アナログ信号による脅威を検知・規制するセキュリティフレームワークの提案と検証

飯島 涼[1,a)]　竹久 達也[2]　森 達哉[1,2]

概要：サイバーフィジカルシステム (CPS) を標的とした，悪意のあるアナログ信号を制御するためのセキュリティフレームワーク「Cyber-Physical Firewall（CPFW)」を開発した．本フレームワークにより，悪意のあるアナログ信号を利用した，CPS 機器に対する攻撃を，ファイアウォールと同様の仕組みで検知・規制することが可能となる．本論文では，CPFW フレームワークの要件・基本仕様を定義した後，フレームワークに基づくプロトタイプを作成した．また，過去の論文で提案されている，3 つの攻撃に対して，プロトタイプを用いてケーススタディを実施し，複数の脅威に対して対策が可能であることを明らかにした．

キーワード：アウトプット，サイバーフィジカルシステム，アナログ信号，音信号

A security framework for detecting and regulating threats caused by analog signals

Ryo Iijima[1,a)]　Tatsuya Takehisa[2]　Tatsuya Mori[1,2]

Abstract: This work developed a new security framework, which provides a generic and flexible access control mechanism for regulating the malicious analog signals that target cyber-physical system (CPS) devices. This framework enables the defeat of various attacks that make use of malicious analog signals against CPS devices; e.g., stealth voice command injection attack using ultrasonic waves or adversarial examples, or attacks to crash drones in flight using malicious sound waves.
Based on relevant previously reported findings, we first defined the requirements and design specifications of the framework. Then, we built a prototype of the framework and demonstrated its feasibility through extensive performance evaluations and case-study experiments using three real-world attacks; DolphinAttack, audio adversarial examples, and WALNUT.

Keywords: Output, Cyber-Physical System, Analog Signal, Audio

1. Introduction

サイバーフィジカルシステム (Cyber Physical System, 以下 CPS) が，物理空間とサイバー空間の相互インタラクションを実現する技術として注目されている．CPS の例として，スマートホーム，自動運転車・ドローンなどの Robotic Vehicles，音声認識を搭載したデバイスなどがある．

CPS の普及が社会に新たな利便性をもらす一方で，CPS に搭載された物理センサに脅威をもたらすセキュリティ攻撃が多数指摘されている [1,8–10]．その脅威は，CPS が持つセンサに不正な入力をもたらし，判断を誤らせるものから，機器そのものの制御を妨害するものまでさまざまである．最新の事例では，IFTTT アプリケーションの命令によって出力されたアナログ信号が他のセンサへの入力として受け取られ，判断のミスを引き起こす脅威 [2]，レーザーの出力に音声信号を変調することにより，音声認識機器に秘密裏にコマンドをインジェクトする脅威 [8] などが報告されている．IFTTT や Voice App などアプリケー

[1] 早稲田大学
[2] 国立研究開発法人　情報通信研究機構
[a)] ryo@nsl.cs.waseda.ac.jp

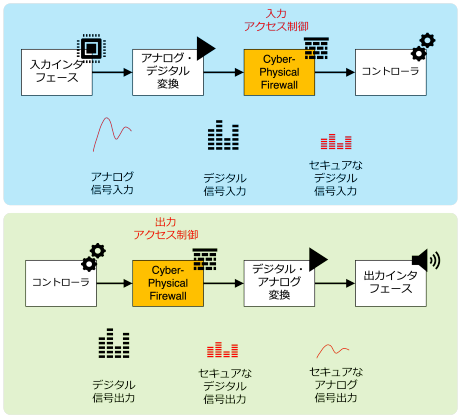

図 1 CPFW フレームワークにおける，CPS 機器の入力側でのアクセス制御 (上) と，出力側でのアクセス制御 (下)

ションベースで CPS のサービスが提供される現在，デバイスの開発者・利用者の意図しない形で CPS の判断・動作が行われる懸念がある．

CPS センサのセキュリティリスクを軽減するため，これまでの研究では，**センサの入力（Input）** を分析することにより攻撃を検知・軽減するアプローチがとられてきた．入力側で行われる対策には，センサの種類によらず，以下に示す問題が存在する．(1) 物理空間上で生じるノイズによって，検知精度が大幅に下がる [7]．(2) コントローラへの入力前に回路の非線形性を利用する攻撃を検知することができない．(3) 攻撃の発生源に対して対策を施すことができない．

上記の問題を解決するため，本研究では，アナログ信号に対するアクセス制御機構を実現するフレームワークとして Cyber-Physical Firewall (CPFW) を提案する．CPFW の特徴は，従来の入力側での対策に加え (Input, 図 1上)，CPS システムの**出力 (Output)** に対するアクセス制御を実現するアプローチを採用する点にある (図 1下)．具体的には，音・光などのアナログ信号によって生じるセキュリティ脅威を，信号の出力前にあらかじめ検知し，規制を行うことを狙いとする．従来の研究がセンサへの入力 (Input) に着目していることに対して，本研究が新たに提示する領域を **Output Security** と定める．出力信号をアクセス制御の対象とすることにより，攻撃の発生源となりうる機器に対して直接に防御メカニズムを埋め込むことにより，根源的な対策を実現できる利点がある．CPFW は，従来の入力側の検知システムを取り込み，入出力の双方でアナログ信号脅威の検知・規制を実現する．

