N ,,y>VãlÕNº0 –û[P`ÅX1• Oá[fO
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

Oá[fb€X1
IEICE Technical Report
SIP2024-53(2024-08)

音声プライバシー保護のための複数のランダム直交行列を用いた秘密鍵による攻撃耐性の向上

田中　康平[†]　貴家　仁志[†]　塩田さやか[†]

[†] 東京都立大学　〒191–0065 東京都日野市旭ヶ丘 6–6

あらまし　本論文では，ランダム直交行列に基づく秘密鍵を用いた音声のプライバシー保護手法の頑健性を強化する手法を提案する．近年，音声対話システムなどの普及からクラウドサーバー上の深層学習モデルを利用するために，クラウドサーバー上に音声データがアップロードされる機会が増えている．アップロードされた音声データがそのままサーバー上に保存される場合，サーバー内外にいる悪意のある攻撃者などによって音声データが漏洩してしまう懸念がある．そのため，音声データをクラウドサーバーへ送信する前にプライバシー保護を施すことでサーバー上に生の音声データを直接送信せずに済む枠組みは重要である．このような背景から，音声データを保護する手法としてランダム直交行列に基づく秘密鍵を用いた音声のプライバシー保護手法が提案された．この手法では，暗号化された深層学習モデルに対して暗号化した音声は復号せず入力でき，かつモデルを正しく利用することができる．しかし，このプライバシー保護法では，秘密鍵に用いるランダム直交行列のサイズが十分な大きさでない場合に発話内容や話者性の秘匿性能が低く，ランダムに生成された秘密鍵を用いる攻撃に対しても頑健性が低いという問題があった．そこで，本研究では複数のランダム直交行列に基づく秘密鍵を用いることで，秘密鍵に用いる行列のサイズが小さい場合でのプライバシー保護性能の低下を防ぐ手法を提案する．複数のランダム直交行列を用いることで一つの行列のサイズが小さくても擬似的に鍵空間が大きくなり，プライバシー保護性能を向上することが可能となる. 実験では，最先端の話者照合モデルに対してランダム直交行列のサイズが小さい場合でも複数のランダム直交行列を用いることでプライバシー保護性能を向上させることができることを確認した．

キーワード　音声プライバシー保護，音声暗号化，話者照合

Enhancing robustness using multiple random orthogonal matrices for voice privacy-preserving

Kohei TANAKA[†], Hitoshi KIYA[†], and Sayaka SHIOTA[†]

[†] Tokyo Metropolitan University　, 6-6 Asahigaoka, Hino-shi, Tokyo, 191–0065 Japan.

1. まえがき

近年，深層学習モデルをクラウドサーバー上に配置し, モデルを実行する機会が増えている．モバイル端末など, ローカルの計算資源が限られている場合にも高性能なモデルを利用することができるなどの利点があり，クラウドサーバー上での深層学習モデル用いた画像認識や音声認識なども一般的に利用されている．一方で，クラウドサーバー上での深層学習モデルの実行にはユーザーがクエリとなるデータをアップロードする必要があり，クラウドサーバーの内部や外部からの攻撃によって，アップロードされたデータが漏洩してしまうことに対する懸念も示されている [1],[2].

音声データには，その発話内容だけでなく言語，年齢，性別など個人のプライバシーに関する情報や，個人を特定しうる話者性などの情報が含まれている [3],[4] そのため，クラウドサーバー上の深層学習モデルを利用するためにアップロードした音声データが漏洩することで個人情報や機密情報が漏洩してしまう恐れがある．こうしたリスクを防ぐためにクラウドサーバー上にアップロードする音声データにプライバシー保護を施し，クラウドサーバー上でプライバシーを保護したまま深層学習モデルを実行することが可能となる枠組みが必要となってくる．このような背景から，音声のプライバシー保護技術が注目され始めており，音声の匿名化や仮名化など様々な技術が提案されている [5]〜[9].

