

# 複数のランダム直交行列に基づく秘密鍵による 音声プライバシー保護法の適用要件緩和と攻撃耐性評価

田中 康平<sup>1</sup> 貴家 仁志<sup>1</sup> 塩田 さやか<sup>1</sup>

**概要:** 本研究では、先行研究である複数のランダム直交行列に基づく秘密鍵による音声プライバシー保護法における適用可能なモデルに関する制約を緩和する手法を提案する。近年、深層学習を用いた音声処理システムをモバイル端末から利用することは一般的になっており、それに伴ってクラウド上で実行される深層学習モデルに音声を送信する機会も増加している。一方で、クラウド上に送信される音声のプライバシーに対する懸念も高まっている。先行研究では、クラウド上に配置される深層学習モデルとアップロードされる音声に対して、複数のランダム直交行列に基づく秘密鍵による暗号化を施し、音声に含まれる発話内容、話者性などを秘匿した状態でクラウド上のモデルによる推論を実現できることが報告されていた。ただし、先行研究の手法には暗号化を適用可能な深層学習モデルに関する制約があり、実際に適用可能なモデルは限られていた。そこで、本研究ではこの制約を緩和し、提案法により一般的に利用される音声処理モデルに対しても暗号化手法を適用可能であることを示す。さらに、音声匿名化技術のコンペティションである voice privacy challenge で用いられた攻撃モデルを拡張し、2つの攻撃シナリオに基づいて提案法に対する攻撃耐性の評価を行った。実験の結果、先行研究で検証されていなかった、より強力な攻撃を想定したシナリオにおいても、提案法では話者性の秘匿に関してはプライバシー保護性能が維持されることが確認できた。

## 1. はじめに

近年、深層学習を用いた音声処理システムが広く普及し、スマートフォンやスマートスピーカーなどから手軽に利用できるようになった。こうした音声処理システムは、モバイル端末における計算資源の制約などから、クラウド上の深層学習モデルを利用して実現されていることが多い。クラウド上の深層学習モデルを利用するシステムでは、音声をクラウドサーバーへ送信する必要があり、音声のプライバシーに関する懸念も示されている。音声には、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 [1] や、個人を特定しうる話者性などの情報が含まれている。そのため、国際的なデータ保護法である General Data Protection Regulation (GDPR) において音声はプライバシー規制の対象となっている [2]。こうした背景から音声プライバシー保護に関する研究は盛んに行われており、音声匿名化技術の国際的なコンペティションである voice privacy challenge (VPC) [3-5] が定期的に開催されている。主流な音声プライバシー保護技術は声質変換やテキスト音声合成に基づく手法で、VPC へ提出されたシステムの多くもこれらの手法に基づいている。VPC で

は、原音声の話者性を秘匿することを目的としており、発話内容の秘匿は対象としていない。

音声に含まれる話者性に加えて、発話内容も対象としたプライバシー保護手法の先行研究として、ランダム直交行列に基づく秘密鍵による暗号化に基づく手法が提案されている [6,7]。この手法では、音声に含まれる話者性および発話内容を秘匿した状態で深層学習モデルによる推論を行えることが報告されている。しかし、この音声暗号化手法は適用可能なモデルの制約の影響で、モデルの再学習なしに最先端の音声処理モデルに適用することはできなかった。また、先行研究では暗号化音声から発話内容および話者性を推測する攻撃において、攻撃者が暗号化システムを利用しない場合のみを想定しており、より強力な攻撃者を想定した実際的な条件での攻撃耐性の評価がされていなかった。

そこで、本研究ではランダム直交行列に基づく秘密鍵による暗号化に基づく手法を拡張し、より多くの深層学習モデルに対して再学習なしに暗号化手法を適用可能にする手法を提案するとともに、VPC で用いられた攻撃モデルを発話内容の秘匿へ拡張し、2つの攻撃シナリオによって、暗号化手法に対するプライバシー保護性能の評価実験を行った。実験の結果、提案法による暗号化手法を一般的な SSL モデルである wav2vec 2.0 [8] をフロントエンドとする ASR モ