代表的なアクセス制御機構であるファイヤウォールや Android OS のパーミッションシステムはいずれもデジタル信号を対象としている．CPFW のチャレンジは，自由度が高い物理空間において，様々な数値を取りうるアナログ信号を対象として，一般的かつ柔軟なアクセス制御機構を実現することにある．まずはじめに，CPS 機器内で生じる悪性アナログ信号を検知・規制する上で必要となる要求仕様と設計仕様を整理し，任意の攻撃シナリオに対してアクセス制御のポリシーを拡張できるよう，ポリシー記述言語およびグラフィカルインタフェースを開発する．さらに CPFW のプロトタイプを実装し，アナログ信号に対するアクセス制御が適切に実現できること，性能上の問題がないことを実証する．また，CPS デバイスをターゲットとした実世界における具体的な攻撃として，DolphinAttack [10], audio adversarial examples (AEs) [1], WALNUT [9] を対象に，これらの攻撃を防止できることを示す．

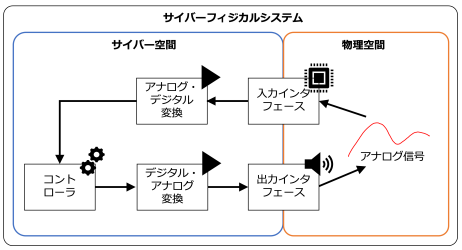

図 2 CPS の概要図

本研究の貢献は以下の通りである．

- アナログ信号向けのアクセス制御機構として，Cyber Physical Firewall (CPFW) を作成し，セキュリティ・プライバシー上脅威となりうるアナログ信号を規制するためのフレームワークを開発した．
- 出力側でのアクセス制御に必要な要素をデザインし，従来の入力側で生じていた検知上の問題（ノイズ，非線形性）などの問題を解決した．
- プロトタイプを実装し，現状提案されている 3 つの脅威: Audio AEs (ノイズ付加), DolphinAttack (超音波攻撃), WALNUT (センサ共振攻撃) が 1 つのフレームワークで対策可能であることを示した．

2. 脅威モデル

2.1 CPS のターゲット領域と具体的な脅威モデル

CPS は，入力インタフェース (マイク・カメラセンサなどセンサ類)，出力インタフェース (スピーカ，ディスプレイ，モータなど)，コントローラの 3 つの要素からなる．図 2に各要素の関係を示す．

入力インタフェースは，物理空間からアナログ信号を読みとり，デジタル化した後にコントローラへ送る役割，出力インタフェースは，コントローラの出力をアナログ信号に変換した信号を受け取り，物理空間へ放出する役割を担う．コントローラは，入力インタフェースから受け取った信号を処理し，処理に応じた出力を出力インタフェースに送り出す役割を持つ．本研究では，コントローラ–クラウド間のセキュリティに関しては対象外としているため，図ではクラウド等との関係を省略している．

本研究では，センサを対象としたアナログ信号の脅威を

扱うことにする．攻撃者が意図して行うボイスコマンド関連のセンサ攻撃の例として，DolphinAttack [10], Light Commander [8], audio adversarial examples (AEs) [1] がある．同様に，市販のスピーカで出せる高周波音によってセンサの内部に共振を起こし，ドローンや自動運転車に危害を加える例などもある [9]．

本研究の最終的なゴールは，上記で示したアナログ信号脅威を規制するためのファイアウォールを作成することである．従来センサセキュリティで行われていた入力側でのアプローチに加えて，出力側で対策をデザインすることにより，市販のデバイスが CPS に与える脅威を予め防ぐことを可能にする．

2.2 攻撃者の仮定

攻撃者は，アナログ信号を放射するための機器として，市販のデバイスを使用するものとする．audio AEs [1]，リプレイ攻撃 [7]， WALNUT [9] など，従来提案されるアナログ信号を用いた攻撃のほとんどは，攻撃の容易性を示すため，市販の機器のみで攻撃が実証可能であることが示されている．

攻撃者が，外部のハードウェア・回路を独自に利用する場合，デバイスの入出力に回路等を接続することは可能だが，直接市販の機器内に回路を埋め込むことはできないものとする．また，CPFW をハードウェアとして実装する場合，CPFW は耐タンパ性を有しており，CPFW の機構そのものを回避することができないものとする．

CPS 上でサードパーティ製アプリケーションがインストールできる場合，攻撃者はアナログ信号を放射するためにマルウェアアプリ等がインストールされた機器を利用することができるものとする．本フレームワークが扱うアナログ信号による脅威は，信号を発生させるデバイスを持つ所有者が意図しているか否かを区別しない．開発者が CPFW を実装できない機器（攻撃者が 1 からデバイスを自作する場合等）については，本研究の対象外とする．