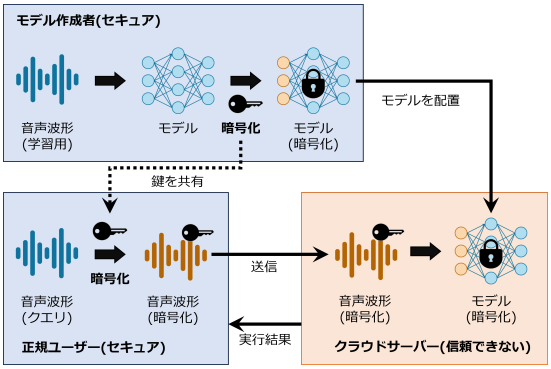

図 1　プライバシー保護シナリオ

音声のプライバシー保護手法の一つとしてランダム直交行列に基づく秘密鍵を用いた手法が提案されている [10]．この従来手法では，入力層に畳み込み層を持つ深層学習モデルの畳み込み層のカーネルおよび音声データに対して，ランダム直交行列に基づく秘密鍵による暗号化を施すことで，音声データのプライバシーを保護したまま正しく深層学習モデルを利用することを可能としている．また，従来法では暗号化の際の秘密鍵の行列サイズを大きくすることでプライバシー保護性能が向上することも確認されている．一方で，秘密鍵として用いるランダム直交行列のサイズは畳み込み層のカーネルサイズと一致している必要があり，小さいカーネルサイズを使用しているモデルでは，高いプライバシー保護性能を得ることが難しいことが報告されている．

そこで本論文では，文献 [10] の従来手法に対して使用するモデルのカーネルサイズに依存しにくく，かつ高いプライバシー保護性能と攻撃耐性が得られるために，複数のランダム直交行列に基づく秘密鍵を用いる手法を提案する．複数のランダム直交行列を秘密鍵に用いることで，ランダム直交行列のサイズが小さい場合でも，秘密鍵の鍵空間を擬似的に大きくできるため，プライバシー保護性能を向上させることができる．実験では，最新の話者照合モデルに対して本手法を適用し，小さいカーネルサイズを使用したモデルであっても従来法と比較して高いプライバシー保護性能と攻撃耐性を得られることが確認できた．

2.　プライバシー保護シナリオ

本論文で想定しているプライバシー保護のシナリオを図 1 に示す．まず，深層学習モデルの作成者は，セキュアな環境下でモデルの学習を行う．この際，モデルの学習に使用する音声データは暗号化を施されていないものとする．次に，モデルの作成者は学習済みのモデルの暗号化を行い，クラウドサーバー上に配置する．その後，モデルの作成者はモデルの暗号化に使用した秘密鍵を正規ユーザーに提供する．モデルの利用者は受け取った秘密鍵を用いてクエリとなる音声データを暗号化してクラウドサーバー上へアップロードして暗号化されたモデルを利用し，実行結果を受け取る．この際，クラウドサーバー上では暗号化された音声データを復号することなく暗号化されたモデルに入力することを想定している．そのため，セキュアではないと仮定しているクラウドサーバー上には暗号化された音声データおよび暗号化されたモデルのみが保存されていることになる．これにより，秘密鍵を知らない第三者がクラウドサーバー上の暗号化された音声を窃取しても，暗号化された音声データからは音声データに含まれている本来の情報を取得することはできず，クラウドサーバー上に保存される音声データの保護が可能となる．この枠組みは画像のプライバシー保護においても研究がなされている [11], [12].

