THE IEICE TRANSACTIONS ON FUNDAMENTALS OF ELECTRONICS, COMMUNICATIONS AND COMPUTER SCIENCES (JAPANESE EDITION)

電子情報通信学会 論文誌 A 基礎・境界

VOL. J105-A NO. 12
DECEMBER 2022

基礎・境界ソサイエティ

一般社団法人 電子情報通信学会
THE ENGINEERING SCIENCES SOCIETY
THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS

スマートフォンのタップ音からの入力内容推測可能性に関する研究

大内　結雲[†]　野崎真之介[†]　佐々木　葵[†]　奥村　紗名[†]
吉平　瑞穂[†]　芹澤　歩弥[†]　大木　哲史[†]　西垣　正勝[†a)]

Study on Possibility of Estimating Smartphone Inputs from Tap Sounds

Yumo OUCHI[†], Shinnosuke NOZAKI[†], Aoi SASAKI[†], Sana OKUMURA[†], Mizuho YOSHIHIRA[†], Ayumi SERIZAWA[†], Tetsushi OHKI[†], and Masakatsu NISHIGAKI[†a)]

あらまし　今日我々はスマートフォンを用いて様々な情報を入力しているが，それらの情報は音として生活の中に漏れ出している．攻撃者は小型盗聴器やスマートスピーカのマイクを用いて生活音や環境音を不正に録音し，我々の個人情報や機密情報を盗取する可能性がある．本研究では，「正規ユーザがスマートフォンにキー入力をする際のタップ音を，攻撃者が外部のマイクで盗聴する」という攻撃シナリオを想定し，その脅威に関する検討を行う．実験では，検討対象のタップ手段としては指とタッチペンの 2 種類を，検討対象の録音機器としてはタブレット端末とスマートスピーカの 2 種類をそれぞれ想定し，タップ操作で生じた音響データを用いて識別精度を評価する．より現実的な攻撃のシチュエーションを想定し，収集した複数名のデータを用いることで，既知のユーザのタップ音で学習した識別器を用いて未知のユーザの入力内容を推定可能であるかを検証した．その結果，最も高い場合で 29.4% の識別率が得られた．また，スマートフォン上のキー配置を変更し，キーとタップ音の対応関係を変化させることによる識別精度を評価することで，この攻撃に対する防御策の検討に資する．

キーワード　サイドチャネル攻撃，情報漏えい対策，スマートフォン，タップ音

1. まえがき

今や我々の日常の全てがスマートフォンを介在して行われていると言っても過言ではない．スマートフォンを用いて個人情報やパスワード等のセンシティブな情報を入力する機会が激増し，これに伴い，スマートフォンに入力された秘密情報を盗取する攻撃も増加の一途を辿っている．そのような攻撃手法の一つとしてサイドチャネル攻撃が存在する [1]．サイドチャネル攻撃は，暗号モジュールが搭載されている機器を外部から観察し，得られる副次的な情報を元に暗号解析を行う攻撃である．サイドチャネル攻撃はログに残らないために，攻撃の証拠が残りにくいという特徴がある．サイドチャネル攻撃の一つにテンペスト攻撃が存在する [2]．テンペスト攻撃はディスプレイやケーブルから漏洩する微弱な電磁波を検知することで，ディスプレイに表示された情報や入力された文字列等を取得する攻撃である．また，スマートフォンやタブレット端末への入力操作に伴い発生する音響を利用して入力内容を推測する攻撃手法が提案されている [3]〜[8]．

近年は，各家庭で見守りカメラやスマートスピーカの普及が進んでいる．これらの IoT 機器にはマイクが内蔵されており，スマートスピーカに至っては，音声アシスタントを備え，音声インタフェースを介して様々な機能を実行することができる．この結果，アタックサーフェスは「音」にまで及ぶこととなった．現に，攻撃者あるいはマルウェアに操られた見守りカメラやスマートスピーカが，会話や生活音の盗み聞きを行い，プライバシーを脅かす事例が発生している [9], [10]．スマートフォンやタブレット端末の操作音を利用したサイドチャネル攻撃の脅威は，既に現実のものとなっている．

本研究では，正規ユーザが自身のスマートフォンに入力した重要な個人情報を，攻撃者が盗み聞きすることができる攻撃環境に対して，その脅威と防御策について調査する．これまでの既存研究では，正規ユーザの端末に対する積極的な干渉 [3]〜[5]，あるいは，正規ユーザの端末操作に関する事前知識 [6], [8] が必要な

[†] 静岡大学，浜松市
Shizuoka University, Hamamatsu-shi, 432–8011 Japan
a) E-mail: nisigaki@inf.shizuoka.ac.jp
DOI:10.14923/transfunj.2022BAP0004

攻撃シナリオが想定されていた．これに対し本論文では，「正規ユーザに対する事前知識をもち合わせていない攻撃者」が，「正規ユーザがスマートフォンにキー入力をする際のタップ音を，外部のマイクで受動的に盗聴する」という環境の下で，そのタップ音から入力内容を再現するという攻撃モデル [7] を扱う．

2. 関連研究

Shumailov らは，正規ユーザのスマートフォンやタブレットの内蔵マイクとタップ音によって入力内容を推測する手法を提案している [3]．正規ユーザの端末に複数のマイクが内蔵されている場合，タップした際に発生する音響は，上部に設置されているマイクと下部のマイクで受信する時間に差が生じる．この音響の到達時間の差から画面上のどこをタップしたときの音であるのかを計算し，正規ユーザのキー入力を約 61% の精度で推測可能であることを報告している．しかしこの攻撃手法は，タップ音を盗聴する録音デバイスが正規ユーザの端末に内蔵されているマイクであり，事前侵入が必要という点で妥当性を欠く攻撃シナリオとなっている．攻撃対象のスマートフォンに侵入できたのならば，攻撃者はキーロガー等を用いて正規ユーザのキー入力を直接取得できる．

Lu らは，攻撃者のスマートフォンの内蔵スピーカから攻撃対象のスマートフォンに向けてソナー音を放射し，タップ入力の際の正規ユーザの指からの反射波を分析することによって，正規ユーザのキー入力を約 90% の精度で推測可能であることを報告している [4]．しかしこの攻撃手法は，攻撃者が正規ユーザの端末に対して積極的に干渉するサイドチャネル攻撃となっており，能動的な攻撃シナリオであると言える．

Zhuang らは，PC の物理キーボードの打鍵音から入力内容を推測する攻撃を検証している [5]．正規ユーザの打鍵音を近辺に設置されているマイクから盗聴し，得られた音響データに対してケプストラム分析で特徴量抽出を行い，クラスタリングを行うことによって，正規ユーザのキー入力を約 96% の精度で推測可能であることを報告している．この事実は，スマートフォンにおいても，タップ入力音を外部マイクによって盗聴するだけで，正規ユーザのスマートフォンへの入力を推測できる可能性があることを意味している．