3. CPFW の設計と実装

はじめに，CPFW の実現に必要となる要求仕様 (requirement specification) を示す．次に，要求仕様に基づき，設計仕様 (design specification) および実装方針を導出する．本稿では，CPFW のプロトタイプとして，音信号出力に対するアクセス制御を実装した例を示す．

3.1 要求仕様

1章で示した CPFW の目的を達成するため，以下の要求仕様 (Requirement, **R1**–**R3**) を設定する．

R1: 処理のリアルタイム性

CPFW は，出力での制御を仮定しているため，従来の音・映像等のサービスに支障が生じないよう，リアルタイム性を確保する必要がある．

表 1 一般 Attribute のリスト

クラス	Attribute 名
Base Info	Amplitude
	FFT
	Sampling Rate
Power	Total Power
	Mean Power (MNP)
Frequency	Mean frequency (MNF) [6]
	Median Frequency (MDF) [6]
	Peak Frequency (PKF) [6]
	Variance of Central Frequency (VCF) [6]
	Max Frequency (MAF)
Rate	Threat Frequency Rate (TFR)
	Zero Crossing Rate (ZCR) (Voice part extraction)
Energy	Short Time Energy (STE)

表 2 プリセット規制手法によって防げる脅威のリスト

規制方法	対策可能な脅威
振幅抑制	Replay Attack, Voice Synthesis
	WALNUT [9], Rocking Drones
ローパスフィルタ	DolphinAttack [10], WALNUT [9]
	Rocking Drones
ノイズリダクション	audio adversarial examples [1]
	Hidden Voice Commands, jamming

R2: アクセス制御の柔軟性

アナログ信号の攻撃は，音信号が加速度センサの共振攻撃に利用されるなど，直接入出力に関係しないアナログ信号がセンサにとって脅威となる可能性がある．脅威となる攻撃・アナログ信号の種類を問わず，柔軟にアクセス制御を行えるフレームワークが必要となる．

R3: 拡張可能性

必要なポリシーを事前に定めて実装し，予めセキュリティを確保したセキュリティ・バイ・デザインが実現していることが理想であるが，デバイスがユーザの手に届いてから脅威が発見されるケースが数多く報告されている．このことから，CPFW は，新たな脅威が発生しても，制御方法をその脅威の特性に合わせて拡張できるように設計されているべきである．

3.2 設計仕様

次に，前述の要求仕様をもとに，設計仕様 (**D1**–**D4**) と，それぞれの実装方針を示す．実装方針では，先行研究との差分となる，出力側で脅威となるアナログ信号を防ぐための実装に絞って話を進める．実際に作成したプロトタイプ実装は，音信号出力を対象として作成している．実装したプロトタイプによって防げる攻撃の種類を 表 2に示している．

D1: ブロック単位でのデータ処理

R1 を満たすため，CPFW で行われる処理はすべて N 個のブロック単位で行われるものとする．N は，機器のス

ペックに応じて，リアルタイム性を満たすように設定するパラメータである．また，周波数領域での解析を可能とするため，N は，FFT で利用される 2 の指数乗の値を利用する．例えば，音信号を Raspberry Pi 3B+上で扱う場合，サンプリング周波数を $f_s = 48$ kHz, または 96 kHz，量子化ビット数を $q = 16$ bit としたとき，$N = 1024, 2048, 4096$ でリアルタイムに動作することが確認されている．

D2: attribute デザイン・抽出処理

R1, **R2**, and **R3** を満たすため，N 単位のブロックから検知に必要な情報を抽出する機構が必要となる．アナログ信号から抽出可能な情報を attribute と呼ぶことにする．

N 単位のブロックからは，基本的な統計量 (平均，分散，中央値，最大値，最長値等) を時系列データに適用した数値，FFT によって得られたスペクトラムに対して統計量を定めた数値，パワーに対して統計量を適用した数値を attribute として取得することができる．表 1に，すべてのアナログ信号に適用可能な一般 attribute の表を示す．一般 attribute は，アナログ信号の種類を問わず利用することができる．

ここで，threat frequency rate (TFR) は，脅威信号とみなされる周波数領域のパワーの割合である．脅威信号の領域を $[f_L, f_H]$ Hz, $X(n)$ を n 番目の FFT の結果としたとき, TFR は,

$$TFR(f_{DL}, f_{DH}) = \sum_{n=F^{-1}(f_{DL})}^{F^{-1}(f_{DH})} X(n) \Big/ \sum_{n=0}^{N} X(n), \quad (1)$$

と計算される．$F^{-1}(f)$ は，周波数の値 f から，該当する配列番号を取得する関数である．

アクセス制御のポリシーは，これらの attribute の候補と組み合わせて定義する．例えば，超音波を用いるボイズコマンド攻撃を防ぐポリシーを定める場合,「mean frequency が 20 kHz を超えていたら，ポリシー違反とみなす」のようにする．ポリシーの例を表 3に示す．超音波を用いてセンサに共振を生じさせる攻撃の場合も同様の手法で定義が可能である．