3.　ランダム直交行列に基づく秘密鍵による暗号化法

本章では従来法となる文献 [10] で提案されたランダム直交行列に基づく秘密鍵を用いた暗号化によるプライバシー保護手法の概要について説明する．本論文では，音声波形のような 1 次元で表現される音声データを対象とするため，本章でも音声データが 1 次元で表現されている場合について記載する．

3.1　音声データの暗号化

音声データに対する秘密鍵を用いた暗号化手法について説明する．

手順 1　1 次元の音声データ $\boldsymbol{X} = [x_1, x_2, \ldots, x_T]$ を式 (1) のようにブロックサイズ M に分割する．

$$\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i, \ldots, \boldsymbol{X}_t] \tag{1}$$

ただし，T はデータ長であり，$t = \lfloor T/M \rfloor$ である．

手順 2　暗号化に用いる秘密鍵 $\boldsymbol{K}$ を生成する．ここで，秘密鍵はサイズ $M \times M$ のランダム直交行列である．$\boldsymbol{K}$ は式 (2) のように表される．

$$\boldsymbol{K} = \begin{bmatrix} k_{11} & \cdots & k_{1M} \\ \vdots & \ddots & \vdots \\ k_{M1} & \cdots & k_{MM} \end{bmatrix} \tag{2}$$

手順 3　手順 1 で用意したブロックに分割された音声データ $\boldsymbol{X}_i$ に対し，手順 2 で生成した秘密鍵 $\boldsymbol{K}$ を用いて式 (3) のように行列積をとり，暗号化されたブロック $\boldsymbol{X}_i^{(\boldsymbol{K})}$ を得る．

$$\boldsymbol{X}_i^{(\boldsymbol{K})} = \boldsymbol{X}_i \boldsymbol{K} \tag{3}$$

音声データ $\boldsymbol{X}$ のブロック $\boldsymbol{X}_i$ すべてに対して式 (3) の計算を行うことで式 (4) のような暗号化された音声データ $\boldsymbol{X}^{(\boldsymbol{K})}$ を得られる．

$$\boldsymbol{X}^{(\boldsymbol{K})} = \left[\boldsymbol{X}_1^{(\boldsymbol{K})}, \ldots, \boldsymbol{X}_i^{(\boldsymbol{K})}, \ldots, \boldsymbol{X}_t^{(\boldsymbol{K})}\right] \tag{4}$$

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 $\boldsymbol{X}^{(\boldsymbol{K})}$ を復号することなく直接モデルに入力するために，モデルの一部に処理を施す必要がある．従来法では，深層学習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ブロックサイズ等しいことを前提としている．この条件を満たすモデルに対して，ブロックサイズ M をカーネルサイズと一致するように設定して音声を暗号化することで，暗号化された音声を復号することなく直接モデルに入力するこ

とができる．これは，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，畳み込み層で行われる内積演算が暗号化されたブロック毎に行われるためである．暗号化を施す第一層目の畳み込み層のカーネル $\boldsymbol{E}$ を式 (5) のように表す．

$$\boldsymbol{E} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_M \end{bmatrix} \tag{5}$$

暗号化されていない音声データ $\boldsymbol{X}$ を畳み込み層に入力するとき，カーネルサイズとストライドサイズがブロックサイズ等しいため，各ブロック $\boldsymbol{X}_i$ とカーネル $\boldsymbol{E}$ を用いて式 (6) のような内積演算が行われる．

$$z = \boldsymbol{X}_i \boldsymbol{E} \tag{6}$$

ここで，入力データとして暗号化された音声データ $\boldsymbol{X}^{(\boldsymbol{K})}$ を暗号化されたモデルへ入力することを考える．モデルの暗号化は，暗号化された音声データ $\boldsymbol{X}^{(\boldsymbol{K})}$ の暗号化の効果を打ち消すために，式 (7) のように音声データの暗号化に用いた秘密鍵の転置と畳み込み層のカーネルの行列積をとる．

$$\boldsymbol{E}^{(\boldsymbol{K})} = \boldsymbol{K}^\top \boldsymbol{E} \tag{7}$$

秘密鍵 $\boldsymbol{K}$ で暗号化された音声データ $\boldsymbol{X}^{(\boldsymbol{K})}$ を暗号化されたモデルに入力すると，第一層目の畳み込み層では式 (8) のような計算が行われる．

$$z^{(\boldsymbol{K})} = \boldsymbol{X}_i^{(\boldsymbol{K})} \boldsymbol{E}^{(\boldsymbol{K})} = \boldsymbol{X}_i \boldsymbol{K} \boldsymbol{K}^\top \boldsymbol{E} = \boldsymbol{X}_i \boldsymbol{E} = z \tag{8}$$