Zarandy らは，スマートスピーカを外部マイクとして用い，スマートフォンに対する正規ユーザのタップ音からその入力内容を推測する攻撃を検証している [6]．正規ユーザのタップ音を学習させた LDA（Linear Discriminant Analysis）識別器と CNN（Convolutional Neural Network）識別器を用い，スマートスピーカを介して収集した音響データを分析した結果，約 80% の精度で音響データ中のタップ音の発生を検出可能であること，正規ユーザが 1 桁の数字を入力した際のタップ音に対して約 40% の精度で入力した数字の識別が可能であることを報告している．ただし，文献 [6] の実験では，実験協力者数は 3 名であった．また，正規ユーザのタップ音を訓練用データから除き，正規ユーザ以外のタップ音を学習させた LDA 識別器，CNN 識別器を用いての実験についても実施しており，約 35% の精度で 1 桁の数字の識別が可能であるとしている．

Zarandy らの実験は，Shumailov らの研究や Lu らの研究とは異なり，「正規ユーザがスマートフォンにキー入力をする際のタップ音を，攻撃者が外部のマイクで受動的に盗聴する」という観点において，より現実的な攻撃モデルを想定している．特に，正規ユーザのタップ音を学習させていない識別器を用いた実験においては，「攻撃者は，正規ユーザに関する事前知識をもち合わせていない」という観点からも，現実的な攻撃モデルとなっている．しかし，Zarandy らの実験は，実験協力者 3 名のみの結果であり，かつ，正規ユーザ以外のタップ音を学習させた場合の実験については詳細な記載がない．

そこで本論文では，Zarandy らよりも更に現実的な観点で新たな攻撃実験を行う．具体的には，「正規ユーザに関する事前知識をもち合わせていない攻撃者（正規ユーザのタップ音は未学習）」が，「正規ユーザがスマートフォンにキー入力をする際のタップ音を，外部のマイクで盗聴する（受動的な盗聴）」という環境 [7] の下で，PIN 入力の内容をタップ音から推測するという攻撃実験を実施した．実験協力者は 11 名である．また，Zarandy らの研究では十分に検討されていなかった防御策についても，基礎的な検討と評価を行った．

3. 攻撃手法

本論文が想定する攻撃モデルを説明する．

3.1 攻撃者のターゲット

本実験における攻撃者のターゲットは，正規ユーザがスマートフォンに入力する文字情報とする．正規ユーザの使用するスマートフォンの機種については，攻撃者も既知であるとする．本研究の現段階では，キー入力を数字入力に限定して調査を行う．正規ユーザのス

マートフォンに，数字入力用のソフトウェアキーボード（図 1）を表示する．正規ユーザは，図 1 のソフトウェアキーボードを，自身の利き手の人差指，あるいは，スマートフォン用のタッチペンを用いてタップする．

3.2 攻撃者が用いる録音デバイス

本実験では，受動的な盗聴を想定する．すなわち攻撃者は，正規ユーザ（攻撃対象）のスマートフォンのタップ音を，その近辺に設置した録音デバイスを用いて盗聴する．具体的な録音デバイスとして，小型盗聴器とスマートスピーカを用いる．小型盗聴器は，「攻撃者が正規ユーザに物理的に近接し，攻撃者が隠しもつ小型盗聴器を用いて，正規ユーザのタップ音を盗聴する」というケースを想定したものである．スマートスピーカは，「攻撃者が何らかの方法(注1)で正規ユーザ宅のスマートスピーカを不正操作し，スマートスピーカに内蔵されているマイクを用いて，正規ユーザのタップ音を盗聴する」というケースを想定したものである．

3.3 攻撃者が用いる識別器のモデル

CNN 識別器を用いる．CNN の入力は，音響データのメル周波数スペクトログラムをヒートマップ化した画像（縦軸がメル周波数スペクトル，横軸が時系列）である．CNN の出力は，正規ユーザが入力したキー情報である．CNN は深層学習の一種で，画像識別で広く用いられている．文献 [11], [12] では，音響分類の手法の一つとして，メル周波数スペクトログラム画像と CNN を用いる手法が挙げられている．本実験は，この手法を採用した．本実験では数字キー入力（「1」，「2」，・・・，「9」，「0」）の識別が目的であるため，CNN の出力層は 10 クラス分類用の softmax 関数となっている．

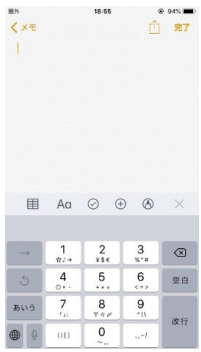


図 1 数字入力用のソフトウェアキーボード
Fig. 1 Numeric software keyboard.

(注1)：例 1：スマートスピーカの OS に脆弱性が存在していたため，攻撃者にその脆弱性を突かれ，スマートスピーカが不正操作されてしまった．例 2：攻撃者（正規ユーザの友人）がスマートスピーカに盗聴スキルを不正にインストールし，そのスマートスピーカを正規ユーザにプレゼントした．

3.4 攻撃者が用いる識別器の構成

本実験では，指とタッチペンの 2 種類のタップ手段，小型盗聴器とスマートスピーカの 2 種類の録音デバイスを用いる．攻撃者は，指によるタップ音，タッチペンによるタップ音，小型盗聴器による録音，スマートスピーカによる録音の全てに対応可能な汎用的な識別器（以降，汎用識別器）を一つ学習することが理想的ではあるが，本実験では実験の簡便化に配慮し，「指によるタップ音を小型盗聴器で録音したデータ」を用いて学習された識別器（以降，指 × 盗聴器用識別器），「タッチペンによるタップ音を小型盗聴器で録音したデータ」を用いて学習された識別器（以降，ペン × 盗聴器用識別器），「指によるタップ音をスマートスピーカで録音したデータ」を用いて学習された識別器（以降，指 × スピーカ用識別器），「タッチペンによるタップ音をスマートスピーカで録音したデータ」を用いて学習された識別器（以降，ペン × スピーカ用識別器）の四つを，それぞれ個別に用意した．

3.2 で述べたように，小型盗聴器を用いての録音は，「攻撃者が正規ユーザに物理的に近接し，攻撃者が隠しもつ小型盗聴器を用いて，正規ユーザのタップ音を盗聴する」というケースを想定したものである．この場合，攻撃者は正規ユーザのタップ手段を目視で確認可能であるため，汎用識別器を用いた場合のタップ手段の判断機構を目視での確認によって代替し，「タップ手段に応じて，指 × 盗聴器用識別器とペン × 盗聴器用識別器を使い分ける」ことによって汎用識別器を模擬できる(注2)．

同様に，スマートスピーカを用いての録音は，「攻撃者が正規ユーザ宅のスマートスピーカを不正操作し，スマートスピーカに内蔵されているマイクを用いて，正規ユーザのタップ音を盗聴する」というケースを想