統計的な数値に加えて，アナログ信号の種類に応じて高度なデータ処理手法を attribute として利用する．例えば，音声信号の場合，音声認識結果として得られたテキストデータを attribute として利用できる．この attribute によって，音声攻撃の起点として必要な “Ok Google” or “Alexa.” のような起動音声を，検知することができる．また，プライバシーに関する単語や NG ワードに該当する単語を規制する役割を果たすこともできる．音声認識結果として，FFT と同じ N フレーム単位で処理を渡して結果を受け取ることができる Google text-to-speech を利用した．

D3: ポリシーベースアクセス制御

R2 を満たすため，ポリシーベースのアクセス制御を採用する．ポリシーベースのアクセス制御は，ネットワークファイアウォールや，Android パーミッションシステムに利用されている．従来のアクセス制御に習って，*if–then* ルールを採用する． ポリシー違反を *if* 構文によって検知し，ポリシーに基づく規制を *then* 構文によって作動させる形で実装を行う．

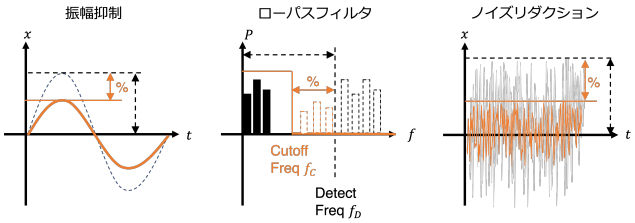

図 3 プリセット規制方法の概要図．%は規制の強さの割合を示す．

規制の方法は柔軟に決定できるものとするが，本論文では，アナログ信号の規制手法として，代表的な 3 つの手法: 振幅減衰, ローパスフィルタ，ノイズ除去について説明・実験を行う．図 3に，規制方法のイメージを示す．

規制の強さを柔軟に決定できるようにするため，規制の強さを定める Enforcement level を 0–1 の間で設定できるようにする．ここで，0 は全く規制を行わない場合，1 は完全に脅威信号を出力しないものとする．開発者やユーザは，CPS が提供するサービスに支障が生じないように，自由に Enforcement Level を設定することができる．

これらの規制方法によって，防ぐことが可能な脅威のリストを表 2に示している．

D4: ポリシー記述インタフェース

R2，**R3** を満たすため，ポリシーを実際の実装に落とし込むためのポリシー記述用インタフェースを導入する．具体的には，GNU Radio, MATLAB Simulink のように，ブロックで記述可能な GUI ベースのインタフェースを採用し，これをアナログ・ポリシー・ダイアグラム (Analog Policy Diagram, APD) と呼ぶことにする．ポリシー記述をブロックベースで行うことにより，ポリシーの拡張や再利用，共有が行いやすくなるほか，アナログ信号の処理に精通していない開発者でも，直感的にポリシーを定めることができる．

APD の主な構成要素を，図 4上，APD を利用したダイアグラムによって mean frequency を取得する例を 4に示す．APD は，内部言語として定義される JSON 形式の中間言語 (Analog Policy Language, APL) と対応しており，APL がシステムに読み込まれることによってシステム内で利用可能な言語に変換される．図 5に，ポリシー記述，APD, APL 及び実際に処理を行うシステムとの関係を示す．

本章のまとめとして，図 6に CPFW の全体像を示す．これまでに示してきたデザインに従って，CPFW は，N フレームごとに分割されたブロック単位でデータを受け取り (D1)，attribute を取得する (D2)．どのような attribute を用いて検知を行い，規制をするかをポリシー記述言語

表 3　音信号を出力側で検知する場合のポリシー例

カテゴリ	脅威クラス	ポリシー文	検知の条件例
音声	超音波 ノイズ リプレイ	20 kHz 以上の音波出力を制限する． 0.1 未満の振幅を持つ出力を制限する． Voice command を含む音波を制限する．	getMNF(Nframes) $\geq$ 20000 getAmplitude(Nframes) $\geq$ 0.1 “OK Google” in [RecognitionText]
プライバシー情報	情報	声を含む音波を制限する． 声を含む音波情報をメッセージカードのみの配信に制限する．	getZCR(Nframes) $\geq$ 0.2 getZCR(Nframes) $\geq$ 0.2
センサ	共振	共振が確認される周波数より上の音波を制限する．	getMDF(Nframes) $\geq$ 4000

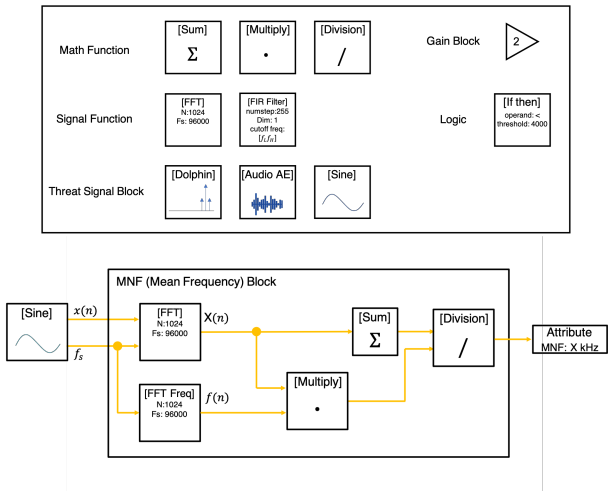

図 4　APD の構成要素 (上図) と，それを利用したダイアグラム構成例 (下図)．f_s はサンプリング周波数，$f(n)$ は n 番目のフレームの周波数を表す．