このように，$\boldsymbol{K}$ と $\boldsymbol{K}^\top$ の積が単位行列となり，秘密鍵 $\boldsymbol{K}$ で暗号化された音声データを同じ秘密鍵 $\boldsymbol{K}$ で暗号化されたモデルに対して入力することで暗号化を施した場合にも，暗号化を施さない場合と全く同じ計算結果を得ることができる．したがって，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，音声のプライバシーを保護したままモデルを使用することが可能となる．

3.3 従来法の問題点

本章で説明した従来法では，正しい秘密鍵を用いた場合には暗号化を行わない場合と全く同じ性能が得られ，正しい秘密鍵を用いない場合には，性能が大幅に低下することが報告されていた．しかし，ブロックサイズが小さい場合には，ランダムに生成した正しくない秘密鍵を用いた場合にも比較的高い性能が得られてしまう問題があった．音声タスクで用いられる深層学習モデルにおいては，性能向上のために一般的に小さいカーネルサイズが用いられることが多い [13],[14]．そのため，ブロックサイズを大きくせずに安定して高いプライバシー保護性能を得ることが課題となっていた．

4. 提 案 手 法

本論文では，3.3 節で説明した従来法の問題点を解決するために，秘密鍵として複数のランダム直交行列を用いる方法を提案する．以降では，複数のランダム直交行列に基づく秘密鍵を用いた，音声データの暗号化および，暗号化した音声データを復号せずにモデルへ入力するための処理について説明する．3.1 節と同様に音声データが音声波形のような 1 次元のデータとして与えられる場合について述べる．

4.1 音声データの暗号化

手順 1　3.1 節の式 (1) と同様に一次元の音声データ $\boldsymbol{X}$ をブロックサイズ M のブロックに分割し，分割された各ブロックを $\boldsymbol{X}_i$ とする．

手順 2　暗号化に用いる秘密鍵を生成する．提案手法では秘密鍵として複数のランダム直交行列を用いる．秘密鍵 $\boldsymbol{K}_{\text{mult}}$ を式 (9) のように表す．

$$\boldsymbol{K}_{\text{mult}} = \{\boldsymbol{K}_1, \ldots, \boldsymbol{K}_n, \ldots, \boldsymbol{K}_N\} \tag{9}$$

ここで，N は秘密鍵に用いるランダム直交行列の個数であり，$\boldsymbol{K}_n \in \boldsymbol{K}_{\text{mult}}$ は $M \times M$ のランダム直交行列である．

手順 3　手順 1 で用意したブロックに分割された音声データ $\boldsymbol{X}_i$ に対し，手順 2 で生成した秘密鍵 $\boldsymbol{K}_{\text{mult}}$ の要素 $\boldsymbol{K}_n$ を用いて，式 (10) のように行列積をとり，暗号化した $\boldsymbol{X}_i^{(\boldsymbol{K}_n)}$ を得る．

$$\boldsymbol{X}_i^{(\boldsymbol{K}_n)} = \boldsymbol{X}_i \boldsymbol{K}_n \quad \text{ただし} \quad n = i \bmod N \tag{10}$$

音声データ $\boldsymbol{X}$ のブロック $\boldsymbol{X}_i$ すべてに対して式 (10) の計算を行い，式 (11) のような暗号化された音声データ $\boldsymbol{X}^{(\boldsymbol{K}_{\text{mult}})}$ を得る．

$$\boldsymbol{X}^{(\boldsymbol{K}_{\text{mult}})} = \left[\boldsymbol{X}_1^{(\boldsymbol{K}_1)}, \ldots, \boldsymbol{X}_N^{(\boldsymbol{K}_N)}, \boldsymbol{X}_{N+1}^{(\boldsymbol{K}_1)}, \ldots, \boldsymbol{X}_t^{(\boldsymbol{K}_{t \bmod N})} \right] \tag{11}$$