<sup>1</sup> 東京都立大学大学院 システムデザイン研究科

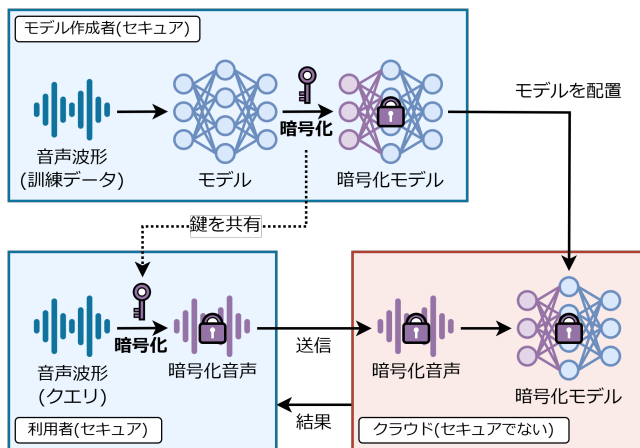


図 1 プライバシー保護シナリオ

デルおよび ASV モデルへ適用可能であることが確認できた。さらに、先行研究では検証されていなかった、より強力な VPC に基づく攻撃シナリオにおいても、特に話者性の秘匿に関して攻撃耐性を維持できることがわかった。

## 2. プライバシー保護シナリオ

本章では、前提となるプライバシー保護のシナリオについて説明する。図 1 はプライバシー保護シナリオの概要図であり、モデル作成者、利用者、クラウドの 3 つの領域に分かれている。モデルの利用者が送信する音声のプライバシーを保護するために、暗号化された音声を復号することなくモデルの計算を行い、音声処理タスクの推論結果を得る流れが示されている。

この一連の流れを実現するために、モデルの作成者は、暗号化されていない音声を用いてモデルを学習し、秘密鍵を用いてモデルの暗号化を行いクラウド上に配置した後、モデルの暗号化に使用した秘密鍵をモデルの利用者に共有する。モデルの利用者は共有された秘密鍵を用いてクエリ音声を暗号化しクラウド上に送信する。クラウド上のモデルの計算は暗号化された音声を復号することなく行えるため、第三者がクラウドサーバー上から原音声を窃取することはできず、秘密鍵を知らなければ暗号化された音声から原音声に含まれる情報を得ることはできない。このシナリオにおいて、モデル利用者のプライバシー保護のためには、暗号化された音声から原音声に関する情報を推測することがより困難であることが求められる。

## 3. 複数の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、従来法である文献 [7] で提案された複数のランダム直交行列を秘密鍵として用いる音声プライバシー保護手法の概要を説明する。本章で説明する手法は、2 章で説明したプライバシー保護シナリオに基づいており、プライバシー保護手法は音声の暗号化とモデルの暗号化で構成

される。

### 3.1 音声の暗号化

音声の暗号化は、1 次元の音声波形  $X$  をサイズ  $M$  のブロックに分割し、それぞれのブロック  $X_i$  に  $M$  次元のランダム直交行列をかけることで行われる。従来法では、秘密鍵  $K_{\text{mult}}$  に  $N$  個のランダム直交行列が含まれており、ブロック  $X_i$  にかけるランダム直交行列  $K_n$  は  $K_{\text{mult}}$  の中から一つ選ばれる。文献 [7] の手法では、ブロック  $X_i$  のインデックス  $i$  を用いて、 $n = i \bmod N$  のようにブロックにかけるランダム直交行列  $K_n$  を選択する。このようにして暗号化された音声は、人が聞いた際に発話内容や発話者を識別することは困難となる。

### 3.2 モデルの暗号化

モデルの暗号化とは共通の秘密鍵で暗号化された音声を入力した場合に、暗号化された音声から正しい推論結果を計算できるよう事前にモデルに施す処理のことである。文献 [7] の手法が対象としているモデルは、1 次元の音声波形を直接入力するモデルであり、その第 1 層目はカーネルサイズとストライドが等しい畳み込み層である必要がある。また、秘密鍵のランダム直交行列の次元  $M$  はカーネルサイズと同じである。モデルの暗号化はモデルの第 1 層目のカーネルに、音声ブロック  $X_i$  にかけられた  $K_n$  を転置した行列  $K_n^T$  をかけることで行われる。これにより、モデルに入力された音声ブロック  $X_i$  にかけられた  $K_n$  は、モデルの第 1 層目で打ち消され非暗号化時と同じ出力が得られる。従来法では、 $N$  個の暗号化されたカーネルが用意されるため、第 1 層目は  $N$  個の暗号化された畳み込み層に分岐し、分岐したそれぞれの第 1 層目は対応する  $K_n$  で暗号化されたブロックを入力とする。暗号化された音声をブロックごとに異なる畳み込み層へ入力する必要がありモデルの構造は変更されるが、それぞれのカーネルは学習済みのモデルのものを用いて計算できるため再学習は不要である。また、異なる秘密鍵を利用する場合も暗号化されていないモデルに新たな秘密鍵を適用することで暗号化を行えるため、モデルの再学習は不要である。