(注2)：汎用識別器は，タップ音が入力として与えられると，「タップ手段の判断（指でタップされた音なのか，タッチペンでタップされた音なのか）」と「入力キーの識別（0 から 9 のどの数字のタップ音なのか）」を同時に行う．一方，指 × スピーカ用識別器とペン × スピーカ用識別器は，いずれも「入力キーの識別」のみを行う．小型盗聴器を用いての録音では，攻撃者自身が目視で正規ユーザのタップ手段を確認できるため，攻撃者が「タップ手段の判断」の機能を代行し，正規ユーザのタップ手段に応じて使用する識別器を選んで利用してやれば，汎用識別器を模擬できる．

定したものである．この場合，攻撃者は遠隔地にいるため，正規ユーザのタップ手段を目視で確認することはできない．しかし，識別対象のタップ音を指 × スピーカ用識別器とペン × スピーカ用識別器の両方に入力し，信頼度（CNN の出力層の各ニューロンの softmax 関数の出力値）が一番高い出力を採用するという方法を採れば，各識別器がタップ手段の特性に特化して学習を行っていることから，汎用識別器の模擬が可能である(注3)．

3.5 攻撃者が用意する訓練用データ

本実験では，正規ユーザに関する事前知識をもち合わせていない攻撃者を想定する．すなわち，識別器の訓練用データの中に正規ユーザのタップ音は含めない．ただし，比較実験として，訓練用データに正規ユーザのタップ音を含めた場合も取り扱う．本論文では，以降，前者を「事前知識なし学習」，後者を「事前知識あり学習」と呼ぶ．

事前知識なし学習のシナリオは次のとおりである．攻撃者は，真の攻撃対象（正規ユーザ）を狙う前に，同じ目的をもった攻撃仲間を集めることや，実験と称して希望者を募ることによって，多数のユーザのタップ音を収集することができる．そして，このようにして収集した（正規ユーザ以外の）多数のユーザのタップ音を訓練用データとして用いて，識別器の学習を行う．

3.6 攻撃環境

騒音環境の中ではタップ音はノイズにかき消されることになるが，タップ音とは異なる周波数帯のノイズに関しては，適正な音源分離技術の適用によってタップ音のみを抽出することが（理論的には）可能である．このため，本実験では，静音環境のみを対象として評価を行うこととした．

(注3)：スマートスピーカを用いての録音では，攻撃者も正規ユーザのタップ手段を確認することができないが，その代わりに，両方の識別器にタップ音を入力するという手段を採ることができる．指 × スピーカ用識別器は「指によるタップ音の識別」をするために学習されているので，指によるタップ音が入力された場合には指 × スピーカ用識別器のほうがより高い精度で入力キーを識別可能であることが期待される．同様に，ペン × スピーカ用識別器は「タッチペンによるタップ音の識別」をするために学習されているので，タッチペンによるタップ音が入力された場合にはペン × スピーカ用識別器のほうがより高い精度で入力キーを識別可能であることが期待される．このため，識別対象のタップ音を指 × スピーカ用識別器とペン × スピーカ用識別器の両方に入力し，一番高い信頼度を出力する識別器の結果を採用してやれば，汎用識別器を模擬できる．

4. 評価実験

4.1 実験環境

実験に使用した機器の諸元を表 1 に示す．実験は静岡大学浜松キャンパス情報学部棟 1 号館内の会議室で行い，会議室に設置されている机の上に，攻撃対象のスマートフォンと攻撃者の録音デバイス（小型盗聴器とスマートスピーカ）を設置した．本実験では，小型盗聴器としてタブレット端末を，スマートスピーカとして円形マイクアレー(注4)を使用した．

正規ユーザ役の実験協力者は，静岡大学の学生 11 名（21〜23 歳，男性 9 名，女性 2 名，右利き 11 名）である．実験協力者には，椅子に座ってスマートフォンを非利き手でもち，利き手の人差指で，あるいは，利き手でタッチペンをもって，数字キーのタップを行うよう指示した．その際，タブレット端末とマイクアレーから 5〜10 [cm] の位置でスマートフォンを操作してもらった．本実験で使用するタッチペンのペン先は

表 1 実験機器
Table 1 Experimental equipment.

種類	名称
スマートフォン	iPhone6
タブレット端末	iPad Pro 2018 11 インチ
マイクアレイ	ReSpeaker 6 マイク円型アレイキット Raspberry Pi 用
マイクアレイ制御用コンピュータ	Raspberry Pi 3
タッチペン	エレコム 電池式アクティブタッチペン（P-TPACST01BK）
分析用 CPU	2.3 GHz クアッドコア Intel Core i7
OS（PC）	macOS Big Sur 11.6
音声編集ソフト	Audacity 2.3.3
プログラミング言語	Python 3.7
音声処理ライブラリ	librosa 0.7.0
深層学習ライブラリ	Keras 2.3.1
	TensorFlow 1.14.0

(注4)：スマートスピーカとして広く知られている Amazon Echo や Google Home では，それらの音響データを外部に取り出すことはできないため，Zarandy らは ReSpeaker 円形マイクアレーを用いてスマートスピーカの録音環境を再現している [6]．本実験もこれに倣った．

1.5 [mm] の細さ，材質はポリアセタールとなっており，タップを行うとコツコツという小さな音が発生する．

実験環境を図 2 に示す．実験時の会議室は静寂であり，普通騒音計（リオン株式会社製 NL-42）を設置し，周波数重みづけ：A 特性，時間重みづけ：Fast 特性のモードで暗騒音を測定したところ，騒音レベルは 35〜45 [dB] であった．

4.2 音響データ収集

攻撃対象のスマートフォンの画面に日本語用ソフトウェアキーボードの PIN 入力インタフェースを表示させた（図 1）．イヤホンを装着した実験協力者 11 名に，10 種類の数字キーを「1」，「2」，・・・，「9」，「0」の順番でタップしてもらった．音響データの分析を簡易にするために，タップの際には，100 [bpm] のメトロノームの音声をイヤホンから流し，実験協力者はそのリズムに合わせてタップを行った．

実験協力者は，はじめにタッチペンを用いて入力を行う．「1」から「0」のタップ入力を 1 セットとし，実験協力者に 100 セット分の入力を繰り返してもらい，合計 1,000 回のタップ音を収集した．ただし，タップミスの発生や突発的な雑音の混入に備え，15〜20 セット余分にタップするよう指示した．このため，実際に収集したタップ音は約 1,200 回分である．実験協力者は，次に指を用いて入力を行う．指を用いた入力に対しても，同様の方法で，約 1,200 回のタップ音を収集した．なお，実験協力者の負担に配慮し，約 600 回のタップ音収集ごとに休憩を挟んだ．約 600 回のタップにおよそ 6 分を要する．どれくらい休憩を取るかは各実験協力者に任せたが，最大 5 分とした．

タップ音は，タブレット端末とマイクアレーの 2 種類のマイクで同時に録音する．タブレット端末では，内蔵されているマイクと録音アプリでタップ音の録音を行い，マイクアレーでは，マイクアレー制御用の Raspberry Pi 3 を用いて録音を行った．なお，マイクアレーはマイク機能のみを用いており，アレー処理は行っていない．録音する音響データの形式は wav である．ただし，タブレット端末は録音アプリの制約で m4a 形式での録音となるため，収録後に m4a から wav に変換した．

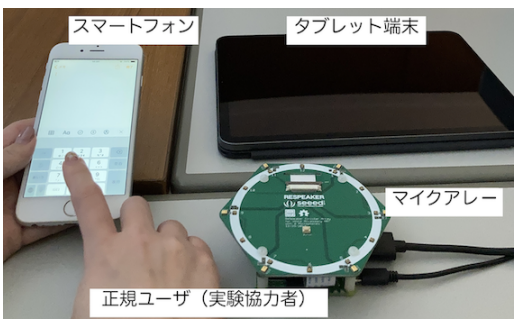


図 2 実験環境
Fig. 2 Experiment configuration.