式 (11) のように $\boldsymbol{K}_{\text{mult}}$ は N 個のランダム直交行列からなるため，N 個の音声ブロック $\boldsymbol{X}_i$ に対して $\boldsymbol{K}_{\text{mult}}$ の要素 $\boldsymbol{K}_1$ から $\boldsymbol{K}_N$ までを順番に変えながら暗号化を施す．音声ブロック $\boldsymbol{X}_i$ が N 個を超えると秘密鍵は再び $\boldsymbol{K}_1$ に戻り，再び $\boldsymbol{K}_1$ から $\boldsymbol{K}_N$ までを順番に用いて暗号化を行う．このように暗号化を行うことで，複数のランダム直交行列を秘密鍵として用いることができる．ランダム直交行列の個数 N を増やすことで鍵空間を擬似的に大きくすることが可能となり，ブロックサイズが小さい場合でも従来法と比較してプライバシー保護性能を向上させることができる．

4.2 モデルの暗号化

4.1 節の手順で暗号化された音声データ $\boldsymbol{X}^{(\boldsymbol{K}_{\text{mult}})}$ を復号することなくモデルへ入力するために，使用するモデルに施す処理について説明する．従来法と同様に本手法も，深層学習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ブロックサイズ等しいことを前提としている．以降では，モデルの第一層目が一次元畳み込み層である場合の手順について述べる．

使用するモデルの第一層目の畳み込み層を，複数ランダム直交行列からなる秘密鍵で暗号化された音声 $\boldsymbol{X}^{(\boldsymbol{K}_{\text{mult}})}$ に対応させるためには，N 個の畳み込みカーネルが必要になる．そのため，モデルの第一層目の畳み込み層を N 個の畳み込み層へ置き換える．暗号化を施す前の畳み込み層のカーネル $\boldsymbol{E}$ を 3.2 節

と同様に式 (5) で表す．秘密鍵 $\boldsymbol{K}_{\text{mult}}$ を用いて暗号化された N 個のカーネルを用意する．$\boldsymbol{K}_n \in \boldsymbol{K}_{\text{mult}}$ で暗号化されたカーネル $\boldsymbol{E}^{(\boldsymbol{K}_n)}$ は式 (12) のように表される．

$$\boldsymbol{E}^{(\boldsymbol{K}_n)} = \boldsymbol{K}_n^\top \boldsymbol{E} \tag{12}$$

ここで，入力データとして $\boldsymbol{K}_{\text{mult}}$ を用いて暗号化された音声データ $\boldsymbol{X}^{(\boldsymbol{K}_{\text{mult}})}$ を暗号化されたモデルへ入力することを考えると，モデルの第一層目の畳み込み層では式 (13) のような計算が行われる．

$$z^{(\boldsymbol{K}_n)} = \boldsymbol{X}_i^{(\boldsymbol{K}_n)} \boldsymbol{E}^{(\boldsymbol{K}_n)} = \boldsymbol{X}_i \boldsymbol{K}_n \boldsymbol{K}_n^\top \boldsymbol{E} = \boldsymbol{X}_i \boldsymbol{E} = z \tag{13}$$

音声データのブロックおよびカーネルがともに $\boldsymbol{K}_n \in \boldsymbol{K}_{\text{mult}}$ によって暗号化されているため，従来法と同様に暗号化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．これにより，ブロックサイズを大きくせずにプライバシー保護性能を向上させることが可能になり，カーネルサイズが小さい深層学習モデルに暗号化を適用した場合にも安定して高いプライバシー保護性能を確保することができる．