### 3.3 従来法の問題点

本章で説明した従来法では、モデル第 1 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必要がある。これは、音声にかけた暗号化がモデルの第 1 層目の畳み込み層で打ち消されるために、畳み込み層で暗号化された音声のブロックと暗号化されたカーネルとの内積計算を行う必要があるためである。しかしながら、この条件により多くの最先端の音声処理モデルを再学習せずに暗号化を適用することが困難であった。また、既存のモデルをカーネルサイズとストライドサイズが等しくなるよう修正し、

再学習を行うことはモデルの性能低下を招く可能性もある。このため、モデルの再学習やモデルの性能を低下させるような変更を必要とせずに暗号化を適用できる手法を検討する必要がある。

#### 4. 提案法

本研究では、3.3節で述べた暗号化が適用可能なモデルの条件を緩和する手法を提案する。3.3節で述べたように、従来法によるモデルの暗号化では第1層目の畳み込み層はカーネルサイズとストライドが等しい必要があった。しかし、最先端のモデルにおいて、この条件を満たしているモデルは少ない。そこで、モデルの暗号化において第1層目の畳み込み層のストライドがカーネルサイズと一致しない場合にも正しい計算が行われるように音声ブロックの分割方法を変更することを提案する。

カーネルサイズ  $M$ 、ストライド  $S$ 、出力チャンネル数  $1$  でバイアス項を持たない簡略化した畳み込み層を考える。畳み込み層のカーネル  $E$  を式 (1) のように表す。

$$E = [e_0, \dots, e_k, \dots, e_{M-1}]^T \quad (1)$$

ここで、 $e_k$  はカーネルの要素を表す。畳み込み層の入力  $Y$  を式 (2)、出力を  $Z$  を式 (3) のように表す。

$$Y = [y_0, y_1, \dots, y_{T-1}] \quad (2)$$

$$Z = [z_0, \dots, z_i, \dots, z_{L-1}] \quad (3)$$

ただし、 $T$  は畳み込み層の入力の長さ、 $L$  は畳み込み層の出力の長さで  $L = \lfloor \frac{T-M}{S} + 1 \rfloor$  である。カーネルサイズとストライドが  $S < M$  の条件において、畳み込み層が音声ブロックとカーネルとの内積で計算できるとすると、畳み込み層の出力の各要素  $z_i$  は式 (4) のように表される。

$$z_i = \sum_{k=0}^{M-1} e_k y_{Si+k} = A_i E \quad (4)$$

ここで、 $A_i = [y_{Si}, y_{Si+1}, \dots, y_{Si+(M-1)}]$  である。 $A_i$  は畳み込み層の出力の要素  $y_i$  を計算する際に参照される入力  $Y$  の要素からなるベクトルである。この、 $A_i$  を  $A_0$  から  $A_{L-1}$  まで入力音声から切り出し、連結した配列が  $A$  である。図 2 には  $M=3, S=2$  の場合の  $A$  が示されている。ここで、 $A$  をモデルへの入力とするために畳み込み層を修正する事を考える。畳み込み層をカーネルは変更せずにストライドを  $S=M$  に変更し、 $A$  を入力した場合元の畳み込み層の出力と一致する出力が得られる。ただし、修正された畳み込み層へ入力される信号の長さ  $L_A$  は式 (5) となり、図 2 のように元の音声よりも長くなる。

$$L_A = \left\lfloor \frac{T-M}{S} + 1 \right\rfloor M \quad (5)$$

これを利用して  $S < M$  の畳み込み層を同一の計算を行う

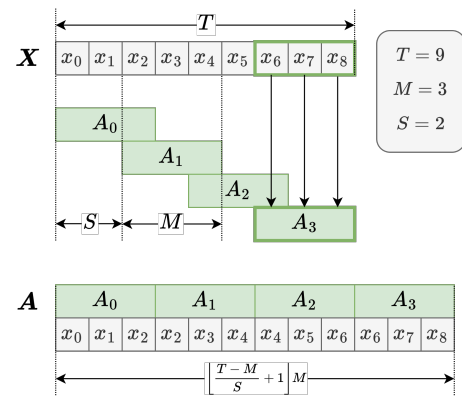


図 2 ブロックの切り出し

$S = M$  の畳み込み層へ修正することが出来るため、従来法をカーネルサイズとストライドが異なる畳み込み層を持つモデルへ適用することが可能になる。音声の暗号化は  $A$  に対して従来法と同様の方法で行える。また、モデルの暗号化は第1層目の畳み込み層を  $S = M$  に変更したうえで、3.2節の手順により行う。

