部に変換を施す必要がある．本手法ではモデルの第一層目が畳み込み層であり，第一層目の畳み込み層のカーネルサイズとストライドサイズが等しいモデルを想定する．カーネルサイズとストライドサイズが等しいことによって，畳み込み処理を暗号化されたブロック毎に行うことができるためである．Figure 2 に示すように，暗号化を施す第一層目の畳み込み層のカーネルを $\boldsymbol{E}$ とおく．音声データ $\boldsymbol{X}$ が 1 次元データの場合，$\boldsymbol{E}$ は $1 \times P$ の行列，2 次元データの場合，$\boldsymbol{E}$ は $P \times P$ の行列である．このとき P は $\boldsymbol{E}$ のカーネルサイズを表し，本論文では $P = M$ である．カーネル $\boldsymbol{E}$ を用いて音声データ $\boldsymbol{X}$ に対して畳み込み処理を行うとき，任意のブロック $\boldsymbol{X}_{\mathrm{b}}$ に対して行う計算は式 (4) のように表せる．

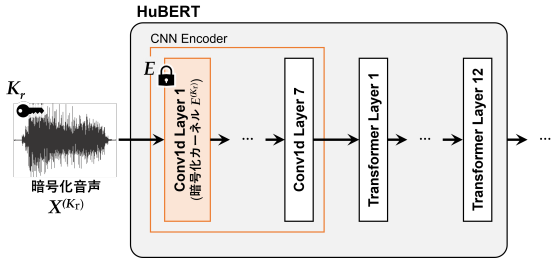


Fig. 2: 自己教師付き学習モデルの 1 つである HuBERT [5] に提案手法を適用した例

$$z = \boldsymbol{X}_{\mathrm{b}} \cdot \boldsymbol{E} \tag{4}$$

カーネル $\boldsymbol{E}$ に暗号化を施す手順は以下の通りである．まず，カーネル $\boldsymbol{E}$ を 1 次元のベクトルに平坦化し，$\hat{\boldsymbol{E}}$ を得る．式 (5) のように転置させた秘密鍵 $\boldsymbol{K}_{\mathrm{r}}$ と $\hat{\boldsymbol{E}}$ との積をとり，暗号化した $\hat{\boldsymbol{E}}^{(\boldsymbol{K}_{\mathrm{r}})}$ を得る．

$$\hat{\boldsymbol{E}}^{(\boldsymbol{K}_{\mathrm{r}})} = \boldsymbol{K}_{\mathrm{r}}^{t} \hat{\boldsymbol{E}}^{t} \tag{5}$$

1 次元ベクトルである $\hat{\boldsymbol{E}}^{(\boldsymbol{K})}$ を暗号化前のカーネル $\boldsymbol{E}$ と等しい形の行列となるように戻し，暗号化されたカーネル $\boldsymbol{E}^{(\boldsymbol{K}_{\mathrm{r}})}$ を得る．秘密鍵 $\boldsymbol{K}_{\mathrm{r}}$ で暗号化された音声データ $\boldsymbol{X}^{(\boldsymbol{K}_{\mathrm{r}})}$ とカーネル $\boldsymbol{E}^{(\boldsymbol{K}_{\mathrm{r}})}$ で畳み込みを行う場合，各暗号化ブロック $\boldsymbol{X}_{\mathrm{b}}^{(\boldsymbol{K}_{\mathrm{r}})}$ に対して式 (6) のような計算が行われる．

$$\begin{aligned} z^{(\boldsymbol{K}_{\mathrm{r}})} &= \boldsymbol{X}_{\mathrm{b}}^{(\boldsymbol{K}_{\mathrm{r}})} \cdot \boldsymbol{E}^{(\boldsymbol{K}_{\mathrm{r}})} = \hat{\boldsymbol{X}}_{\mathrm{b}}^{(\boldsymbol{K}_{\mathrm{r}})} \hat{\boldsymbol{E}}^{(\boldsymbol{K}_{\mathrm{r}})} \\ &= \hat{\boldsymbol{X}}_{\mathrm{b}} \boldsymbol{K}_{\mathrm{r}} \boldsymbol{K}_{\mathrm{r}}^{t} \hat{\boldsymbol{E}}^{t} = \boldsymbol{X}_{\mathrm{b}} \cdot \boldsymbol{E} = z \end{aligned} \tag{6}$$

したがって，提案手法を適用した前と後で全く同じ計算結果を得ることができるため，暗号化した音声を復号することなくモデルに入力することが可能となり，プライバシー保護が可能となる．

3.3　ランダム直交行列

先行研究 [4] で提案された暗号化手法である Shuffling や Flipping では，暗号化手順は 3.1, 3.2 と等しいものであったが，使用している秘密鍵の鍵空間が非常に小さいという問題があった．例えば，Shuffling による暗号化で使用される秘密鍵 $\boldsymbol{K}_{\mathrm{s}}$ は各ブロック

