戦略的創造研究推進事業　CREST
研究領域「人間と情報環境の共生インタラクション基盤技術の創出と展開」
研究課題「VoicePersonae: 声のアイデンティティクローニングと保護」

研究終了報告書

研究期間　2018年10月～2024年03月

研究代表者：山岸　順一
（情報・システム研究機構　国立情報学研究所　教授）

§1 研究実施の概要

(1)実施概要

国立情報学研究所(NII)、アビニョン大学、Eurecom 研究所からなる日仏合同チームで構成される VoicePersonae プロジェクト(2018.10-2024.3)では、以下の4テーマについて同時に研究を実施する構想を立てた。

テーマ1:音声合成・声質変換・音声強調・明瞭性強調等の声のアイデンティティに関する生成モデリング技術の高精度化

テーマ2:音声による生体認証(話者認識)、および、合成音声・変換音声等による偽音声の検知技術(プロジェクト開始後にディープフェイク検知という言葉が社会で定着)の高度化による音声インターフェースの安全性と頑健性の向上

テーマ3:再識別攻撃に頑健な新しい音声プライバシー保護技術の実現

テーマ4:画像や文章等の他のモダリティ情報へ研究成果の適用

本プロジェクトにおける特にユニークな着眼点は、個人特徴の生成モデルと個人認証モデル、個人性の匿名化手法と再識別化攻撃と言った目的が相反する技術をどちらも研究し、加速させることを狙う点である。生成モデルと識別モデルの両方を利用し最適化する敵対的ネットワーク(GAN)になぞらえて、「敵対的研究」と呼んでいる本スタイルはサブグループ内では気が付かない視点や考えをもたらし、新たな研究テーマを生み出すと期待した。

それぞれのテーマに関して以下の様な成果を挙げた。

テーマ1:声のアイデンティティに関する生成モデリング技術の融合と高精度化

声のアイデンティティ、つまり、声の個人性に関する生成タスク内の融合と高精度化に関しては、音声合成と声質変換の両方のタスクを実施可能な生成モデルや、音声合成の発話様式模倣技術を活用した音声強調等の提案を行なった。さらに、生成タスクと話者認識技術の融合も検討し、話者識別の中間表現(話者ベクトル)により生成モデルの出力を制御する方法の提案も行い、また多様な登場人物が登場する落語実演を話者ベクトルを利用した音声合成モデルにより再現する試みも行なった。さらに、信号処理と深層学習を融合させ、信号処理演算を深層学習の演算として行うニューラルボコーダも複数提案した。これらの手法はその後微分可能デジタル信号処理(DDSP)と呼ばれる分野になった。そして、音声合成のネットワーク構造により、ピアノ音やギター音を高品質に再現するシステムへとも発展し、音楽情報処理との融合も可能にした。これに加え、環境雑音に負けないよう音声を変換する音声明瞭性強調でも成果を挙げ、プロジェクト後半には東海道新幹線駅において利用されはじめた。コンペティションも開催し、声質変換に関する Voice Conversion Challenge 2020 と合成音声の主観評価値を予測する VoiceMOS Challenge 2022 を開催した。

テーマ2:音声による生体認証(話者認識)、および、ディープフェイク検知高度化

音声の生成モデルは、個人性の再現性の高さゆえに、悪用された場合にはセキュリティー上の問題を起こす。この様なディープフェイク等のなりすまし攻撃に対する防御モデルに関しても成果を挙げた。まず、ディープフェイク音声検知モデル学習用の大規模音声データベースの構築を行った。本データベースは標準データベースとして広く利用され、現在までに 60 万回もダウンロードされた。さらに、電話越しディープフェイク音声検知を行うシナリオ、圧縮された音声に対してディープフェイク音声検知を行うシナリオのための評価データも公開した。検知アルゴリズムの性能指標も複数提案し、統一評価を行うための ASVspoof チャレンジを 2019 年と 2021 年に世界規模で開催した。そして 50 組織が構築した検知モデルの分析から、高精度な防御に必要な条件を明らかにし、その知見をまとめたガイドラインとオープンソースコードも公開した。加えて、未知手法によるディープフェイクの検知性能を向上させるために有効なデータ拡張法、特徴量、データベース拡張法、および、改ざん領域を推定するローカライゼーション技術も提案した。

テーマ3:再識別攻撃に頑健な新しい音声プライバシー保護技術の実現

ウェブ上の音声から個人特定すること、ディープフェイク音声を作ることも容易になった現在、音

声公開前に話者情報を事前に加工することでプライバシー保護を行う「話者匿名化」が必要である。研究開始当時、どの様に話者匿名化を行うのか、どの様に評価を行うのか、どの様に他手法と比較するのか全く確立されていなかったため、複数の話者匿名化手法の提案、評価指標の提案、話者匿名化手法を相互比較するための学習データベースと評価セット定義を行なった。これに加えて、SIG 設立、チャレンジ開催、特集号企画、ワークショップ開催等を行い、研究分野の創設にも尽力した。本テーマについてもコンペティションを実施し、Voice Privacy Intiative 1 & 2 をそれぞれ 2020 年と 2022 年に開催した。10 数の組織が提案した話者匿名化手法の分析を実施し、各手法の個人識別可能性、ダウンストリームタスクでの有用性、再識別攻撃に対する耐性等をプライバシー保護のレベルが異なる複数のオペレーションポイント毎に解析した。