#### 5. 攻撃耐性の評価

暗号化手法の攻撃耐性を評価するために用いる攻撃シナリオについて説明する。本研究では VPC2020 [3] および VPC2024 [5] で用いられた攻撃モデルに基づく 2 つの攻撃シナリオを用いて暗号化手法を評価する。

##### 5.1 攻撃モデル

暗号化された音声に対する攻撃とは、暗号化された音声から原音声に含まれる情報を推測する行為である。本研究では、暗号化された音声から原音声の発話内容と話者性を推測することを目的とした攻撃者を想定する。そのために、話者性の秘匿のみを目的としていた VPC の攻撃シナリオを、攻撃者が暗号化された音声から発話内容も推測するシナリオに拡張する。攻撃者による暗号化された音声からの話者性の推測には、VPC と同様に ASV モデルが用いられる。また、暗号化された音声からの発話内容の推測には ASR モデルが用いられる。本研究では、以下の 2 つの情報にアクセス可能な攻撃者を想定した攻撃シナリオで提案法の攻撃耐性を評価する。

- (1) 攻撃者はクラウド上のモデル利用者が音声に暗号化を施して送信した暗号化された音声（クエリ音声）にアクセスできる
- (2) 攻撃者はクエリ音声の暗号化に利用された暗号化システムのアルゴリズムにアクセスできる

本研究で用いる 2 つの攻撃シナリオにおいて、攻撃者は上記のアクセス可能な情報と ASR モデルおよび ASV モデルを用いて、暗号化された音声から発話内容と話者性を表現する話者埋込みベクトルを得る。ただし、攻撃者はユーザーが使用した秘密鍵を知ることはできない。攻撃者は、

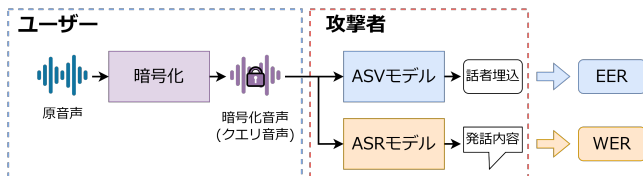


図 3 攻撃シナリオ 1

アクセス可能な情報の範囲内において、暗号化された音声に対して前処理を行い攻撃に使用するモデルの精度を改善したり、暗号化システムを利用して攻撃に使用するモデルを暗号化された音声へ適応させたりすることができる。

## 5.2 攻撃シナリオ 1

攻撃シナリオ 1 は攻撃者が 5.1 節の (1) の情報のみ、つまり、暗号化されたクエリ音声の情報のみを利用して攻撃を行うことを想定している。図 3 は攻撃シナリオ 1 の概要図である。ユーザーは自身の発話を暗号化してクエリ音声を作成する。クエリ音声を入手した攻撃者は、事前学習済みの ASR モデルおよび ASV モデルを用いてクエリ音声から発話内容および話者埋込みベクトルを得る。攻撃者がクエリ音声に ASR モデルを適用して得られた書き起こしと、原音声の書き起こしとの単語エラー率 (Word error rate; WER) は攻撃者がクエリ音声からどの程度発話内容を推測できたかを示す評価指標となる。さらに、攻撃者がクエリ音声に ASV モデルを適用して得られた話者埋込みベクトルと、原音声と同一話者の暗号化されていない登録発話の話者埋込みベクトルの類似度に基づいて等価エラー率 (Equal error rate; EER) を算出する。この EER は、攻撃者がクエリ音声からどの程度に原音声の話者を推測できるかを示す評価指標となる。暗号化手法の攻撃耐性は WER および EER によって評価され、WER および EER が高いほど暗号化手法の攻撃耐性は高いと考えられる。