Table 1: VoxCeleb1 を用いて話者照合モデルを構築した際の EER (%) とモデルおよびクエリに対しての暗号化の有無による性能評価

モデル 暗号化	クエリ		
	暗号化なし	正しい鍵	誤った鍵 (5 回平均)
なし	7.91	-	-
あり	35.3	7.91	35.1

に含まれる要素のインデックスを任意の順番に入れ替えるため，入力データが 1 次元，$M = 3$ のとき，$\boldsymbol{K}_{\mathrm{s}} = [3, 1, 2]$ などになる．つまり，秘密鍵 $\boldsymbol{K}_{\mathrm{s}}$ は音声データが 1 次元の場合には $M!$ 通り，2 次元の場合は $(M \times M)!$ 通りの鍵しか使用できない．また，Flipping で使用される秘密鍵 $\boldsymbol{K}_{\mathrm{f}}$ は 0 または 1 からなるビット列であるため，入力データが 1 次元，$M = 3$ のとき，$\boldsymbol{K}_{\mathrm{f}} = [0, 0, 1]$ などになる．つまり，秘密鍵 $\boldsymbol{K}_{\mathrm{f}}$ は音声データが 1 次元の場合には 2^M 通り，2 次元データの場合には $2^{M \times M}$ 通りの鍵しか使用できない．特に音声データの無音区間に着目すると，鍵空間が小さい場合，第三者によって秘密鍵が推定されやすくなるという危険性がある．秘密鍵の鍵空間を大きくし，予測を難しくするために，本研究ではランダム直交行列を鍵空間に用いることを提案する．ランダム直交行列では，入力データが 1 次元，$M = 3$ のとき，式 (7) のような秘密鍵が作成できる．

$$\boldsymbol{K}_{\mathrm{r}} = \begin{bmatrix} 0.9898 & -0.0661 & -0.1264 \\ 0.1309 & 0.7732 & 0.6205 \\ 0.0568 & -0.6307 & 0.7740 \end{bmatrix} \tag{7}$$

このように，負値を含むランダムに生成される実数値の要素を用いるため，秘密鍵の鍵空間が大きくなり，秘密鍵の予測が困難になる．

4 実験

4.1 実験条件

本実験では提案手法のプライバシー保護性能を評価するため，話者照合タスクでの実験を行った．実験に使用したモデルは自己教師付き学習モデルの 1 つである HuBERT モデル [5] で，LibriSpeech コーパス [6] を用いて学習を行った．モデルの構造は HuBERT BASE と同じであるが，第一層目の畳み込み層のカーネルサイズ P とストライドサイズが等しくなるように，ストライドサイズを 10 とした．入力は 1 次元データの音声波形であり，暗号化に用いるブロックサイズ M はカーネルサイズ P やストライドサイズと等しくなるように 10 に設定した．HuBERT から出力される音声表現を元に x-vector [7] に基づく話者照合モデルを構築した．学習には VoxCeleb1 コーパス [8] を用いた．評価指標には等価エラー率（Equal error rate; EER）を用いる．

Table 2: ランダム直交行列と Shuffling [4] を用いて暗号化した LibriSpeech [6] に対する音声認識モデルの WER(%) の比較

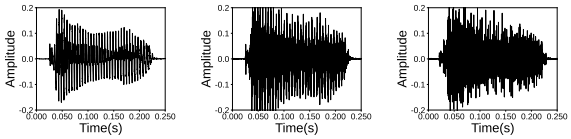
M	ランダム 直交行列	Shuffling
暗号化なし	2.7	
5	28.9	13.9
10	57.6	30.6
20	85.4	63.9
128	94.8	94.9

提案手法によって暗号化された音声に含まれる発話内容がどのくらい秘匿されているかを評価するための実験を行った．ブロックサイズ $M = 5, 10, 20, 128$ の条件下で，提案手法または先行研究 [4] の 1 つである Shuffling を用いて暗号化した音声波形を文献 [9] で提案された学習済みの音声認識モデルに入力して評価した．評価データには LibriSpeech [6] の test_clean サブセットを用い，評価指標は単語誤り率 (Word error rate; WER) を用いる．

4.2 実験結果

話者照合実験において，ブロックサイズ $M = 10$ とした場合のモデル及びトライアルデータであるクエリの暗号化の有無による性能結果を Tab. 1 に示す．「正しい鍵」はクエリが暗号化されたモデルと同じ秘密鍵を用いて暗号化されている場合，「誤った鍵」はクエリが暗号化されたモデルとは異なる秘密鍵を用いて暗号化されている場合である．また「誤った鍵」の EER は，正しい秘密鍵とは異なるランダムに生成した 5 つの秘密鍵を用いてクエリを暗号化したときの EER の平均を取ったものである．正しい鍵を用いてクエリを暗号化したとき，提案手法を適用する前と全く同じ EER が得られている．一方，誤った鍵を用いてクエリを暗号化したとき，提案手法を適用する前よりも大きく EER が上がっていることが分かる．また，「暗号化なし」も誤った鍵のうちの 1 つとみなすことができるため，大きく EER が下がっている．これらの結果から，提案手法を適用することで，正しい鍵を知っている正規ユーザのみが暗号化されたモデルを正しく使用することができることが確認できた．