5. 実　　験

5.1 実験条件

提案手法のプライバシー保護性能およびランダムに鍵を生成した際の攻撃耐性についての評価を行うため，話者照合での実験を行った．話者照合とは，入力された音声が登録されている話者本人の音声であるか否かを判定する二値分類タスクである．本実験では先端手法の一つである話者埋め込みに基づく話者照合モデルで，かつ入力が 1 次元の音声波形である RawNeXt [13] を用いた．実験の際には，VoxCeleb2 コーパス [15] を用いて学習され，文献 [13] で配布されている事前学習済みのモデルに対して暗号化を施した．話者照合の評価には VoxCeleb1 コーパス [16] の評価セットを使用し，評価指標には等価エラー率（Equal error rate; EER）を用いた．EER とは本人棄却率と他人受入率が等しくなる値であり，この値が小さいほどモデルの精度が高いことを示す指標である．

2 章で説明したシナリオに基づいた提案手法のプライバシー保護性能と，ランダムに生成された秘密鍵を用いた攻撃に対する頑健性を検証するために，話者照合モデルを秘密鍵で暗号化し，そのモデルに対して 200 通りの正しくない秘密鍵（モデルの暗号化に用いた秘密鍵とは異なる秘密鍵）を用いて暗号化した音声を入力した．暗号化に用いるブロックサイズは $M = 3$ とし，モデルのカーネルサイズと等しくなるように設定した．これは従来法の実験においてプライバシー保護性能が十分に得られなかったときの条件となる．秘密鍵に使用するランダム直交行列の個数 N は，$N = 3, 9, 27, 81, 243$ の場合を試した．なお，従来法では単一のランダム直交行列を秘密鍵として用いるため，提案手法の $N = 1$ の場合と同じ条件となる．本実験では，音声の暗号化に正しい鍵を用いた場合には EER が暗号化なしの場合から変化せず，正しくない鍵を用いたときには EER が大きくなることを話者性が秘匿されているとみなしており，これをプライバシー保護性能が向上したかどうかの指標としている．

表 1　暗号化された RawNext に対して 200 通りの正しくない鍵で暗号化した音声を入力した際の平均 EER (%) (暗号化なしの場合の EER は 1.95%)

ランダム直交行列の個数 N	クエリ	
	正しい鍵	正しくない鍵
1 (従来法)	1.95	16.50
3		30.26
9		37.69
27		38.93
81		38.86
243		38.76

5.2 実験結果

表 1 に，RawNeXt に対してモデルの暗号化を施し，200 通りの正しくない秘密鍵を用いて暗号化した音声を入力した際の平均の EER を示す．正しい鍵の場合の EER は音声およびモデルの暗号化を行わない場合の EER と同一の値である．これは従来法の性質として文献 [10] でも説明されており，本実験においても同じ結果であった．一方で，正しくない鍵の場合の EER は，秘密鍵に用いるランダム直交行列の個数 N がいずれの場合にも，正しい鍵の場合の EER と比較して大幅に高くなっており，正しい鍵を知っているユーザーのみが正しくシステムを利用できる，つまり音声のプライバシー保護が行えていることが確認できる．従来法の $N = 1$ の場合と比較して，提案手法では $N = 3$ から 243 まで N を増加させるにつれて EER がより高くなった．これより，複数のランダム直交行列を秘密鍵に用いる提案手法では，ブロックサイズを大きくすることなく擬似的に鍵空間を広げることが可能であり，その結果プライバシー保護性能が向上することがわかる．

次に，ランダムに生成された 200 通りの正しくない秘密鍵を用いて暗号化した音声を入力した際の EER の分布を図 2 に示す．従来法の $N = 1$ の場合には，EER の分布の幅が広く，ランダムに生成された秘密鍵であっても，暗号化された音声の話者性が観測されてしまう場合がある．これは従来法の問題点として述べられており，特に本実験のようなブロックサイズ M が小さい場合にはこの問題が顕著に現れている．一方，複数のランダム直交行列を用いた提案手法の場合には，N が増加するほど，EER の分布の幅が狭くなっており，特に十分大きい N の場合には，ランダムに生成された秘密鍵で暗号化された音声で話者照合システムを突破することが，従来法と比較して大幅に難しくなることが示された．