## 5.3 攻撃シナリオ 2

攻撃シナリオ 2 は攻撃者が 5.1 節の (1) と (2) の情報、つまり、クエリ音声の情報に加えてクエリ音声の暗号化に利用したシステムのアルゴリズムも利用して攻撃を行うことを想定している。図 4 は攻撃シナリオ 2 の概要図である。ユーザーが自身の発話を暗号化してクエリ音声を作成する点は攻撃シナリオ 1 と同様である。ただし、攻撃シナリオ 2 では、攻撃者がクエリ音声の暗号化に用いられた暗号化システムを用いて暗号化されたデータセットを作成し、暗号化されたデータセットでファインチューニングされた ASR モデルおよび ASV モデルを攻撃に使用する。攻撃シナリオ 1 と同様に WER および EER を算出し、暗号化手法の攻撃耐性を評価する。ただし、EER の算出にはクエリ発話と暗号化された登録発話から得られた話者埋込みベクトルを用いる。暗号化された登録発話を用いる理由は、

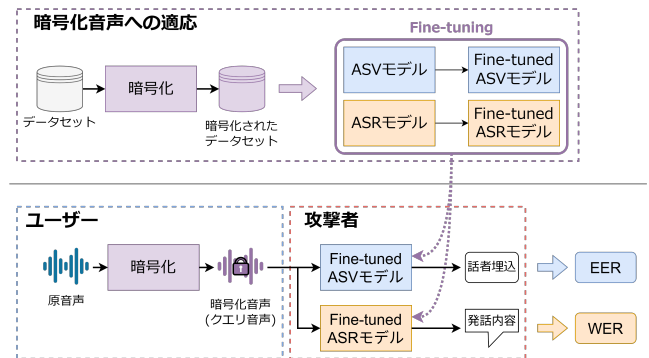


図 4 攻撃シナリオ 2

ASV モデルがファインチューニングによって暗号化された音声に適応しているためである。攻撃シナリオ 2 は攻撃者が暗号化された音声に適応したモデルを利用してクエリ音声から発話内容と話者性を推測する事ができるため、攻撃シナリオ 1 と比較してより強力な攻撃である。

## 6. 実験

従来法による暗号化を適用できなかった、第 1 層目の畳み込み層においてカーネルサイズとストライドが異なるモデルにおいて、提案法による暗号化を適用できることを示すために実験を行った。また、暗号化手法に対して 5 章で説明した攻撃シナリオに基づいて攻撃耐性を評価する実験を行った。

### 6.1 事前学習モデル

攻撃耐性の評価実験に用いる ASR モデルと ASV モデルのフロントエンドにはそれぞれ、wav2vec 2.0 Large と wav2vec 2.0 Base [8] を用いる。また、ASR モデルのバックエンドは 1024 のユニットを持つ 2 層の全結合層、ASV モデルのバックエンドは x-vector [9] を用いる。本研究で用いる x-vector モデルは 5 層の TDNN を持ち、文献 [9] と同様の構造であるが、wav2vec 2.0 Base モデルが出力する 768 次元の特徴フレームを入力とする変更がされている。wav2vec 2.0 の第 1 層目の畳み込み層のカーネルサイズは 10、ストライドは 5 であり先行研究では暗号化を適用できない条件である。wav2vec 2.0 Large モデルと wav2vec 2.0 Base モデルでは、Transformer ブロックの数と各 Transformer ブロックが持つアテンションヘッドの数が異なっているが、その前段の CNN 特徴抽出器は共通である。ASR モデルおよび ASV モデルの事前学習に用いるデータセットはそれぞれ LibriSpeech [10] と voxceleb1 [11] である。

攻撃シナリオ 2 の場合には、この事前学習モデルを暗号化されたデータセットでファインチューニングし、暗号化された音声に適応したモデルを作成して攻撃に用いる。暗号化されたデータセットは、事前学習モデルの学習に使用されたデータセットと同じものを発話ごとにランダムに生

表 1 提案法による暗号化を適用した事前学習モデルの性能  
(暗号化なしの場合, WER=1.76[%], EER=3.66[%])

N	正しい鍵		正しくない鍵	
	WER [%]	EER [%]	WER [%]	EER [%]
1	1.76	3.66	67.1	30.8
3	1.76	3.66	94.2	34.6
5	1.76	3.66	96.6	36.8

成した秘密鍵で暗号化して作成する。ただし, ASR モデルのファインチューニングでは, LibriSpeech データセットの train-clean-100 および train-clean-360 サブセットのみを使用して暗号化されたデータセットを作成した。ASR モデルおよび ASV モデルは, フロントエンドの学習率  $1e-4$ , バックエンドの学習率  $5e-3$  でファインチューニングを行った。