提案手法がどのくらい音声に含まれる発話内容を秘匿できるのかを評価するために，提案手法または先行研究の 1 つである Shuffling [4] を用いて暗号化した音声波形を音声認識モデルに入力した結果を Tab. 2 に示す．暗号化していない音声を入力したときの WER



(a) オリジナル音声　(b) 暗号化音声 ($M = 10$)　(c) 暗号化音声 ($M = 128$)

Fig. 3: 暗号化した音声波形

と比較すると，提案手法または Shuffling を用いて暗号化した音声を入力したときの WER が高くなっていることが分かる．さらに，Shuffling と比較すると M が小さい場合は，ランダム直交行列で暗号化したときの方が WER が高く，よりプライバシーの保護ができていることが分かる．また，ランダム直交行列を用いた暗号化では，M が小さい場合でも鍵空間は十分に大きくなるため，秘密鍵の予測は困難である．例えば $M = 5$ の場合，先行研究 [4] で用いられる秘密鍵 $\boldsymbol{K}_{\mathrm{s}}$ は整数値からなる 1×5 の行列であり $5! = 120$ 通りの鍵しか生成できないが，本手法の秘密鍵 $\boldsymbol{K}_{\mathrm{r}}$ は実数値からなる 5×5 の直交行列であるため，圧倒的に多くのパターンの秘密鍵を生成することができる．したがって，提案手法はブロックサイズが小さい場合でも，大きな鍵空間を保ちながら音声波形に含まれる発話内容を秘匿できることが確認できた．実際に提案手法によって暗号化した音声を Fig. 3 に示す．Figure 3(a) は暗号化がされていないオリジナル音声の波形，Fig. 3(b) は $M = 10$ でオリジナル音声を暗号化した例，Fig. 3(c) は $M = 128$ でオリジナル音声を暗号化した例である．これらの図からもブロックサイズが小さくても音声が大きく変化していることが確認できる．

提案法ではブロックサイズ M とモデルのカーネルサイズ P，ストライドサイズが等しくなるよう設計する必要がある．そのため，P の大きさが暗号化なしの状態で精度にどの程度影響が出るかを確認するために，$P = 10, 20, 64, 128$ の条件下で HuBERT モデルを学習したときの EER の変化を Tab. 3 に示す．この結果から，P が大きくなるほど EER が高くなることが分かる．したがって，音声を用いた分類タスクでは，カーネルサイズが小さい方が性能が高くなる傾向があるため，ブロックサイズが小さくても保護性能が高いことは，暗号化手法を評価する点において非常に重要である．

5　まとめ

本稿では，ランダム直交行列を用いた秘密鍵による音声のプライバシー保護手法を提案した．提案手法では秘密鍵としてランダム直交行列を用いることで，先行研究と比較して同じブロックサイズであっても秘密鍵の鍵空間が大きくなることを示した．実験より，提案手法を適用することで，正規ユーザのみが正しくモデルを使用することができ，暗号化された音声から発話内容を認識することが難しいことも確認された．音声分類モデルはモデルのカーネルサイズが小さいほど高い性能が出やすい傾向がある．したがって，先行研究と比較すると，提案手法はカーネルサイズやブロックサイズを大きくすることなく，音声のプライバシー保護性能の向上や鍵空間の拡大を達成することができたため，より安全性が向上したといえる．今後の課題として，提案手法を一般的な機械学習モデルでも使用できるような枠組みの検討が挙げられる．

Table 3: カーネルサイズ P の変化と話者照合モデルの EER (%) の変化

カーネルサイズ P	10	20	64	128
EER	7.9	10.3	16.3	21.8

謝辞　本研究は，JSPS 科研費 21H01327 と ROIS-DS-JOINT(047RP2023)，セコム財団挑戦的研究助成の助成を受けたものである．

参考文献

[1] Hamed Tabrizchi, et al. A survey on security challenges in cloud computing: issues, threats, and solutions. *The journal of supercomputing*, Vol. 76, No. 12, pp. 9493–9532, 2020.

[2] Natalia Tomashenko, et al. The voiceprivacy 2022 challenge evaluation plan. [Online]. Available: `https://www.voiceprivacychallenge.org/vp2020/docs/VoicePrivacy_2020_Eval_Plan_v1_4.pdf`, 2020.

[3] Hiroto Kai, et al. Robustness of Signal Processing-Based Pseudonymization Method Against Decryption Attack. in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, pp. 287–293, 2022.

[4] Niwa Shoko, et al. A privacy-preserving method using secret key for convolutional neural network-based speech classification. in *European Signal Processing Conference (EUSIPCO)*, 2023, accpected.

[5] Wei-Ning Hsu, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 29, pp. 3451–3460, 2021.

[6] Vassil Panayotov, et al. Librispeech: An asr corpus based on public domain audio books. in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, 2015.

[7] David Snyder, et al. X-vectors: Robust dnn embeddings for speaker recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[8] Arsha Nagrani, et al. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, Vol. 60, p. 101027, 2020.

[9] Shigeki Karita, et al. A comparative study on transformer vs rnn in speech applications. in *2019 IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 449–456, 2019.