6. ま と め

本論文では，ランダム直交行列に基づく秘密鍵を用いた音声プライバシー保護手法に対して，プライバシー保護性能を向上させるために複数のランダム直交行列を用いる手法を提案した．提案手法は，従来法で高いプライバシー保護性能を得られなかった，畳み込み層のカーネルサイズが小さいモデルにおいても，秘密鍵に用いるランダム直交行列の個数を増やすことで

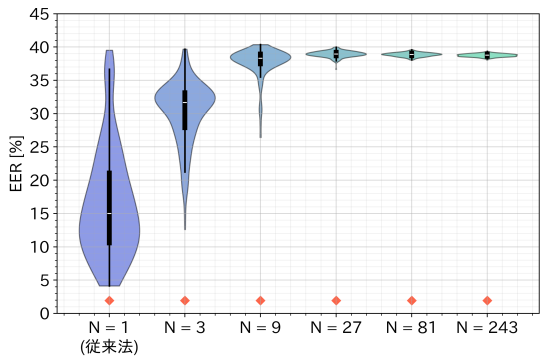

図 2　暗号化された RawNext に対してランダムに生成された 200 通りの正しくない秘密鍵を用いて暗号化した音声を入力した際の EER の分布 (ひし形: 暗号化なしの場合の EER)

秘密鍵の鍵空間を擬似的に広げることが可能となり，高いプライバシー保護性能を得られた．また，実験では話者照合において話者性がどの程度秘匿されるかを指標として実験を行った．実験結果より，提案手法では秘密鍵に用いるランダム直交行列の個数を増やすことで攻撃に対する頑健性が向上することも確認できた．今後の課題として，発話内容の秘匿性能や他のモデルについても検証することが挙げられる．

7. 謝　　辞

本研究の一部は，ROIS-DS-JOINT(022RP2024) と JSPS 科研費 JP24K14993 の助成を受けたものである．

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] Ashish Singh, et al. Cloud security issues and challenges: A survey. *Journal of Network and Computer Applications*, Vol. 79, pp. 88–115, 2017.

[3] Jacob Leon Kröger, et al. Privacy implications of voice and speech analysis–information disclosure by inference. *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers 14*, pp. 242–258, 2020.

[4] Gundeep Singh, et al. Spoken language identification using deep learning. *Computational Intelligence and Neuroscience*, Vol. 2021, No. 1, p. 5123671, 2021.

[5] 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.

[6] Hiroto Kai, et al. Lightweight and irreversible speech pseudonymization based on data-driven optimization of cascaded voice modification modules. *Computer Speech & Language*, Vol. 72, p. 101315, 2022.

[7] 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.

[8] Shi-Xiong Zhang, et al. Encrypted speech recognition using deep polynomial networks. *In Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5691–5695, 2019.

[9] Francisco Teixeira, et al. Towards End-to-End Private Automatic Speaker Recognition. In *Proc. Interspeech 2022*, pp. 2798–2802, 2022.

[10] Niwa Shoko, et al. Speech privacy-preserving methods using secret key for convolutional neural network models and their robustness evaluation. *APSIPA Transactions on Signal and Information Processing 2024 (Accepted)*, 2024.

[11] Hitoshi Kiya, et al. An overview of compressible and learnable image transformation with secret key and its applications. *APSIPA Transactions on Signal and Information Processing*, Vol. 11, No. 1, 2022.

[12] AprilPyone Maungmaung, et al. Privacy-preserving image classification using an isotropic network. *IEEE Transactions on MultiMedia*, Vol. 29, No. 2, pp. 23–33, 2022.

[13] Ju-Ho Kim, et al. Rawnext: Speaker verification system for variable-duration utterances with deep layer aggregation and extended dynamic scaling policies. In *Proc. 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7647–7651, 2022.

[14] 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.

[15] J Chung, et al. Voxceleb2: Deep speaker recognition. In *Proc. Interspeech 2018*, 2018.

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