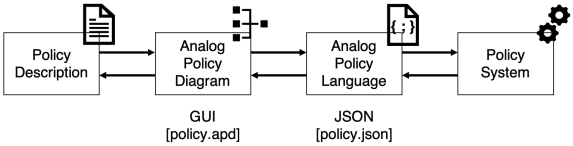

図 5　ポリシー文 (Policy Description) と，ダイアグラム言語 (APD)，中間言語 (APL), 及びシステムとの関係図

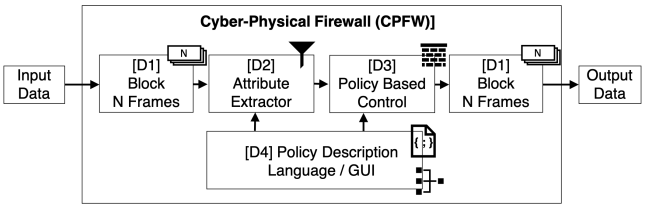

図 6　CPFW の全体像．D1–D4 は 3.2節と対応．

(APD) によって記述し (D4)，その記述に従って違反した信号の規制を行う (D3)．処理が行われた後の信号も，N フレーム単位でリアルタイムに出力が行われる (D1)．

4.　CPFW の基礎性能評価

本章では，CPFW の有効性を確認するために，リアルタイム性の評価と，attribute 取得機構のテストを行う．

4.1　リアルタイム性評価

はじめに，CPFW を実装することによって生じるオーバーヘッドを計測し，対策のリアルタイム性を評価する (**R1**)．パフォーマンス測定は，(1) attribute 取得にかかる時間，(2) 規制手法ごとにかかる処理時間，(3)End-to-end の測定として， attribute 取得から，ポリシーを検知し，規制が完了するまでの時間，の 3 つの観点で行う．(1), (2) の処理時間の測定は，1 分間の間に 0–30 kHz を推移するチャープ信号を，(3) の処理時間の測定は，20kHz 以上の超音波がランダムに生成された 1 分間の信号を利用し，各 N フレームごとに要した時間を記録した．(3) の処理時間の測定には，超音波を含む音を規制するポリシーを適用し，規制方法はローパスフィルタを用いた．

表 4に，(1) の結果として，それぞれの attribute 取得に要した平均時間とその標準偏差，表 5に，(2) の結果として，各規制手法ごとに要した平均時間とその標準偏差を示す．(3) の End-to-end の測定では，平均時間が 5.42 ms, 最大所要時間が 26.34 ms となった．ITU-T が定める音声転送時間の許容遅延の定義によれば [4]，0–150 ms の遅延はほとんどのユーザにとって許容範囲となることが示されている．このことから，CPFW の処理のオーバーヘッドは，極めて小さいと判断できる．

表 4　ブロック単位を $N = 1024$ フレームとした場合の attribute の処理時間計測．

attribute	平均 [ms]	標準偏差
Amplitude	0.121	0.019
Sampling Rate	0.057	0.013
Zero Crossing Rate	0.260	0.048
FFT	1.258	0.240
Mean Frequency	1.385	0.271
Median Frequency	2.677	0.451
Threat Frequency Rate	2.192	0.452

表 5　ブロック単位を $N = 1024$ フレームとした場合の規制方法の処理時間計測．

規制方法	平均 [ms]	標準偏差
振幅抑制	1.694	0.082
ローパスフィルタ	3.219	0.140
ノイズリダクション	8.471	0.689

4.2　Attribute の有効性の確認

CPFW によって取得される attirbute が，脅威の検知に有効であるかを確認する．本評価の目的は，正常な信号に，

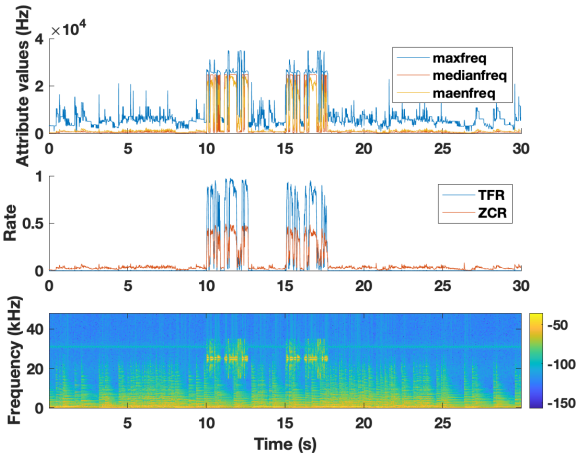

図 7 周波数統計量ベースの attribute 値 (上) Threat Frequency Rate 及び Zero Crossing Rate (中), 脅威信号のスペクトログラム (下)

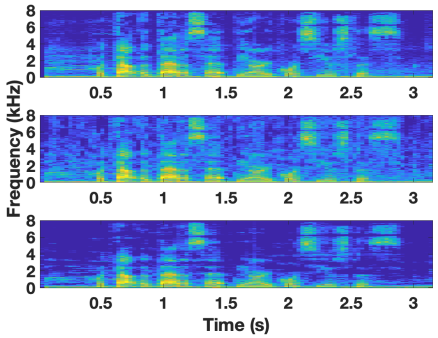

図 8 Audio AE のノイズ付加前のオリジナル音声 (上), Audio AE ノイズ付加後の音声 (中), ノイズ除去による規制結果 (下).