テーマ4:画像や文章等の他のモダリティ情報へ研究成果の適用

ディープフェイク顔映像検知に関しても顕著な成果を複数挙げた。例えば、カプセルネットワークによるディープフェイク顔映像検知モデル、真贋判定と改ざんされたピクセル領域特定を同時に行うモデル、複数顔画像に対する検知および領域特定を行うモデル、ディープフェイク画像をオリジナルの顔画像へ復元するモデル等を提案した。更に、日本でも Deepfake の被害が発生してきたため、ディープフェイク検知に係る全処理を API 化したプログラム“Synthetiq Vision”を開発し、日本企業数社へライセンスした。サイバーエージェント社では既にサービスの一部として利用中である。これに加え、文章の内容が事実であるかどうかを、文字もしくは表形式の知識データベースと照らし合わせて判断する、自動ファクトチェックの研究も実施した。

(2)顕著な成果

＜優れた基礎研究としての成果＞

1. ディープフェイク音声検知と ASVspoof チャレンジ

概要: 音声の生成モデルは、産業価値をもたらす一方、個人性の再現性の高さゆえに、セキュリティー上の問題も起こす。そこでディープフェイク音声検知モデル構築用の大規模音声データベースの構築を行った(現在 60 万回ダウンロード)。検知アルゴリズムの統一評価を行う ASVspoof チャレンジも世界規模で開催した。また電話符号化や圧縮がある劣悪条件でも頑健に検知を行うモデルを構築するための知見をまとめたガイドラインとプログラムも公開した。

2. 信号処理と深層学習の融合

概要:音の生成モデルではボコーダを利用し、音響特徴量を音声波形に変換する。研究開始当時、信号処理によるボコーダとニューラルネットワークによるボコーダが提案されていたがどちらも問題を抱えていた。そこで、信号処理と深層学習を密に融合させ、信号処理演算を深層学習の演算として行うニューラルボコーダを複数提案した。この様な手法はその後、微分可能デジタル信号処理(DDSP)と呼ばれ、音声処理と音楽処理の融合にも繋がった。

3. ディープフェイク顔映像検知に関する先駆的研究

概要:ディープフェイク顔映像検知技術に関し顕著な成果を複数挙げた。2019 年には、その後 484 回も引用される事になるディープフェイク顔映像検知技術を発表した。同年には、真贋判定と改ざんされたピクセル領域特定を同時に行う技術も発表し、2023 年には IEEE Biometrics Council の 5-Year Highest Impact Award を受賞した。さらに、保護したい画像の顔以外の背景領域に、オリジナル顔情報を予めハイディングしておき、ディープフェイク画像をオリジナルの顔画像へ復元する技術も提案した。

＜科学技術イノベーションに大きく寄与する成果＞

1. 話者匿名化分野の創設

概要:ウェブ上の音声から個人特定すること、ディープフェイク音声を作ることも容易になった現在、音声公開前に話者情報を事前に加工することでプライバシー保護を行う「話者匿名化」が必要である。研究開始当時、どの様に話者匿名化を行うのか、どの様に評価を行うのか、どの様に他手法と比較するのか全く確立されていなかった。そこで話者匿名化手法の提案、評価指

標の提案、話者匿名化手法を相互比較するための学習データベースと評価セット定義、SIG 設立、チャレンジ開催、特集号企画、ワークショップ開催等を行い、研究分野を創設した。

2. **音声明瞭性強調の提案と実社会利用**

概要:聞き手側の環境雑音に負けないよう音声を予め変換する「音声明瞭性強調」は、単純な教師あり学習では解決できない難題である。我々は、微分不可能な音声明瞭性の指標を敵対的生成ネットワークの識別器の出力値と見做し識別器に音声明瞭性指標を近似させ、その近似指標に基づき、不明瞭な音声を聴きやすい音声に自動変換する「iMetricGAN」という方法を提案した。本技術は東海道新幹線駅構内アナウンス強調技術として導入された。

3. **Deepfake 顔映像検知技術と日本企業へ技術還元**

概要:2020 年頃から日本でも Deepfake の被害が発生してきた事を踏まえ、Deepfake 顔映像検知技術の研究成果を社会還元した。我々のディープフェイク検知技術を他社が容易に導入できる様に、全処理を API 化したプログラム Synthetiq Vision を開発し、特許・商標も取得し、日本企業数社へ有償ライセンスした。

