安心安全なユビキタス社会を構築するための情報セキュリティ技術

【代表者】 黄　緒平　島根大学　総合理工学部　准教授

【研究の目的と内容】

人工知能（AI）技術の進展は、サイバー攻撃を巧妙化・自動化させ、既存の防御技術だけでは対応が難しい課題が残っている。例えば、フィッシングサイトは年々その手口が洗練され、従来のブラックリスト方式や静的な特徴分析では検出ができなくなっている。同様に、深層学習を悪用したディープフェイク技術は、映像や音声の信憑性を根本から揺るがし、個人のプライバシー侵害が課題になっている。更に、AI 技術は自動運転やスマートシティといった次世代の社会基盤においても中核になっているが、標識の誤認識や人間と機械の感度の違いによる標識への攻撃が巧妙化になっている。特に、物理世界に存在する交通標識などへの敵対的標識攻撃は、自動運転システムの誤認識を誘発し、人命に関わる重大な事故を引き起こす可能性が指摘されている。以上の通り、情報セキュリティについて、スマート社会を支えるサーバーインフラにおいても、未知の異常通信を迅速に検知する技術は、安定したサービス提供に不可欠である。

これらの課題に対し、安全で持続可能な社会の実現に貢献することを目指し、本研究では、最新の AI 技術を駆使して、フィッシングサイトやディープフェイクの高精度な自動検出、敵対的攻撃に対する防御技術の確立、サーバーの異常通信検知といった一連の技術開発に取り組んできた。

主に以下の内容で研究を展開し、研鑽を積んできた。

Web の自動巡回と DNS 情報を活用し、大規模言語モデル（LLM）を用いてフィッシングサイトを自動検出する手法を新たに提案した。主に Web サイトのテキストや構造、DNS 情報を分析する手法になる。具体的には、クローラーがウェブサイトを自動的に巡回し、HTML ソース、テキスト内容、URL 構造などを収集する。更に、ドメインの登録情報や IP アドレスなど、DNS レベルの情報を組み合わせることで、攻撃の痕跡を見逃さず検知する。これらの情報を入力データとし、ChatGPT などの大規模言語モデルを解析エンジンに用い、収集された情報からフィッシングの意図や不正なパターンを高度に検出する。従来のシグネチャベースの手法に比べ、未知のフィッシングサイトに対しても、98%以上の精度でフィッシングサイトを判定できる結果が明かになった。

また、音声認識に対する攻撃についての検討も行った。深層学習を用いた音声認識技術（ASR）の普及に伴い、その脆弱性を突く敵対的サンプル（Adversarial Examples）の脅威が顕在化している。特に，音声波形に人間には知覚できない微小な摂動を加えることで、AI モデルに攻撃者が意図した特定のコマンドとして認識させる「標的型攻撃」は、スマートスピーカー等の音声インターフェースにとって深刻な脅威となる。従来の手法（C&W 攻撃など）は強力であるが、最適化の過程で局所解に陥りやすく、生成が困難な場合がある。本研究では、モデルの決定境界を探索的に推定する「境界探索（Boundary Attack）」のアプローチを音声認識に応用する。雑音生成メカニズムの設計により、高品質の標的テキストへの誤認識を誘発可能であることを確認した。

【研究の成果（本研究によって得られた知見、成果、論文、学会発表、外部資金への応募見込み等）】

以下の内容で技術を新たに開発し、実証実験を行い、その成果をまとめ、研究論文の 6 報を国内外の学術会議にて論文公開及び登壇発表を行った。

1)音声認識モデルに対する境界探索を用いた標的型敵対的サンプルの生成に関する検討
森山 修輝，黄緒平，伊藤彰則，情報処理学会研究会論文集(暗号と情報セキュリティシンポジウム SCIS 2026) 2026 年 1 月

2)多言語音声と音韻修復効果を応用した高堅牢性音声 CAPTCHA の提案と評価
森山 修輝，黄 緒平，伊藤 彰則，情報処理学会研究会論文集(暗号と情報セキュリティシンポジウム SCIS 2026) 2026 年 1 月

3)道路標識認識に対する敵対的攻撃による誤認識誘発手法
高橋 征那，黄 緒平，伊藤 彰則，電子情報通信学会技術研究報告（パターン認識・メディア理解 PRMU）vol.125(229),pp. 97-101 2025 年 11 月

4)Web の自動巡回と DNS 情報を利用した大規模言語モデルを用いたフィッシングサイトの検出
竹下 駆，黄 緒平，伊藤 彰則，コンピュータセキュリティシンポジウム 2025 論文集 2025 年 10 月

5)整数ウェーブレット変換と差分拡張に基づく軽量かつ可逆な音声電子透かし方式
黄 緒平，伊藤 彰則，コンピュータセキュリティシンポジウム 2025 論文集 2025 年 10 月

6)A Lightweight and Reversible Audio Watermarking Scheme Based on Integer Wavelet Transform 黄緒平，伊藤彰則，IEEE Proc. of The 17th Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ACS 2025)) 2606(2607) 2025 年 10 月

外部資金への応募について、上記研究内容をまとめ、主に音声信号の信憑性を中心に、研究申請書を作成し、科学研究費への申請 1 件を行った。