SCIS 2020 2020 Symposium on
Cryptography and Information Security
Kochi, Japan, Jan. 28 – 31, 2020
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
A Study on Possibility for Guessing Smartphone Inputs from Tap Sounds

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††
Yumo OUCHI* Ryosuke OKUDERA* Yuya SHIOMI†
Kota UEHARA† Ayaka SUGIMOTO† Tetsushi OHKI† Masakatsu NISHIGAKI††

あらまし スマートフォンのキー入力盗取に関するサイドチャネル攻撃として，タップ音響からその入力内容が傍受可能であるという脅威が存在する．Ilia らは，正規ユーザのスマートフォン(タブレット端末)の内蔵マイクによって盗聴したタップ音から，キー入力を 61%の精度で推測可能であることを報告している．Li らは，攻撃者のスマートフォンの内蔵スピーカからソナー音を放射し，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， キー入力を 90%の精度で推測可能であることを報告している．しかし，Ilia らの方法は，攻撃対象のスマートフォンに事前侵入をする必要があるという点で，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，不正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)． また，Li らの方法は，攻撃者が正規ユーザの端末に対して積極的に干渉するタイプの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．そこで本稿では，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単純に盗聴するという受動的な攻撃シナリオにおいて，正規ユーザのキー入力に関する情報が攻撃者にどの程度漏れるのかを検証する．今回は，本研究の第一歩として，キー入力を PIN 入力に限定して調査を行った．タップ音の音声データのメル周波数ケプストラム係数を画像データとして表現し，畳み込みニューラルネットワークを用いてキー入力の推定を行った結果，高い精度で入力の識別が可能であることが判明した．

キーワード タップ音，キー入力盗取，機械学習，MFCC

1 はじめに

近年，スマートフォンの普及やキャッシュレス決済サービスの普及により，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増加している．そのような秘密情報を盗む攻撃手法の 1 つとしてサイドチャネル攻撃という攻撃が存在する[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 1 つにテンペスト攻撃という攻撃が存在する[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や音を検知することで，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．このようなテンペスト攻撃の一手法として，スマートフォンやタブレット端末への入力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][4][5]．しかし，既存の手法は正規ユーザの端末に対して，積極的な干渉が必要な攻撃シナリオを想定しており，現実的な脅威にはなり得ない．

そこで本研究ではスマートフォンのタップ音を用いて入力内容を推測する受動的な攻撃を想定し，その脅威の深刻度を評価するとともに防御策を検討する．

2 関連研究

スマートフォンのキー入力盗取に関するサイドチャネル攻撃として，タップ音響からその入力内容が傍受可能であるという脅威が存在する．

Ilia 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，

* 静岡大学情報学部情報科学科,Faculty of Informatics, Shizuoka University.
† 静岡大学大学院総合科学技術研究科，Graduate School of Integrated Science and Technology ， Shizuoka University
†† 静岡大学創造科学技術大学院, Graduate School of Science and Technology, Shizuoka University

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出ることが示されている．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なのかを計算し，キー入力を 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，不正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．

Li らは攻撃者のスマートフォンの内蔵スピーカからソナー音を放射し，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，キー入力を 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タイプの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．

Li 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提案している[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，得た音声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，キー入力を 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．そこで本研究では，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単純に盗聴するという受動的な攻撃シナリオを想定し，その脅威の深刻度を評価するとともに防御策を検討する．

3　攻撃手法

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単純に盗聴するという受動的な攻撃シナリオにおいて，正規ユーザのキー入力に関する情報が攻撃者にどの程度漏れるのかを検証する．今回は，本研究の第一歩として，キー入力を PIN 入力に限定して調査を行った．

攻撃者は攻撃対象のスマートフォンのタップ音を，その近辺に盗聴器を設置することで盗聴する．収集した音声データは特徴量を抽出し，機械学習を用いて識別させる．識別器は畳み込みニューラルネットワーク(CNN: Convolutional Neural Network)を用いた．CNN の入力は，音声データをメル周波数ケプストラム係数(MFCC: Mel-Frequency Cepstrum Coefficients)の時間遷移をヒートマップ化した 2 次元画像である．CNN の出力は，正規ユーザが入力したキー情報である．MFCC は，音声認識で一般的に用いられているケプストラムの手法である． CNN は深層学習の一種で，画像識別で広く用いられているモデルである．また，MFCC の特徴量ベクトルを画像化し，CNN を用いて分類を行う研究が報告されている[6]．