＜代表的な論文＞

1. Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi Kinnunen, Ville Vestman, Massimiliano Todisco, Hector Delgado, Md Sahidullah, Junichi Yamagishi, Kong Aik Lee, “ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech” IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 3, no. 2, pp. 252-265, April 2021, doi: 10.1109/TBIOM.2021.3059479.【Impact factor 5.81】

概要:生体認証分野のトップジャーナルである IEEE T-BIOM に採択された本論文では、ASVspoof 2019 チャレンジで提案された 50 種類以上のディープフェイク検知手法の詳細な分析を行ない、高精度な検知モデルに必要な条件を明らかにした。例えば、学習データに含まれない未知の手法や派生手法にも頑健に検出を行うためには、異なる音響特徴量に基づき学習した複数の検知モデルのアンサンブル学習が必須であるなど、有意義な知見を多く示した。

2. Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O’Brien, Anaïs Chanclu, Jean-François Bonastre, Massimiliano Todisco, Mohamed Maouche, “The VoicePrivacy 2020 challenge: Results and findings,” Computer Speech & Language, Volume 74, 101362, July 2022, doi: 10.1016/j.csl.2022.101362.【Impact factor 4.3】

概要:VoicePersonae と仏 Inria 研究所が協力し 2020 年に実施した Voice Privacy Intiative の知見を集約した 40 ページにわたるジャーナル論文。話者匿名化手法を相互比較できる様、音声データベース、評価セット、評価手順を規定し、10 数の大学・企業・研究組織が提案した話者匿名化手法の分析を実施し、各手法の個人識別可能性、ダウンストリームタスクでの有用性、各手法の再識別攻撃に対する耐性などを緻密に解析した。

3. Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, Natalia Tomashenko, “Speaker Anonymization using Orthogonal Householder Neural Network,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, Sept 2023 doi: 10.1109/TASLP.2023.3313429.【Impact factor 4.364】

概要:VoicePrivacy Intiative の取り組みから判明した、話者匿名化変換に必要な複数の条件を満たすように、話者ベクトル空間の変換を行うニューラルネットワークを構築し、最適化する手法を提案したジャーナル論文。英語と中国語実験から有用性確認。与えられた音声データベース全体を話者匿名化する事が可能になり、プライバシフリーデータベース生成に繋がった重要な成果。話者匿名化はメディア報道において活用が期待される。日本放送協会(NHK)が話者匿名化技術をインタビュー音声の匿名化ツールとして導入し、2024 年から実際に利用され始めた。

§2 研究実施体制

(1)研究チームの体制について

① 日本 NII グループ

・声のアイデンティティのモデル化と統合
・他の生体情報におけるディープフェイク検出の研究
・話者認識の安全性と頑健性の向上に関する研究
・音声のプライバシー保護に関する研究

② 仏 Eurecom グループ

・話者認識の安全性と頑健性の向上に関する研究
・音声のプライバシー保護に関する研究

③ 仏 Avignon 大グループ

・音声のプライバシー保護に関する研究
・話者認識の安全性と頑健性の向上に関する研究

(2)国内外の研究者や産業界等との連携によるネットワーク形成の状況について

- End-to-end 音声合成に関し、米 MIT とコラボ
- 声質変換とテキスト音声合成の融合に関して、シンガポール国立大学(NUS) とコラボ
- 信号処理と深層学習を融合したニューラルボコーダに関し、フィンランドアールト大学とコラボ
- ピアノ音の生成モデルに関し、米南カルフォルニア大学とコラボ
- アコスティックギター音の生成モデルに関し、スウェーデン KTH とアールト大学とコラボ
- 声質変換の性能を共通 DB 上で競い合う Voice Conversion Challenge 2020 を、名古屋大、中国科学技術大学、シンガポール国立大学、東フィンランド大と共同運営
- 合成音声の主観評価値を予測する技術およびその性能を共通 DB 上で競い合う Voice MOS challenge に関し、名古屋大、台湾アカデミアシニカとコラボ
- 音声のディープフェイク検知性能を競い合う ASVspoof challenge 2019 の巨大データベースを、フィンランドアールト大、東フィンランド大、台湾アカデミアシニカ、トリニティ・カレッジ・ダブリン、独 DFKI、Google、HOYA、中国 iFlytek、名古屋大、NTT、独ザールラント大、東フィンランド大、中国科学技術大と協力し作成・無償公開
- 上記 ASVspoof challenge 2019 の運営は東フィンランド大、NEC、仏フランス国立情報学自動制御研究所(Inria)と共同運営
- ディープフェイク検知と生体認証の同時学習および統合指標の提案に関して、東フィンランド大および米 MIT とコラボ
- 話者匿名化手法を相互に比較できる様、仏 Inria 研究所、および EU H2020 プロジェクト COMPRISE とコラボし、話者匿名化手法の学習および評価に利用する音声データベース、評価セット、評価手順を共同で策定
- 顔認証の「マスター顔」への脆弱性調査に関して、スイス IDIAP 研究所とコラボ