4.3 音響データ処理

収録後，音声編集ソフトウェア Audacity [13] を用いて，音響データ全体を 0.6 [s] ごとに時分割した上で，一つの音響データが 1 タップ分（約 0.35 [s]）になるように切り出す(注5)ことにより，実験協力者 11 名のデータセットを作成した．実験実施者（著者ら）が個々の音響データを確認し，実験協力者のタップミスや突発的な騒音が認められた場合は，当該時間区間のタップ音をデータセットから除外した．タップ音の除外が生じた際には余分に録音しておいたタップ音を補填し，各実験協力者のデータセットが全て「各数字キーのタップ音 ×100 セット = 1000 個のタップ音」になるようにした．

本実験では，10 種類の数字キーを「1」，「2」，・・・，「9」，「0」の順番でタップしてもらっているため，実験協力者がどのキーを入力した際の音響データであるか自明である．1000 個のタップ音のそれぞれに対し，入力したキーの情報を正解ラベルとして付与した．

データセット中のすべてのタップ音に対し，Python の音声処理ライブラリ librosa を利用して，各 1 タップ分の音響データのメル周波数スペクトログラム画像を作成した．作成した画像例を図 3 に示す．横軸が時間，縦軸が周波数のヒートマップ画像である．実際の学習・識別においては，カラーバー，x 軸，y 軸，ラベルの表示は削除し，640 × 480 ピクセルの画像情報として CNN に入力した．メル周波数スペクトログラム画像はカラー画像のため，RGB の 3 チャネルの画像情報として CNN に入力される．

指によるタップ音をタブレット端末（小型盗聴器）

(注5)：Zarandy ら [6] は「録音された音響データからタップ音を切り出す処理（前処理）」と「切り出したタップ音からキー入力を推定する処理（識別処理）」の両方を扱っているのに対し，本論文では後者の識別処理に焦点を当てている．このため，本実験では，前者の前処理に対しては簡易な方法でこれを行っている．具体的には，**4.2** で説明したとおり，実験協力者に 100 [bpm] のメトロノームの音声に合わせてタップを行ってもらうことによって，音響データを 0.6 [s] ごとに時分割すれば個々のタップ音を切り分けることができるようにしている．そして，切り分けられた 0.6 [s] ごとの各音響データに対し，両端の無音部分を削除することによって 1 タップ分の音響データ（約 0.35 [s]）を切り出している．

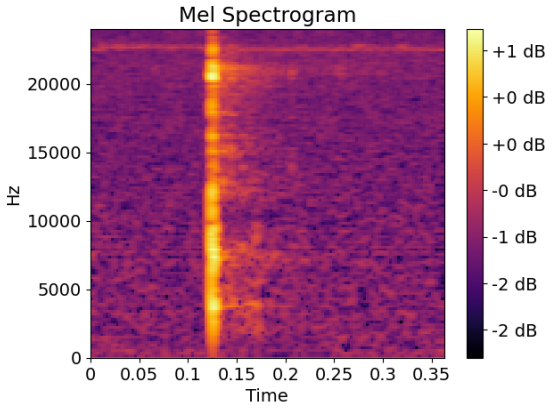


図 3　メル周波数スペクトログラム画像
Fig. 3　Mel-frequency spectrogram.

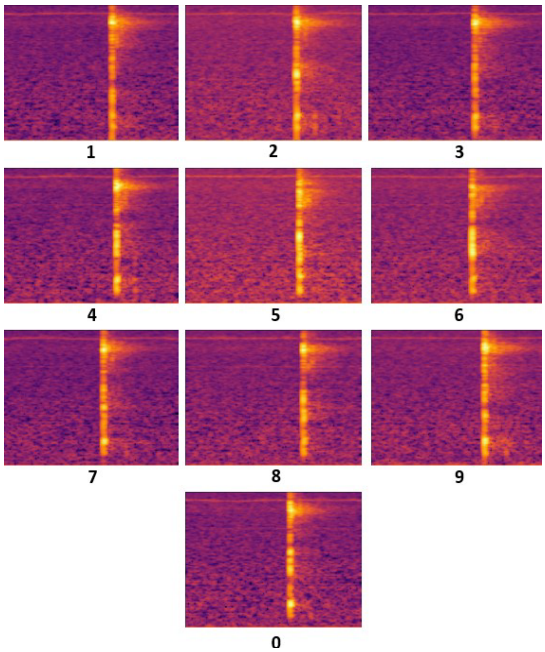


図 4　ある実験協力者が各数字をタップした場合のメル周波数スペクトログラム画像
Fig. 4　Mel-frequency spectrograms of each number tapped by a user.

で録音した音響データに対して上記の処理を適用することによって，指 × 盗聴器用のデータセットが得られる．同様に，指によるタップ音をマイクアレー（スマートスピーカ）で録音した音響データに対して，上記の処理を適用することによって，指 × スピーカ用のデータセットが得られる．タッチペンによるタップ音をタブレット端末（小型盗聴器）で録音した音響データに対して上記の処理を適用することによって，ペン × 盗聴器用のデータセットが得られる．タッチペンによるタップ音をマイクアレー（スマートスピーカ）で録音した音響データに対して，上記の処理を適用することによって，ペン × スピーカ用のデータセットが得られる．

数字ごとのタップ音の違いを示すために，ある実験協力者が各数字をタップした際のメル周波数スペクトログラム画像を図 4 に示す．図 4 は，一具体例として，ペン × スピーカ用のデータセットの中から，ある 1 名の実験協力者の「1」から「0」のタップ音を任意に抽出したものである．

4.4　機 械 学 習

指 × 盗聴器用データセット，指 × スピーカ用データセット，ペン × 盗聴器用データセット，ペン × スピーカ用データセットのそれぞれに対し，Python の深層学習ライブラリ Keras，TensorFlow を用いて CNN の学習及び識別を行った．CNN のネットワーク構成は文献 [14] を参考にした．本実験で使用した CNN モデルを図 5 に示す．