今回のタップ音からのスマートフォン入力の推測可能性の検証の流れを図 1 に示す．攻撃対象のスマートフォンと盗聴用のマイクの距離を変化させて，図 1 の検証を繰り返す．

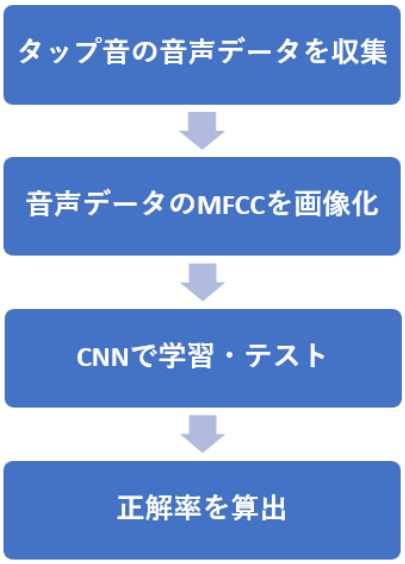

図 1 推測可能性検証の流れ

4　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．録音時には防音室を使用し，その暗騒音を測定した．攻撃対象のスマートフォンを置いた位置に普通騒音計（リオン株式会社製 NL-42）を設置し，周波数重みづけを A 特性，時間重みづけを Fast 特性で用いたところ，暗騒音レベルは 30-35dB であった．

表 1 実験機器

種類	名称
スマートフォン	iPhone6
タブレット	iPad Pro
分析用 CPU	Intel® Core™ i7-6500U@2.50GHz
OS(PC)	Windows 10 Pro
音声編集ソフト	Audacity
言語	Python3.7
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

4.2　音声データ収集

実験環境を図 2 に示す．防音室にて，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．イヤホンを装着し実験実施者（著者）1 名が，爪が画面に当たるように「0」から「9」の 10 種類の数字をそれぞれ連続で 100 回タップし，合計 1,000 回のタップを行った．音声データの分析を簡易にするために，タップの際には，100bpm のメトロノームの音声をイヤホンから流し，実験実施者はそのリズムに合わせてタップを行うようにした．

受動的な攻撃シナリオを想定した場合，録音用のマイクは精密なマイクや盗聴器よりも，攻撃者が普段使用しているモバイルデバイスを用いたほうが適していると考えられる．そこで今回は，同機種のタブレット端末を 4 台用意し，タブレット端末に内蔵されているマイクを録音用デバイスとして使用した．今回は，タブレット端末のサイズや防音室の規模の制限があるため，4 台のタブレット端末（録音用マイク）を攻撃対象のスマートフォンからそれぞれ 10cm，30cm，50cm，70cm の距離に設置し，4 台同時にタップ音を録音した．録音する音声データの形式は m4a である．

図 2 実験環境

4.3　音声処理

収録後，各数字の音声データを m4a 形式から wav 形式に変換した．音声編集ソフトウェア Audacity [7]を用いて，1 つの音声データが 1 タップ分になるように音声データ全体を約 0.35sec ごとに時分割した．**図 3** に音声データの分割の例を示す．

その後，Python の音声処理ライブラリ librosa を利用して，各 1 タップ分の音声データの MFCC から 2 次元画像（以降「MFCC 画像」と呼ぶ）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 20 次元の MFCC 係数のヒートマップである．なお，実際の識別においては，カラーバー，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出力した．MFCC 画像はカラー画像であるので，RGB の 3 チャネルの画像情報として CNN に入力される形になる．