## 6.2 前処理

5.1 節で述べたとおり, 攻撃者は攻撃モデルで定義された範囲内で暗号化された音声に対して前処理を適用できる。本研究では, 攻撃者が用いる ASR モデルおよび ASV モデルの精度の改善と, 攻撃シナリオ 2 で行う暗号化されたデータセットでのファインチューニング時における学習の安定性を改善するために 2 つの前処理を行う。

1 つ目の前処理は, 暗号化された音声の時間スケール調整である。4 章で述べたとおり, 式 (5) で表される暗号化された音声の長さは図 2 のように隣接するブロック  $A_i$  と  $A_{i+1}$  が  $M-S$  だけ重複した音声サンプルを持つため引き伸ばされている。そこで, 暗号化された音声の隣接するブロックの重複を削除する。処理結果  $\tilde{A}$  は式 (4) のブロック  $A_i$  を用いて  $\tilde{A}_0 = A_0, \tilde{A}_i = (A_i[M-S], A_i[M-S+1], \dots, A_i[M-1])$  とすると, この  $\tilde{A}_i$  を  $\tilde{A}_0$  から  $\tilde{A}_L$  まで連結したものととなる。長さは,  $|\tilde{A}| = |\tilde{A}_0| + \sum_{i=1}^{L-1} |\tilde{A}_i| = M + (L-1)S = T$  となり, 暗号化された音声にこの処理を施すと原音声と長さが一致する。この前処理は, 攻撃シナリオ 1 および 2 で適用される。また, 攻撃シナリオ 2 でのファインチューニングの際にも行う。

2 つ目の前処理は, ローパスフィルタの適用である。これは, 攻撃シナリオ 1 でのみ適用され, 1 つ目の前処理を適用したあとに行う。暗号化された音声に攻撃者が使用するモデルへ入力する前に, 暗号化された音声に含まれる高周波成分をカットするために用いられる。カットオフ周波数 4[kHz] のローパスフィルタを使用する。

## 6.3 実験結果

表 1 に事前学習モデルおよび入力音声に暗号化を施した場合の ASR および ASV の性能を示す。N は秘密鍵に用いるランダム直交行列の数を表す。表 1 において正しい鍵, つまり音声およびモデルを共通の秘密鍵で暗号化した場合に WER および EER は, 事前学習モデルに暗号化を施さない場合と同等で, 性能の低下はなかった。これは, 3 章

表 2 攻撃シナリオ 1 の攻撃耐性評価

N	前処理なし		LPF	
	WER [%]	EER [%]	WER [%]	EER [%]
1	40.6	33.1	38.3	30.8
3	90.0	44.8	87.1	45.1
5	93.5	46.3	90.0	46.0
7	96.8	46.5	94.4	46.7
9	97.5	46.8	94.9	47.0

表 3 攻撃シナリオ 2 での攻撃耐性評価

N	WER[%]	EER[%]
1	3.66	11.4
3	7.53	22.3
5	9.95	22.8
7	11.4	26.9
9	11.2	27.0

で述べたとおり, 音声とモデルが共通の秘密鍵で暗号化されている場合には, 第 1 層目の畳み込み層の出力は暗号化を施さない場合と一致するからである。一方で正しくない鍵, つまり音声とモデルを暗号化した秘密鍵が異なる場合には, 正しい鍵を用いた場合と比較して大幅に WER および EER が増加している。このことから, 従来法では暗号化を適用できなかった条件のモデルにおいても, 2 章で説明したプライバシー保護シナリオに従ってクエリ音声およびモデルを暗号化することで, 目的の音声タスクの性能を損なわずに音声を暗号化した状態で推論が可能であることが示された。