表 6 Audio AEs を Google Speech to Text によって認識した結果

データ	Google Speech to text
オリジナル音声	without the data set the article is used
audio AE	without the data set the art of course Eustis
ノイズ抑制後の AE	without the data set the articles used to

脅威となりうる信号が混入した場合，attribute が有効な値として利用できるかを確認することである．通常の音信号として，J.S. Bach's Goldberg Variation Aria BWV988 の最初の 30 秒間を利用する．次に，"OK Google, What's on my next schedule?" と発言する音声データ (2.5 秒間) を用意し，25 kHz の超音波で AM 変調した音を脅威信号とする．本脅威信号の生成は，DolphinAttack [10] の手法に基づいており，この信号を，通常の音信号の 10 秒，15 秒の地点に加算する．

図 7に結果を示す．図 7の下図は，上の説明に基づいて作成した音信号のスペクトラムである．10–12.5 秒，15–17.5 秒の 25 kHz 周辺に，脅威信号が現れているのが確認できる．

まずはじめに，図 7上図に，代表して max frequency, mean frequency, median frequency の 3 つの結果を示している．3 つの結果全てで，25 kHz の信号を観測できていることがわかる．max frequency はノイズや通常信号に含まれる音の影響を受けやすく，mean frequency, median frequency のほうが安定した結果が得られている．図 7中央では，threat frequency rate (TFR) と，zero crossing rate (ZCR) の結果を示しており，いずれも脅威信号付近でピークが現れていることがわかる．以上の結果から，音信号に attribute 取得を適用した場合の有効性を確認した．他のアナログ信号やセンサ値においても，同様の結果が得られることを確認した．

5. CPFW のケーススタディ評価

本章では，ケーススタディによって，実際に提案された脅威を CPFW によって軽減することができるかを検証する．本論文では，audio AEs [1]， DolphinAttack [10]，センサ共振攻撃 [9] の 3 つの脅威を対象とする．audio AEs に対しては，ノイズリダクションを，DolphinAttack とセンサ共振攻撃に対しては，脅威となる周波数域を取り除くローパスフィルタを用いて規制を行った．

以下の実験では，CPFW を出力側で実装し，デバイスからアナログ信号が生じる前に防ぐことができているかを基準として評価を行う．

5.1 ノイズリダクションによる audio AEs の規制

本実験では，audio AE として生成された信号に対して，ノイズリダクションを適用した結果を示す．audio AEs の攻撃成功率が高い状態を保つため，audio AEs の生成から CPFW までの適用まではソフトウェア上で完結させた状態でシミュレーション実験を行う．Audio AEs の信号として，Carlini らが提供する音声信号のデータセットを用いた [1]．データセット上で，付加されたノイズの量を解析すると，平均振幅は，4.5×10^{-4} で，どの周波数領域も同程度に現れることがわかった．この結果から，Audio AEs に付加されるノイズの特徴はホワイトノイズと似た特徴を持つと判断できる．4.5×10^{-4} 程度のホワイトノイズの除去を行うためのリファレンス信号を用意し，Audio AEs に対してノイズリダクションを実行した．

図 8 の上図に元の音声信号，中央に audio AEs の信号，下図にポリシー規制を行った後の波形を示す．下図を見ると，中央図に現れていた，AE 特有のランダムなノイズが取り除かれていることがわかる．表 6 に，それぞれの波形を Google Speech to Text に適用した結果を示す．ノイズを除去した信号の認識結果が，ほとんど元の音声信号の結果と一致していることがわかる．本実験によって，CPFW が Audio AEs の対策に有効であることを示した．

5.2 DolphinAttack の規制

DolphinAttack を防ぐため，20 kHz 以上の信号をローパ

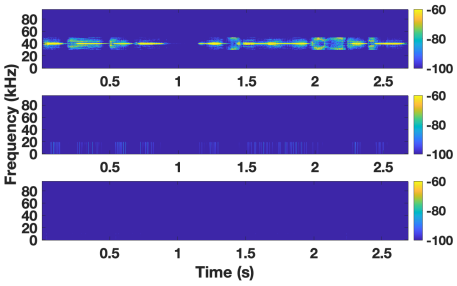

図 9 DolphinAttack の攻撃信号のスペクトログラム (上), 物理空間に放射する前の規制信号 (中), 空間中に放射された規制信号を録音したスペクトログラム (下).