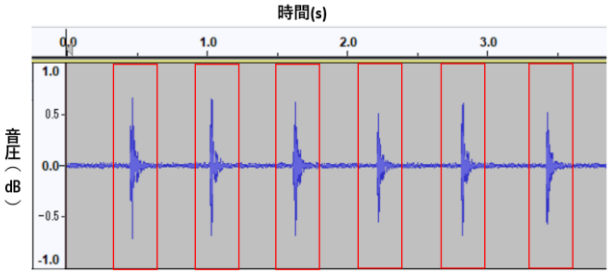

図 3 音声データの時分割

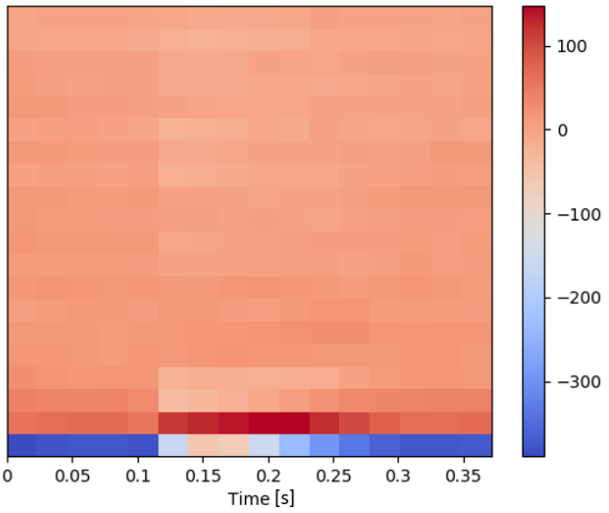

図 4 MFCC 画像

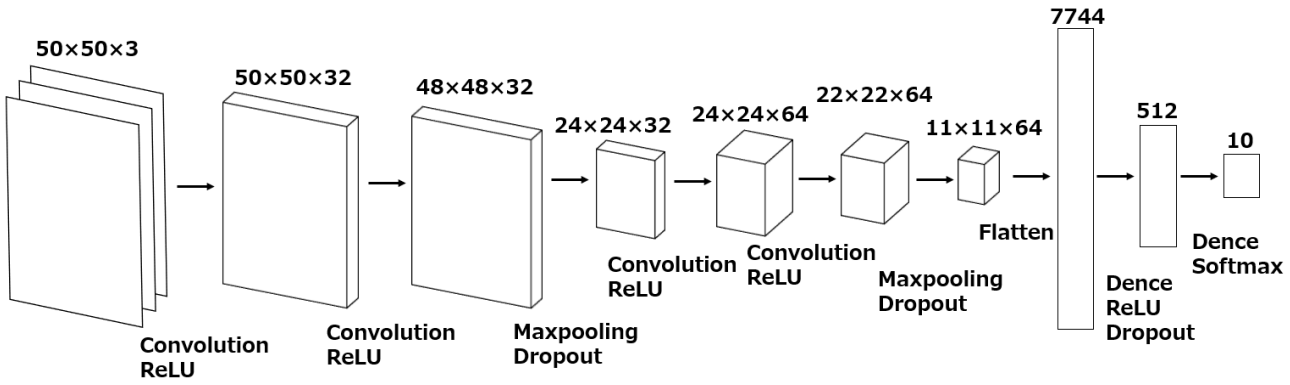

図 5　CNN モデル

4.4　機械学習

本研究では Python の深層学習ライブラリ Keras, TensorFlow を用いて CNN で学習および識別を行う．CNN のネットワーク構成は文献[8]を参考にした．今回使用した CNN モデルを図 5 に示す.

4.3 節で作成した MFCC 画像（RGB の 3 チャネルの画像）を CNN の入力として与える．CNN は，まず，これを 50×50 ピクセルに圧縮した上で，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（小領域に対して最大となる要素を取り出すプーリングの演算方式）により画像サイズを半分に小さくする．今回は小領域サイズを 2×2 で行った．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 次元に平滑化し，全結合層につなげた．今回は PIN 入力の推測（10 クラス分類）が目的であるため，最後に 10 個のノードを持つ全結合層につなげた. 活性化関数は出力層ではソフトマックス関数，それ他の層ではランプ関数を用いた.

今回の識別器の学習と評価においては，4.3 節で作成した MFCC 画像の全データを，テスト用と訓練用に 8:2 の割合で分割し，更に訓練用データを検証用と学習用に 8:2 の割合で分割して使用した.