4.3 で作成したメル周波数スペクトログラム画像（RGB の 3 チャネルの画像）を CNN の入力として与える．まず，これを 50 × 50 ピクセルに圧縮した上で，CNN は 3 × 3 のフィルタを用いて 2 連続で畳み込みを行い，32 枚の特徴量マップを得る．次に，Max プーリングにより画像サイズを半分に縮小する．本実験では Max プーリングを適用する際の小領域サイズは 2 × 2 とした．更に，畳み込みを 2 連続で行い，Max プーリングを行った．この結果得られた 3 次元の配列を 1 次元のベクトルに並び替え，全結合層につなげた．本実験では数字キー入力の識別（10 クラス分類）が目的であるため，最後に 10 個のノードをもつ全結合層につなげた．活性化関数は出力層では softmax 関数，その他の層では ReLU 関数を用いた．

3.5 で述べたように，本実験では，正規ユーザに関する事前知識をもち合わせていない攻撃者を想定する．すなわち，各実験協力者に対して「事前知識なし学習」型の評価実験を行う必要がある．事前知識なし学習の具体的な手順は次のとおりである．

1. 実験協力者 11 人分のデータセットを「1 番目の実験協力者のデータセット」と「2〜11 番目の実験協力者のデータセット」に分割し，前者をテスト用データとして，後者を訓練用データと

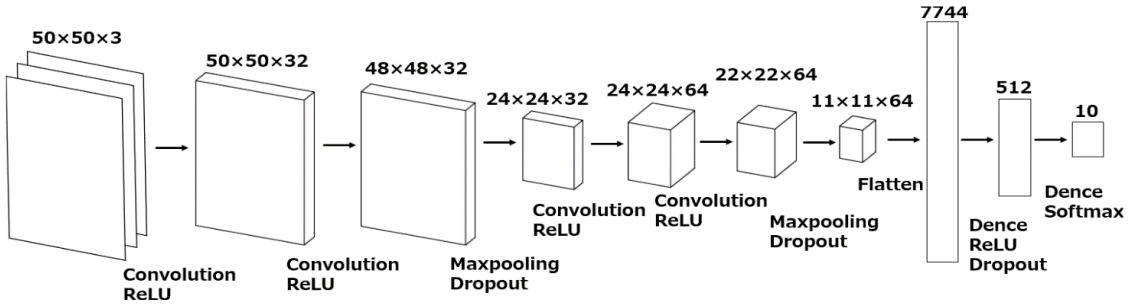


図 5　CNN モデル [14]
Fig. 5　CNN model.

して，それぞれ使用する．

2. 訓練用データ（2〜11 番目の実験協力者のデータセット）を用いて CNN を学習する(注6)．その際，訓練用データを実験協力者ごとに更に 8:2 の割合で分割し，前者を学習用データとして，後者を検証用データとして，それぞれ使用する．
3. テスト用データ（1 番目の実験協力者のデータセット）を用いて CNN をテストすることによって，1 番目の実験協力者に対する「事前知識なし学習」型の評価実験を行う．
4. 2〜11 番目の実験協力者のデータセットに対しても，手順 1〜3 と同様の処理を行うことによって，各実験協力者に対する「事前知識なし学習」型の評価実験が実行される．

一方，比較のために実施する「事前知識あり学習」型の評価実験の手順は次のとおりである．

1. 各実験協力者のデータセットを 8:2 の割合で分割し，前者を訓練用データとして，後者をテスト用データとして，それぞれ使用する．訓練用データについては更に 8:2 の割合で分割し，前者を学習用データとして，後者を検証用データとして，それぞれ使用する（すなわち，学習用データ，検証用データ，テスト用データは 16:4:5 の割合）．
2. 実験協力者 11 人分の訓練用データ（学習用データ，検証用データ）を用いて CNN を学習する．
3. 1 番目の実験協力者のテスト用データを用いて CNN をテストすることによって，1 番目の実験協力者に対する「事前知識あり学習」型の評価実験を行う．
4. 2〜11 番目の実験協力者のデータセットに対しても，手順 3 と同様の処理を行うことによって，各実験協力者に対する「事前知識あり学習」型の評価実験が実行される．

4.5 実験結果

指 × 盗聴器用データセット，指 × スピーカ用データセット，ペン × 盗聴器用データセット，ペン × スピーカ用データセットのそれぞれに対し，**4.4** で説明した「事前知識なし学習」型の評価実験を行った．

3.4 で述べたように，タブレット端末（小型盗聴器）録音の場合は，攻撃者が正規ユーザのタップ手段を目視で確認可能であるという想定である．攻撃者は，正規ユーザが指でタップしていた場合には指 × 盗聴器用識別器を用いて識別を行い，正規ユーザがペンでタップしていた場合にはペン × 盗聴器用識別器を用いて識別を行う．したがって，「指 × 盗聴器用データセット中のテスト用データ」に対する識別精度は「テスト用データを指 × 盗聴器用識別器によって識別する」という実験によって，また，「ペン × 盗聴器用データセット中のテスト用データ」に対する識別精度は「テスト用データをペン × 盗聴器用識別器によって識別する」という実験によって，それぞれ測る形となる．

一方，マイクアレー（スマートスピーカ）録音の場合は，攻撃者は遠隔にいるという想定である．攻撃者は，識別対象のタップ音を指 × スピーカ用識別器とペン × スピーカ用識別器の両方に入力して識別を行う．したがって，「指 × スピーカ用データセット中のテスト用データ」に対する識別精度，並びに，「ペン × スピーカ用データセット中のテスト用データ」に対する

(注6)：指 × 盗聴器用データセットを用いて CNN を学習することによって得られる識別器が，**3.4** で説明した指 × 盗聴器用識別器である．同様に，指 × スピーカ用データセット，ペン × 盗聴器用データセット，ペン × スピーカ用データセットのそれぞれを用いて CNN を学習することによって得られる識別器が，指 × スピーカ用識別器，ペン × 盗聴器用識別器，ペン × スピーカ用識別器である．

表 2　タップ音からの入力キー識別精度
Table 2　Key input identification accuracy estimated by tap sounds.

	録音デバイス ／ タップ手段	小型盗聴器（タブレット端末）	スマートスピーカ（マイクアレイ）
事前知識あり学習	タッチペン	76.5%	85.6%
	指	66.1%	84.3%
事前知識なし学習	タッチペン	29.4%	26.1%
	指	23.3%	19.1%

識別精度は，どちらも，「テスト用データを指 × スピーカ用識別器とペン × スピーカ用識別器の両方に入力し，信頼度（CNN の出力層の各ニューロンの出力値）の高いほうの識別結果を採用する」という実験によって測る形となる．

それぞれの評価実験における入力キー識別精度を表 2 に示す．表 2 の数値は，5-fold Cross Validation を用いて 5 回分の評価で得られた識別率の平均値である．表 2 より，正規ユーザ本人のタップ音を学習していない識別器（事前知識なし学習）を用いた場合であっても，キー入力のタップ音から正規ユーザが入力したキーの情報が漏れることが分かる．具体的には，タブレット端末で録音したタップ音からは，タッチペン入力の場合は 29.4%，指入力の場合は 23.3% の精度で，また，マイクアレーで録音したタップ音からは，タッチペン入力の場合は 26.1%，指入力の場合は 19.1% の精度で，それぞれキー入力が識別されることが判明した．

比較のために，指 × 盗聴器用データセット，指 × スピーカ用データセット，ペン × 盗聴器用データセット，ペン × スピーカ用データセットのそれぞれに対し，**4.4** で説明した「事前知識あり学習」型の評価実験も行った．表 2 にはその結果も併記してある．正規ユーザ本人のタップ音を事前に入手して識別器の学習に利用できる環境に攻撃者がいる場合には，タップ音からのキー入力の漏洩は深刻（66.1%〜85.6% の識別精度）となることが示された．

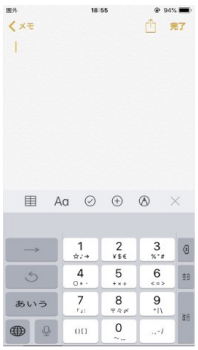


図 6　右配置ソフトウェアキーボード
Fig. 6　Right-aligned software keyboard.