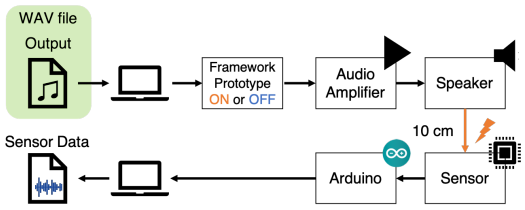

図 10 センサ共振攻撃の再現実験セットアップ

スフィルタで規制する場合の結果を示す．DolphinAttack は，実際の物理空間上で，空気中を通して放射することで脅威信号が生成されるため，空気中を通した音波をマイクで受け取ることにより検証する．DolphinAttack の録音は，防音性に優れた音響室で，マイクと攻撃用スピーカが 50 cm 離れた状態で行う．

DolphinAttack の音波は，音声認識上で “OK Google, what’s on my next schedule” と認識される音声データを，40 kHz の超音波で AM 変調することによって生成した．AM 変調された信号が非線形性を持つ回路に入力されることで，可聴域に信号を生じさせる攻撃となる．

図 9に観測された信号を示す．上図が元の攻撃信号，中央図が，ローパスフィルタによって規制された信号，下図が，実際に放射された後，大気中で録音されたスペクトログラムとなる (信号の減衰を明らかにするため，カラーバーの表示を揃えている)．中央図では，わずかにノイズが現れるものの，ポリシーによって規制を行った後，DolphinAttack のために生成した信号は観測されなくなっていることがわかる．CPFW の規制適用によって，出力された音信号が音声認識機器に対して放射された音として検知されることがなくなることがわかる．本実験によって，CPFW が DolphinAttack の対策に有効であることを示した．

5.3 センサ共振攻撃の規制

本実験では，音を規制する CPFW が，マイクや音声認識機器だけではなく，センサに対する脅威にも対応可能であることを示す．音信号を用いてセンサに共振を生じさせる攻撃のうち，市販のスピーカでの脅威が確認されている (最も容易に模倣される可能性がある) 加速度センサ共振攻撃 WALNUT [9] の対策について検証を行う．WALNUT によって，加速度センサに制御を頼るドローンや自動運転車などの機器の操作を狂わせる恐れがある．

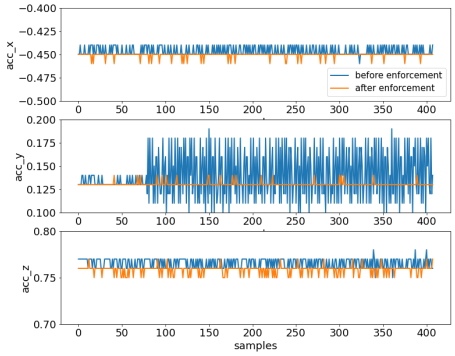

図 11 5,650 Hz の音信号による共振攻撃の結果 (青) と，規制後のセンサ値 (オレンジ).

実験セットアップを図 10に示す．まずはじめに，WALNUT で行われる共振攻撃を再現するため，加速度センサ・ジャイロセンサ・磁気センサを含む MPU9250 を用意する．アンプとつないだスピーカを用意し，MPU9250 の 10 cm 上にセットする．スピーカからは，50 Hz–30 kHz の間のサイン波を，50 Hz のインターバルでそれぞれ放射する．先行研究に示された実験セットアップを参考に，音圧レベルは平均して 100 dB に調節する [9]．音波が放射された状態のセンサ値を記録する．MPU9250 のセンサ値は，arduino を通じて取得する．

上の手続きによって，手元にあるセンサの共振周波数を明らかにし，その共振周波数をターゲットにして検証を行う．観測された共振周波数は，5.2–5.8, 14.0–14.1, 20.25–20.6, 21.3–21.95, and 22.15–22.6 kHz であった．それぞれの共振周波数について，CPFW を適用した場合，適用していない場合の波形を取得し，結果を比較する．4 kHz 以上の音波を検知した場合，ローパスフィルタによって規制するポリシーを適用する．

図 11に， 共振周波数 5650 Hz の場合の結果を示す. 青線が，規制を行う前，オレンジ線が規制を行った後の結果を示している．この結果では，Y 軸のセンサ値が共振を起こしている．オレンジ線を見ると，青線で共振していた Y 軸の値が制限されていることがわかる．センサ値は常に 0.127 付近を遷移しており，安定した結果が得られた．本実験によって，CPFW がセンサ共振攻撃の対策にも有効であることを示した．

6. 議論

6.1 音以外のアナログ信号の評価

3章では，CPFW をどのアナログ信号に対しても適用可能な形で定義した．本研究では，アナログ信号の中でも，提示されている脅威の深刻度や種類の多さから，音信号を対象にして実験を行った．他のアナログ信号に対する詳細

の評価は今後の課題とする．

また，アナログ信号の生成に必要なアクチュエータでは，アナログ信号を生成する際に二次的な影響を生じさせることがある．例えば，音や光を生成するアクチュエータからは，熱や振動など，本来の出力から想定していない出力が生じ，センサなどに影響を与える可能性がある．アクチュエータから生じる二次的な影響は，機器ごとに異なる特徴を有するため，システム内部からその出力を予想することが難しく，機械工学的なアプローチが求められる．二次的な影響によるセキュリティ脅威の検証については，今後の課題とする．