表 2 および表 3 は, 攻撃シナリオ 1 および攻撃シナリオ 2 における攻撃耐性の評価結果を示す。N は秘密鍵に用いるランダム直交行列の数を表す。表 2 より, ランダム直交行列の数 N の増加に伴い, WER および EER がともに増加する傾向が確認できた。具体的には, N=1 のとき WER は 40.6%, EER は 33.1%であったが, N=9 では WER が 97.5%, EER が 46.8%に達した。これは, ランダム直交行列の数を増やすことで, 暗号化された音声から発話内容や話者性を推定することが困難になり, プライバシー保護性能が向上したことを示している。秘密鍵に用いるランダム直交行列の数が増加するほど ASV における EER が増加する傾向は, 本研究と異なる ASV モデルを用いた文献 [7] でも確認されており, 同様の傾向が示された。また, ローパスフィルタを適用した場合も同様の傾向が見られ, 前処理なしの場合と比較して若干 WER および EER が低下したが, 攻撃耐性をおおむね維持しているといえる。表 3 の攻撃シナリオ 2 の結果では, 攻撃者が暗号化システムにアクセス可能なより強力な攻撃条件下で評価を行った。その結果, N=1 では WER が 3.66%, EER が 11.4%であったのに対し, N=9 では WER が 11.2%, EER が 27.0%となり, こちらも N の増加による攻撃耐性の向上が確認できた。ただし, 攻撃シナリオ 1 と比較すると WER および EER は

低く、暗号化された音声に適応したモデルを用いた攻撃がより有効であることが示唆された。以上の結果から、秘密鍵に用いるランダム直交行列の数を増やすことで、どちらのシナリオにおいても攻撃耐性を向上させる効果があることが示された。先行研究では評価されていなかった、攻撃シナリオ2のようなより強力な攻撃に対しても、特に話者性の秘匿に関しては一定のプライバシー保護性能を維持していることが確認できた。

## 7. まとめ

本研究では、複数のランダム直交行列に基づく秘密鍵による音声プライバシー保護法において、従来法における適用可能なモデルに関する制約を緩和する手法を提案した。また、先行研究で攻撃耐性の評価が行われていなかった発話内容の秘匿に関する実験や、VPCで用いられた攻撃モデルに基づく、より強力な攻撃を想定したシナリオでの攻撃耐性の評価実験を行った。実験の結果、従来法では暗号化を適用できなかった条件のモデルにおいても、提案法により暗号化を適用可能であることが確認できた。より強力な攻撃を想定した攻撃シナリオ2の結果においても、話者性の秘匿に関してはプライバシー保護性能が一定程度維持される事が確認できた。今後の課題として、本研究で攻撃耐性の評価に用いたモデル以外にも、様々なアーキテクチャや学習データで作成されたモデルを用いて暗号化手法の攻撃耐性の評価を行う必要がある。また、本研究では発話内容と話者性の秘匿に焦点をあてて評価を行ったが、音声に含まれる年齢、性別などの他の情報についてもプライバシー保護性能を評価する必要がある。

## 参考文献

- [1] Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Müller, C., Huber, R., Andrassy, B., Bauer, J. G. et al.: Comparison of four approaches to age and gender recognition for telephone applications, *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 4, IEEE, pp. IV-1089 (2007).
- [2] Nautsch, A., Jasserand, C., Kindt, E., Todisco, M., Trancoso, I. and Evans, N.: The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding, *arXiv preprint arXiv:1907.03458* (2019).
- [3] Tomashenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.-F., Noé, P.-G. and Todisco, M.: The VoicePrivacy 2020 Challenge Evaluation Plan (2022).
- [4] Tomashenko, N., Wang, X., Miao, X., Nourtel, H., Champion, P., Todisco, M., Vincent, E., Evans, N., Yamagishi, J. and Bonastre, J.-F.: The VoicePrivacy 2022 Challenge Evaluation Plan (2022).
- [5] Tomashenko, N., Miao, X., Champion, P., Meyer, S., Wang, X., Vincent, E., Panariello, M., Evans, N., Yamagishi, J. and Todisco, M.: The VoicePrivacy 2024 Challenge Evaluation Plan (2024).
- [6] Shoko, N., Shiota, S., Kiya, H. 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*, Vol. 13, No. 1 (2024).
- [7] 田中康平, 貴家仁志, 塩田さやか: 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, *信学技報*, Vol. 124, No. 162, pp. 45-49 (2024).
- [8] Baevski, A., Zhou, Y., Mohamed, A. and Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations, *Advances in neural information processing systems*, Vol. 33, pp. 12449-12460 (2020).
- [9] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S.: X-vectors: Robust dnn embeddings for speaker recognition, *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, pp. 5329-5333 (2018).
- [10] Panayotov, V., Chen, G., Povey, D. and Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books, *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, pp. 5206-5210 (2015).
- [11] Nagrani, A., Chung, J. S. and Zisserman, A.: Voxceleb: a large-scale speaker identification dataset, *arXiv preprint arXiv:1706.08612* (2017).