5. 防 御 手 法

5.1 キーボード位置の変更

剛体であっても，力が加わると微視的なたわみが生じる．スマートフォンの物理的な構造上，ディスプレイはその中心部ほどたわみ易い．また，スマートフォン内部には様々な部品が密集しており，ディスプレイ直下の空間（ディスプレイから部品までの距離）は場所によってわずかに異なる．このような理由で，ディスプレイのどこをタップするかによって，発生する音（タップ音）に図 4 に見られるような差が生まれる．識別器は，このタップ音の違いを利用し，入力されたキーの識別を行っていると考えられる．つまり，識別器の中でタップ音とタップ位置が対応付けられることで，キー入力の識別が行われる．

したがって，キーボードの表示位置を変更するという措置は，単純でありながら，大きな効果が期待される対策であると考えられる．キー操作のたびに全てのキーの位置をシャッフルすることで，タップ音と各キーの対応関係を完全に崩すことが可能である．しかし，キー配置の動的な変更は，ユーザの利便性低下に直結する．そこで本研究の現段階では，キーボード上の全てのキーの位置を一律に移動する方法を試すこととした．

4. の実験で用いたキーボード（iPhone の iOS 標準キーボード）に対し，全てのキーの配置を，1 つのキーの半分の幅だけ右側にシフトさせたキーボードを自作した（図 6）．正規ユーザによるキー操作が発生するたびに，キー入力のセッション単位で（キーを 1 文字入力するたびにキーボードの配置が変更されるのではなく，エンターキーが入力されたタイミングで），標準

表 3　右配置キーボードの識別精度（事前知識なし学習）
Table 3　Key input identification accuracy using right-aligned keyboard. (Learning without previous knowledge)

録音デバイス / タップ手段	小型盗聴器（タブレット端末）	スマートスピーカ（マイクアレイ）
タッチペン	20.0%	16.9%
指	17.6%	16.3%

キーボード（図 1）が表示されるか，右配置のキーボード（図 6）が表示されるかが，ランダムに決定される．例えば「1」のキーは，「1」と「2」のキーの中間地点に配置されることになるため，標準キーボードと右配置キーボードのどちらが表示されているか分からない攻撃者（識別器）は，タップ音からだけでは「1」と「2」の識別が不可能（2 分の 1 の確率）となる．

5.2　右配置キーボードに対する識別精度

5.1 で説明した右配置キーボードの効果を検証するために，「右配置キーボードを利用した際のタップ音」を「標準キーボード用の識別器」を用いて識別させる実験を行った．右配置キーボードを利用した際のタップ音は，**4.1**〜**4.3** と同様（実験協力者が操作するスマートフォンのディスプレイに表示されているキーボードのみが異なる）の方法・手順で，同じ 11 名の実験協力者から新たに収集した．標準キーボード用の識別器は，**4.4** の識別器をそのまま利用した．**4.5** の「事前知識なし学習」型の評価実験と同様の実験を行った結果を表 3 に示す．

「指 × 盗聴器用データセット」の識別精度どうしを比較すると，表 3 の数値は表 2 よりも低い．「指 × スピーカ用データセット」，「ペン × 盗聴器用データセット」，「ペン × スピーカ用データセット」の個々の識別精度どうしを比較した結果も同様である．以上より，キー配置の変更が，タップ音からのキー入力の識別精度を減少させることが示された．

6. 考　　察

6.1　PIN コードに対する識別精度

4.5 及び **5.2** では，1 桁の数字入力に対する識別精度を評価した．実際の攻撃シーンでは，攻撃者は，スマートフォンに入力された複数桁の入力を識別する必要がある．そこで本節では，4 桁の PIN コードを対象に，4 桁全てを識別するために必要な試行回数を評価する．

標準キーボードを用いて 4 桁 PIN コードを入力した場合のタップ音は，**4.1**〜**4.3** のデータセットを合成（4 のタップ音を連結）することによって，0000 から 9999 の 1 万パターンの PIN コードを作成した．同様に，右配置キーボードを用いて 4 桁 PIN コードを入力した場合のタップ音は，**5.2** のデータセットを合成（4 のタップ音を連結）することによって作成した．識別器は，**4.4** の識別器をそのまま利用した．実験の内容は，基本的には **4.5** の「事前知識なし学習」型の評価実験と同じであるが，4 桁 PIN コードの推測アルゴリズムは次のとおりとした．

1. 攻撃者は，**4.4** の識別器を用い，4 桁のキー入力をそれぞれ個別に識別する．
2. 手順 1 の操作によって，各桁のキー入力に対する信頼度（CNN の出力層の各ニューロンの出力値）を得る．桁ごとに信頼度の高い順に数字を並べる．1 桁目の数字を，信頼度の高い順に，第 1 桁第 1 候補，第 1 桁第 2 候補，・・・と呼ぶ．第 2〜4 桁目の数字も同様である．
3. 第 1 桁第 1 候補，第 2 桁第 1 候補，第 3 桁第 1 候補，第 4 桁第 1 候補の数字を選び，この 4 桁の組合せを PIN コード第 1 候補とする．PIN コードが正解であった場合は終了．
4. 第 1 桁第 2 候補，第 2 桁第 2 候補，第 3 桁第 2 候補，第 4 桁第 2 候補の数字の内で，信頼度が一番大きい数字を選ぶ．例えば第 2 桁第 2 候補の信頼度が一番大きかったとする．第 1 桁第 1 候補，第 2 桁第 2 候補，第 3 桁第 1 候補，第 4 桁第 1 候補の数字を選び，この 4 桁の組合せを PIN コード第 2 候補とする．PIN コードが正解であった場合は終了．
5. 第 1 桁第 2 候補，第 2 桁第 2 候補，第 3 桁第 2 候補，第 4 桁第 2 候補の数字の内で，信頼度が 2 番目に大きい数字を選ぶ．例えば第 4 桁第 2 候補の信頼度が 2 番目に大きかったとする．第 1 桁第 1 候補，第 2 桁第 1 候補，第 3 桁第 1 候補，第 4 桁第 2 候補の数字を選び，この 4 桁の組合せを PIN コード第 3 候補とする．PIN コードが正解であった場合は終了．
6. 正解の PIN コードが見つかるまで，信頼度を 1 段階ずつ下げながら手順 5 を繰り返すことによって，PIN コードの次の候補を試していく．