5　実験結果

学習曲線を図 6, 図 7 に示す. 図 6 は縦軸が正解率，横軸がエポックの推移であり，acc は訓練データの正解率, val_acc は検証用データの正解率を示している. 図 6 は縦軸が損失，横軸がエポックの遷移であり，loss は訓練データの損失，val_loss は検証用データの損失を示している．図 7，図 7 から，100%に近い正解率で訓練データを学習できていることが分かる．検証用データの正解率も高いことから，過学習を起こすことなく，モデル作成を行えたと判断した.

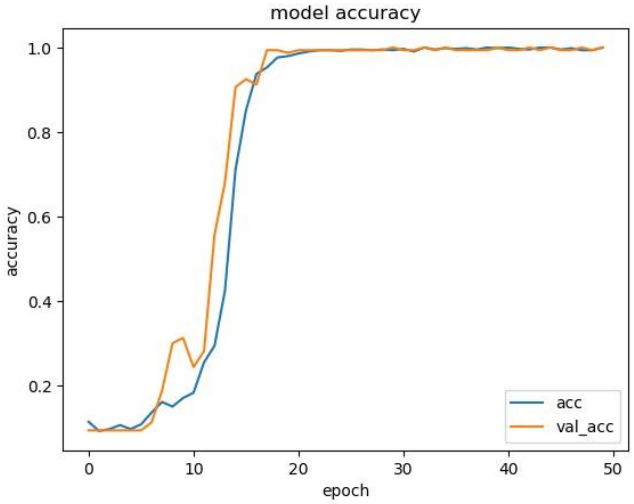

図 6 正解率の学習曲線

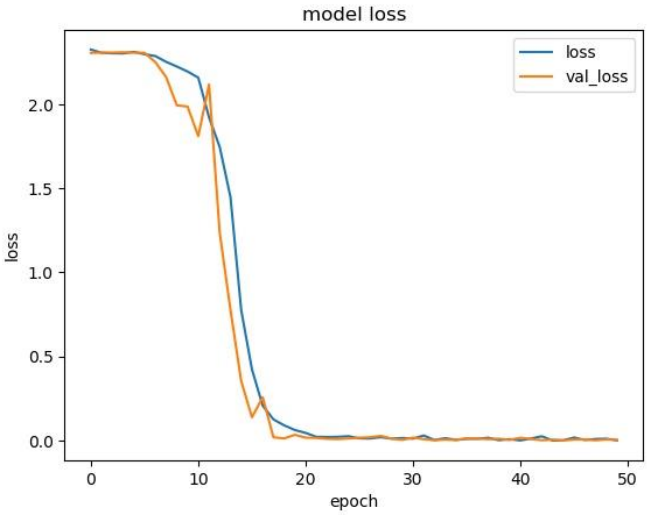

図 7 損失の学習曲線

学習を終えた CNN に対して評価用データを入力し，識別率を評価した．識別率の算出には 5-fold Cross Validation を用い，5 回分の評価で得られた正解率の平均を用いて評価を行った．結果を表 2 に示す.

表 2 距離ごとの正解率

距離	正解率
10cm	98.9%
30cm	99.0%
50cm	98.4%
70cm	96.3%

以上の結果から，本攻撃手法を用いた PIN 入力の識別は高い精度で可能であることが確認できた．また，今回の実験で調査した距離に対しては，正解率に大きな差異は見られなかった．

6 考察

6.1 制限

CNN を用いた入力内容予測攻撃は表 2 に示された通り，高い確率で実現可能であると推測できるが，今回の評価には多くの制限がある．

第一にタップの仕方である．今回の実験では爪を立ててスマートフォンにタップした際の音声を収集したため，正規ユーザの行動シナリオに対して限定的である．したがって，今後は爪を立てずに指の腹のみでタップした際の推測可能性について検証する必要がある．また，今回の実験では，「0」～「9」の各数字 1 つに対して，100 個のタップを連続かつ等時間間隔で行っている．そのため，今後はユーザの任意なタップしないようにした音声データを収集し，学習する必要がある．