6.2 異常検知と機械学習モデル

本論文では，フレームワークの有効性を示す例として，基本的な統計量を用いて検知を行った．本研究で示したattribute 以外にも，異常検知や機械学習モデルで用いられるスコアをポリシー言語に従って記述することで，さらに複雑な条件（例えば特定の異常波形を対象としたパターンマッチングなど）を用いたアクセス制御が可能となる．検知に用いる attribute の複雑さと，リアルタイム性のトレードオフの評価は今後の課題である．

6.3 ポリシー共有システムの提案

本研究では，ポリシーの例をいくつか提示し，提示したポリシーに従った検知・規制を実施した．3章で示したダイアグラム言語は，新しくポリシーを定めることが可能であると同時に，言語化することによって共有・再利用することが容易となる．将来の展望として，IDS 上で不正アクセスのルールを記述し，共有する Snort のシステムや，Arduino 上で作成したコード類を広く Arduino Libraries などのシステムのように，各開発者が作成したポリシーを共有できるシステムを開発することで，CPS のアナログ信号・センサ周辺のセキュリティ技術を容易に実装できる．

7. 関連研究

脅威となりうるアナログ信号を入力側で防ぐしくみとして，Giechaskiel らは，不正なアナログ信号の入力を防ぐフレームワークを開発している [3]．また，Ding らは，IoT 機器間で生じる物理的な相互作用を発見するためのフレームワークを提案した [2] ．先行研究の多くが入力側での対策に注力しているのに対して，本研究では，出力側にもアクセス制御を追加し，ファイアウォールの考え方をアナログ信号・センサセキュリティの分野に適用したことに違いがある．フレームワークの設計にあたっては，AR デバイスの出力のセキュリティを保つために作成された Arya OS [5] のデザインを参考にした．

8. まとめ

本研究では，柔軟に脅威アナログ信号を検知・規制するフレームワークとして Cyber-Physical Fiwewall を開発した．本フレームワークの独自性は，従来の入力側での制御に加え，出力側で，予め生じうる脅威を検知・規制する点にある．この工夫によって，センサの入力後に生じていたため検知が困難であった脅威（例えば，回路の非線形性によって生じる脅威信号を発生させるなど）に対して，脅威が生じる前に対策を行うことが可能となる．フレームワークデザインに加えて，実際に音信号を脅威とした場合のプロトタイプを作成し，リアルタイム性，検知の有効性を確認した後，先行研究で想定される 3 つのセンサ攻撃として，DolphinAttack (超音波攻撃)，audio adversarial example，センサ共振攻撃の再現を行い，ケーススタディとして CPFW による対策の有効性を示した．ケーススタディによって，音信号による脅威から，マイク・加速度センサなど，複数のセンサを同時に対策することができることを示した．CPFW フレームワークを用いた，更に複雑なアクセス制御の実現と評価は今後の課題である．

謝辞 本研究の一部は JSPS 科研費 18K19789，および 19H04111 の助成を受けたものです．

参考文献

[1] Carlini, N. and Wagner, D. A.: Audio Adversarial Examples: Targeted Attacks on Speech-to-Text, *CoRR* (2018).

[2] Ding, W. and Hu, H.: On the Safety of IoT Device Physical Interaction Control, *The 25th ACM SIGSAC Conference on CCS*, pp. 832–846 (2018).

[3] Giechaskiel, I. et al.: A Framework for Evaluating Security in the Presence of Signal Injection Attacks, *Computer Security - ESORICS 2019*, Vol. 11735, Springer, pp. 512–532 (2019).

[4] ITU-T: G.114 : One-way transmission time (2003). `https://www.itu.int/rec/T-REC-G.114-200305-I/en`.

[5] Lebeck, K., Ruth, K., Kohno, T. and Roesner, F.: Arya: Operating System Support for Securely Augmenting Reality, *IEEE Secur. Priv.*, Vol. 16, No. 1, pp. 44–53 (2018).

[6] Phinyomark, A. et al.: The Usefulness of Mean and Median Frequencies in Electromyography Analysis, *Computational Intelligence in Electromyography Analysis* (Naik, G. R., ed.), IntechOpen, Rijeka, chapter 8 (2012).

[7] Sahidullah, M. et al.: Introduction to Voice Presentation Attack Detection and Recent Advances, *Handbook of Biometric Anti-Spoofing - Presentation Attack Detection*, 2nd ed. edition, pp. 321–361 (2019).

[8] Sugawara, T. et al.: Light Commands: Laser-Based Audio Injection Attacks on Voice-Controllable Systems, *the 29th USENIX Security Symposium* (2020).

[9] Trippel, T. et al.: WALNUT: Waging Doubt on the Integrity of MEMS Accelerometers with Acoustic Injection Attacks, *2017 IEEE EuroS&P*, pp. 3–18 (2017).

[10] Zhang, G. et al.: DolphinAttack: Inaudible Voice Commands, *the 24 th ACM SIGSAC Conference on CCS*, pp. 103–117 (2017).