図 7 に，標準キーボードの場合の推測試行回数ごとの PIN コード識別成功率を示す．図 7 から，500 番目の

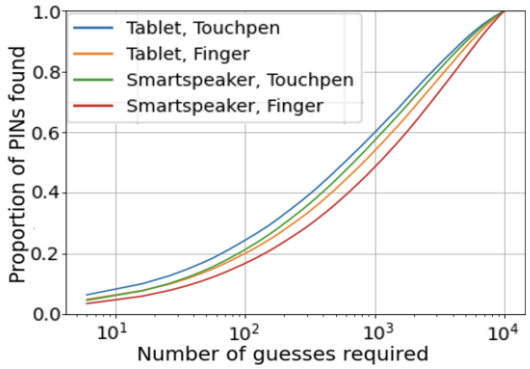


図 7　各推測試行回数の PIN コード識別成功率（標準キーボード）

Fig. 7　Proportion of PINs found with a limited number of guesses. (Standard keyboard)

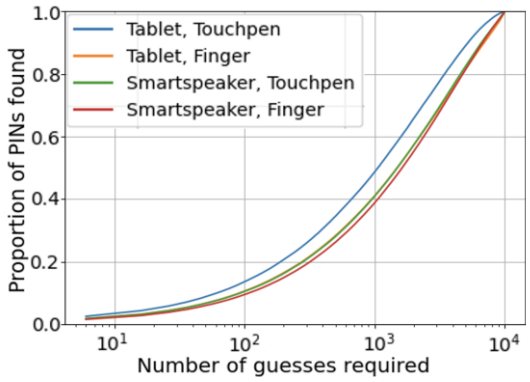


図 8　各推測試行回数の PIN コード識別成功率（右配置キーボード）

Fig. 8　Proportion of PINs found with a limited number of guesses. (Right-aligned keyboard)

候補まで試す機会が攻撃者に与えられたならば，攻撃者は約 50% の確率で PIN コードを識別可能であることが分かる．また，図 8 に，右配置キーボードの場合の推測試行回数ごとの PIN コード識別成功率を示す．図 8 から，標準キーボードと右配置キーボードを混在させることにより，PIN コード推測成功率 50% に対する推測試行回数の期待値を約 500 回から約 1,000 回に増加させることが可能であることが確かめられた．

6.2　制　　限

本論文の実験からは，スマートフォンのタップ音からキー入力を識別するという攻撃が，CNN を用いることによって，チャンスレートを超えた確率で実行可能であるという結果が得られた．ただし，本論文で行った実験においては，多くの前提を設けており，今後検討すべき点が幾つか存在する．

第一に端末依存性である．本実験では特定のスマートフォン端末と録音デバイス（タブレット端末，マイクアレー）の組み合わせに対して評価を行った．しかし，タップ音の特性は正規ユーザが使用しているスマートフォン，攻撃者が使用する録音デバイスに依存すると考えられる．今後は多様なスマートフォン，録音デバイスに対する評価を検討する必要がある．

第二に攻撃対象のデバイスである．本実験ではキー入力を数字入力に限定している．今後は，フリック入力型 50 音キーボードや QWERTY キーボードに対する識別精度についても検討していく必要がある．QWERTY キーボードについては，数字入力用キーボードと比べて一つ一つのキーの表示領域が小さいため，タップ音からの入力キー識別は難しくなると考えられる．また，キーの配置を左右にシフトさせることによってキーの位置が完全に一つ隣にずれるため，入力キー識別攻撃に対する対策効果も大きいことが期待される．

第三に攻撃対象のタップ方法である．本実験では，実験協力者には椅子に座り，左手（利き手の逆）でスマートフォンをもち，右手（利き手）の人差し指で，あるいは，右手にもったタッチペンでタップを行ってもらった．しかし実際の攻撃環境では，正規ユーザが立っている場合や，両手でスマートフォンを操作する場合，右手でもったまま右手で操作を行う場合など，様々な状況が考えられる．攻撃対象である正規ユーザの状況が多様になるほど，攻撃成功率は低下するものと思われるため，今後はこれらの状況での評価を行う必要がある．

第四に実験協力者数である．本実験では 11 名による実験，評価を行った．今後は実験協力者数を増やして評価を行っていく．

第五に攻撃環境である．本実験では，タップ音とノイズは適切な音響処理によって音源分離可能であるという前提の下に，静音環境下での実験のみを行った．今後は屋外など様々な騒音環境においても実験を行い評価する必要がある．

第六に攻撃対策である．本実験は，キーの配置を右側に半分シフトさせたキーボードの導入効果のみを検証した．今後は様々な攻撃対策を検討していく必要がある．例えば，1 桁・1 文字のキー入力の識別精度から，タップ音を利用したサイドチャネル攻撃に対して安全な PIN コードの桁数やパスワードの文字数を逆算することや，キー入力の際にスマートフォンがタップ音と同じ周波数帯の音を発するような改良を提案して

いくなどのアプローチが考えられる．

7. 研究倫理

本研究は提案シナリオにおけるタップ音識別の脅威を調査することを目的とし，実験を通じて相応の危険性が確認される結果となった．しかし，本研究はまだ多くの実験条件の制約を前提としており，「実際の製品に対する現実的な脅威」レベルとは大きな隔たりがある．本研究の目的は「タップ入力」自体の安全性に関する一般的な性質を調査することにあり，特定の機種に対する攻撃を狙ったものではない．今後は調査結果の進展によって，脅威レベルに応じて関係者と連携して対応を進めていく予定である．

8. むすび

本研究では，「正規ユーザがスマートフォンに数字入力（0 から 9 の数字キー入力）を行う際のタップ音を，攻撃者が外部のマイクで受動的に盗聴する」という攻撃シナリオにおいて，正規ユーザの入力情報が攻撃者にどの程度漏れるのかについて検証を行った．複数名の既知のユーザのタップ音で学習した識別器を用いて，未知のユーザの数字入力を識別する事前知識なし学習において，約 30% の精度で数字入力の識別が可能であることが明らかとなった．そして，防御環境を模擬した実験からは，キーボードの適切な変更が攻撃の防御に貢献することが示唆される結果が得られた．今後は，現実的な状況を考慮した実験環境での実験を繰り返し，タップ音によるキー入力の識別精度を更に精査していく必要がある．

文　献

[1] 本間尚文，青木孝文，“知っておきたいキーワード（第 58 回）サイドチャネル攻撃，”映像情報メディア学会誌，vol.64, no.11, pp.1576–1579, Nov. 2010. DOI: 10.3169/itej.64.1576
N. Homma and T. Aoki, “Keywords you should know (vol.58) Side-channel Attack,” J. Institute of Image Information and Television Engineers, vol.64, no.11, pp.1576–1579, Nov. 2010. DOI: 10.3169/itej.64.1576

[2] National Security Agency, “TEMPEST fundamentals,” NACSIM 5000, Feb. 1982.