第二に攻撃環境である．今回の実験では防音室にてデータ収録を行い，雑音が極力入らないような環境での評価を行った．今後は様々な騒音レベルの環境においても実験を行い評価する必要がある．また，今回は防音室の床にタップ用のスマートフォン，録音用のタブレットを設置した．しかし，スマートフォンをテーブルに置く場合や，手に持つ場合等，実際の攻撃環境では様々な状況が考えられる．今後は，ユーザが実際にスマートフォンを利用する状況で実験を行い評価する必要がある．

第三に端末依存性である．今回の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）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可能性がある．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．

第四に被験者数である．今回は，本研究の基礎検討の段階であったため，被検者 1 名による実験，評価を行った．今後は被検者を増やして評価を行っていく．

第五に攻撃対象である．今回は，本研究の基礎検討の段階であったため，キー入力を PIN 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 キーボードも検討していく必要がある．

第六に攻撃方法である．今回は，CNN の学習時において攻撃者が正規ユーザのキー入力の正解データを入手できている前提となっている．今後は，攻撃者が攻撃者自身のキー入力を用いて CNN の学習を済ませ，攻撃時にはその CNN を使用して正規ユーザのキー入力を推測するというシナリオでの実験を行っていく必要がある．

6.2 防御策の検討

今回の実験では，攻撃対象のスマートフォンから攻撃者のタブレット端末（録音デバイス）までの距離が違っても，キー入力推測の正解率は大きく変化しなかった．そのため，単純に「攻撃者からの距離をとる」という手段は効果的な防御策とはならない可能性がある．Adversarial Examples などを応用して，CNN の識別を効果的に妨害する環境音を合成するなどの方法を検討する必要があると考えている．

Li らはキーボード打鍵音のキーごとの特徴からキー入力を推測する攻撃手法に対する防御策として，入力環境と入力内容の 2 つの観点から考察を行っている[5]．入力環境の観点からは，盗聴装置が部屋に存在しないこと，部屋の外からも音が傍受可能でないことを確認することが重要である．入力内容の観点からは，単純なパスワード入力だけでなく，ワンタイムパスワードや生体認証との組み合わせによる 2 要素認証が防御になり得るとしている．これらの対策は，今回の攻撃手法に対する防衛策としても有効である．

7 まとめ

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単純に盗聴するという受動的な攻撃シナリオにおいて，正規ユーザのキー入力に関する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．本稿では，本研究の第一歩として，キー入力を PIN 入力に限定して調査を行った．タップ音の音声データのメル周波数ケプストラム係数を画像データとして表現し，畳み込みニューラルネットワークを用いてキー入力の推定を行った結果，高い精度で入力の識別が可能であることが判明した．防音室環境での実験ではあるが，10cm～70cm の範囲においては，攻撃者と攻撃対象の間の距離による推測精度には差が見られなかった．今後は，被検者を増やしながら，現実的な状況を考慮した実験環境でデータを収集し，タップ音によるキー入力の推測精度を更に精査していく．また，防御策についても検討を深めていく．

参考文献

[1] 本間尚文，青木孝文 : 知っておきたいキーワード 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-pp.1576, 2010.
[2] National Securiy Agency: TEMPEST fundamentals, NACSIM 5000, Feb 1982
[3] Ilia Shumailov, Laurent Simon, Jeff Yan and Ross Anderson: Hearing your touch: A new acoustic side channel on smartphones, arXiv : 1903.11137,2019.
[4] Li Lu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Xiangyu Xu, Guangto Xue and Minglu: KeyListener: Inferring Keystrokes on QWERTY Keyboard of Touch Screen through Acoustic Signals, IEEE INFOCOM 2019.
[5] Li Zhuang, Feng Zhou and J. D. Tygar: Keyboard Acoustic Emanations Revisited, ACM Conference on Computer and Communications Security, November 2005, pp. 373-382.
[6] Leon Mak An Sheng, Mok Wei Xiong Edmund: Deep Learning Approach To Accent Classification, CS229 2017.
[7] Audacity: The Free, Cross-Platform Sound Editor, available from
〈 http://audacity.sourceforge.net 〉 (accessed 2019-12-12)
[8]Keras Documentation: Train a simple deep CNN on the CIFAR10 small images dataset(online), available from
〈 https://github.com/keras-team/keras/blob/master/examples/cifar10_cnn.py〉(accessed 2019-12-12)