[3] I. Shumailov, L. Simon, J. Yan, and R. Anderson, “Hearing your touch: A new acoustic sidechannel on smartphones,” arXiv: 1903.11137, https://arxiv.org/abs/1903.11137, March 2019. DOI: 10.48550/arXiv.1903.11137

[4] L. Lu, J. Yu, Y. Chen, Y. Zhu, X. Xu, G. Xue, and M. Li, “KeyListener: Inferring keystrokes on QWERTY keyboard of touch screen through acoustic signals,” IEEE INFOCOM 2019-IEEE Conf. Computer Communications, pp.775–783, April-May 2019. DOI: 10.1109/INFOCOM.2019.8737591

[5] L. Zhuang, F. Zhou, and J.D. Tygar, “Keyboard acoustic emanations revisited,” ACM Conf. Computer and Communications Security, pp.373–382, Nov. 2005. DOI: 10.1145/1609956.1609959

[6] A. Zarandy, I. Shumailov, and R. Anderson, “Hey Alexa what did I just type? Decoding smartphone sounds with a voice assistant,” arXiv: 2012.00687, https://arxiv.org/abs/2012.00687, Dec. 2020. DOI: 10.48550/arXiv.2012.00687

[7] 大内結雲，奥寺瞭介，塩見祐哉，大木哲史，西垣正勝，“スマートフォンのタップ音からの入力内容推測可能性に関する研究（その 2），”暗号と情報セキュリティシンポジウム 2021（SCIS2021）予稿集，2D2-1, Jan. 2021.
Y. Ouchi, R. Okudera, Y. Shiomi, T. Ohki, and M. Nishigaki, “Study on possibility of estimating smartphone inputs from tap sounds (part2),” Proc. 2021 Symposium on Cryptography and Information Security, 2D2-1, Jan. 2021.

[8] 大内結雲，奥寺瞭介，塩見祐哉，上原航汰，杉本彩歌，大木哲史，西垣正勝，“スマートフォンのタップ音からの入力内容推測可能性に関する研究，”暗号と情報セキュリティシンポジウム 2020（SCIS2020）予稿集，1E2-4, Jan. 2020.
Y. Ouchi, R. Okudera, Y. Shiomi, K. Uehara, A. Sugimoto, T. Ohki, and M. Nishigaki, “Study on possibility of estimating smartphone inputs from tap sounds,” Proc. 2020 Symposium on Cryptography and Information Security, 1E2-4, Jan. 2020.

[9] Fox19 Digital Media Staff, “Hacker hijacks baby monitor,” Fox19NOW, https://www.fox19.com/story/25310628/hacked-baby-monitor/, 参照 March 16, 2022.

[10] S. Wolfson, “Amazon’s Alexa recorded private conversation and sent it to random contact,” The Guardian, https://www.theguardian.com/technology/2018/may/24/amazon-alexa-recorded-conversation, 参照 March 16, 2022.

[11] H. Purwins, B. Li, T. Virtanen, J. Schluter, S. Chang, and T. Sainath, “Deep learning for audio signal processing,” Journal of Selected Topics in Signal Processing, vol.13, no.2, pp.206–219, May 2019. DOI: 10.1109/JSTSP.2019.2908700

[12] K.J. Piczak, “Environmental sound classification with convolutional neural networks,” IEEE International Workshop on Machine Learning for Signal Processing, pp.1–6, Sept. 2015. DOI: 10.1109/MLSP.2015.7324337

[13] Audacity, “The free, cross-platform sound editor,” The Audacity Team, https://www.audacityteam.org/, 参照 March 16, 2022.

[14] Keras Documentation, “Train a simple deep CNN on the CIFAR10 small images dataset,” Keras, https://github.com/kerasteam/keras/blob/master/examples/cifar10_cnn.py, 参照 Dec. 12, 2019.

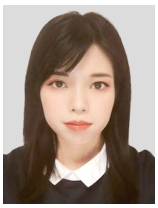
（2022 年 3 月 16 日受付，7 月 8 日再受付，
8 月 26 日早期公開）

大内　結雲

2020 静岡大学情報学部情報科学科卒．2022 同大学院総合科学技術研究科情報学専攻修士課程了．在学中は情報セキュリティに関する研究に従事．

野崎真之介

2022 静岡大学情報学部情報科学科卒．2022 同大学院総合科学技術研究科情報学専攻修士課程在学中．現在は情報セキュリティに関する研究に従事．

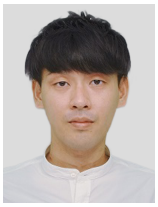


佐々木　葵

2022 静岡大学情報学部情報科学科卒．在学中は情報セキュリティに関する研究に従事．

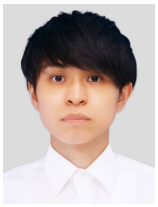
奥村　紗名

2022 静岡大学情報学部情報科学科卒．2022 同大学院総合科学技術研究科情報学専攻修士課程在学中．現在は情報セキュリティに関する研究に従事．



吉平　瑞穂

2021 静岡大学情報学部情報科学科卒．2022 同大学院総合科学技術研究科情報学専攻修士課程在学中．現在は情報セキュリティに関する研究に従事．



芹澤　歩弥

2021 静岡大学情報学部情報科学科卒．2022 同大学院総合科学技術研究科情報学専攻修士課程在学中．現在は情報セキュリティに関する研究に従事．

大木　哲史　（正員）

2002 早稲田大学理工学部電子情報通信学科卒．2004 同大学大学院理工学研究科電子・情報通信学専攻修士課程了．2010 早稲田大学理工学術院情報・ネットワーク専攻博士（工学）取得．2010 早稲田大学理工学総合研究所次席研究員，2013 産業技術総合研究所特別研究員を経て，2017 より静岡大学大学院総合科学技術研究科講師，2020 同大学准教授．情報セキュリティ全般，特に個人認証を中心としたネットワークセキュリティに関する研究に従事．情報処理学会会員．

西垣　正勝　（正員）

1990 静岡大学工学部光電機械工学科卒業．1995 同大学大学院博士課程了．日本学術振興会特別研究員（PD）を経て，1996 静岡大学情報学部助手．同講師，助教授の後，2010 より同創造科学技術大学院教授．博士（工学）．情報セキュリティ全般，特にヒューマニクスセキュリティ，メディアセキュリティ，ネットワークセキュリティ等に関する研究に従事．2013～2014 情報処理学会コンピュータセキュリティ研究会主査，2019～2020 情報環境領域委員長，2020 調査研究運営委員長．2015～2016 電子情報通信学会バイオメトリクス研究専門委員会委員長．2016～2020 日本セキュリティマネジメント学会編集部会長，2021 より副会長．情報処理学会フェロー．