CL-003

IoT セキュリティと不快感の緩和を考慮したテレビ視聴ロボットにおける安全・安心対策
Safety and Security Measures for the TV-watching Robot Considering IoT Security and Mitigation of Discomfort

村﨑 康博† 星 祐太† 萩尾 勇太† 上村 真利奈† 金子 豊† 山本 正男†
Yasuhiro Murasaki Yuta Hoshi Yuta Hagio Marina Kamimura Yutaka Kaneko Masao Yamamoto

1. はじめに

気軽な会話をしながら親しい人とテレビを視聴する（以下，共時視聴）ことは，かつてテレビが一家に一台しかなかった時代から，携帯端末で一人一人視聴できるようになった現在においても，日常的なテレビの楽しみ方の一つである[1]．我々はこの共時視聴を人間に代わってロボットが一緒のテレビを視聴するテレビ視聴ロボットで再現することを検討している．

テレビ視聴ロボット（以下，ロボットとする）は視聴者が見ているテレビ番組の映像と音声および字幕情報等から特徴を示すキーワードを抽出し，それをもとにロボットの発話文を生成する．ロボットは発話文に従い，視聴者に対して感情語を含んだ言葉を発したり，問いかけたりしながら対話するしくみである[2][3]．

このロボットは，キーワード抽出の一部に商用クラウドの映像音声認識サービスを利用している．そのためロボットもインターネット等のネットワークに接続されている IoT（Internet of Things）デバイスの 1 つであるとみなせ，情報漏えいや外部攻撃に対するセキュリティ対策は欠かせない．また IoT デバイスのセキュリティ対策は，運用中に実施することは難しいため，設計段階から考慮に入れる Security by Design の考え方が求められる[4]．

したがってロボットを安全に利用できるように設計・実装し，ユーザに安心して利用してもらえるような対策が必要である．そこでロボットにおける安全・安心対策に取り組みとして「ユーザの個人情報保護に基づいた情報セキュリティ対策」と「ロボットが起こす不適切な行動により，ユーザに与える不快感の緩和への対応」について，先行調査とシステム試作および動作確認を行った．

情報セキュリティ対策では IoT 推進コンソーシアムが総務省と経済産業省と連名で 2016 年に策定した「IoT セキュリティガイドライン」を利用し，特に IoT 設計時の指針である「守るべきものを守る設計を考える（当該ガイドライン 2.3 節参照）」を参考にして開発時におけるセキュリティ設計要件をまとめ実装した．不適切な行動の防止に向けては不快を与える「不快用語」の個別管理システムの試作と動作検証を行った．

2. ロボットの安全・安心対策

2.1 外部クラウドインターフェース

ロボットにおけるクラウドの認識サービス処理の利用は，外部クラウドインターフェース（以下，クラウド IF）が担っている[5]．図 1 はクラウド IF の基本概要図である．クラウド IF の仕組みは次の通りである．まずテレビ視聴ロボットに搭載したカメラ・マイクで視聴している映像音声を収録し，クラウド IF に送信する．クラウド IF では，映像は一定時間間隔にサンプリングした画像を，音声はストリーミング方式で連続的に，クラウドへ送信する．クラウドでは，送られた映像音声をそれぞれ画像認識用・音声認識用の API によって認識処理を行い，それぞれの処理で得られたキーワードをクラウド IF に返信する．クラウド IF では受信したキーワードを指定時間間隔で集計し，ロボット側が実装している発話生成部と動作生成部に送信する．

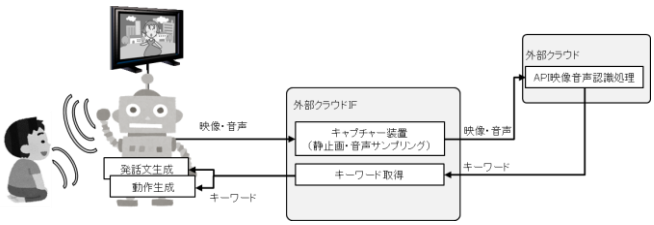

図 1　外部クラウドインターフェース基本概要

2.1.1 入力データ（映像・音声）

クラウド IF の入力データは，ロボットに搭載することを想定したデオカメラから出力される映像音声信号（SDI，HDMI 等）である．これをキャプチャーし，一定間隔でサンプリングすることで，キーワード等特徴量抽出の対象となる静止画データ・音声データ，動画データを生成している．

2.1.2 出力（キーワード）

クラウド API による認識処理によって抽出したキーワードは日本語で出力する．そのため認識結果のキーワードが英語の場合は，クラウドの翻訳 API により翻訳されて出力している，その際，翻訳結果が英日辞書としてクラウド IF 内で保持され，翻訳結果を修正したい場合には，当該辞書にて手動で修正することもできる．なおクラウド IF から出力されるデータはロボット以外でも汎用的に利用できるような形式（JSON ファイル，CSV ファイル，XML ファイルなど）で出力している．

キーワードを含むクラウド IF から出力するデータは以下とした．丸数字が付与されている時刻は，図 2 中に示す各処理を実行した時刻である．

- 抽出サービス名（クラウドの認識処理 API）
- コンテンツ ID（番組名など）
- コンテンツ開始時刻（オンエア時刻，ビデオ撮影時刻，デモ開始時刻）①
- コンテンツ転送開始時刻（クラウド IF からクラウドへの転送開始時刻）②
- コンテンツ処理開始時刻（クラウド内処理開始時刻）③
- キーワード：クラウドの認識 API によって提供される，抽出されたキーワードへのスコア（評価得点）
- コンテンツ処理完了時刻（キーワード抽出時刻）④

† NHK 放送技術研究所
NHK Science & Technology Research Laboratories

- 出力データ転送完了時刻(クラウドからクラウド IF への転送完了時刻)⑤

2.1.3 映像音声認識処理 API

認識技術は既存のクラウドサービスが提供する画像認識および音声認識を用いている．複数のクラウドサービスを活用し，クラウド IF から送信された映像音声からキーワードを抽出する処理を実行している．表 1 は本研究で使用した認識処理 API である．

表 1　認識処理 API 一覧

	サービス名	目的	実装機能
A	AWS Rekognition Image	画像分析	・対象物体，シーン，アクティビティ検出
B	AWS Rekognition Video	動画分析	・対象物体，シーン，アクティビティ検出 ・有名人の認識
C	AWS Elastic Transcoder	メディア変換	・メディアファイル形式の変換 ・サムネイル画像の生成
D	Azure Computer Vision	画像分析	・対象物体，シーン，アクティビティ検出 ・文字の認識
E	Azure Speech Service	音声分析	・音声をテキストに変換
F	Azure Translator Text	翻訳	・英語から日本語の翻訳

なお認識 API（A～F）それぞれの公開されている仕様（ドキュメント）は表 A（文末付録）を参照されたい．

次に認識処理過程の流れを図 2 に示す．認識処理は次の Step1 から Step5 に従って実行される．図中の(A)から(E)は表 1 の API を指し，丸数字は 2.1.2 項の該当する時刻の測定場所を示す．

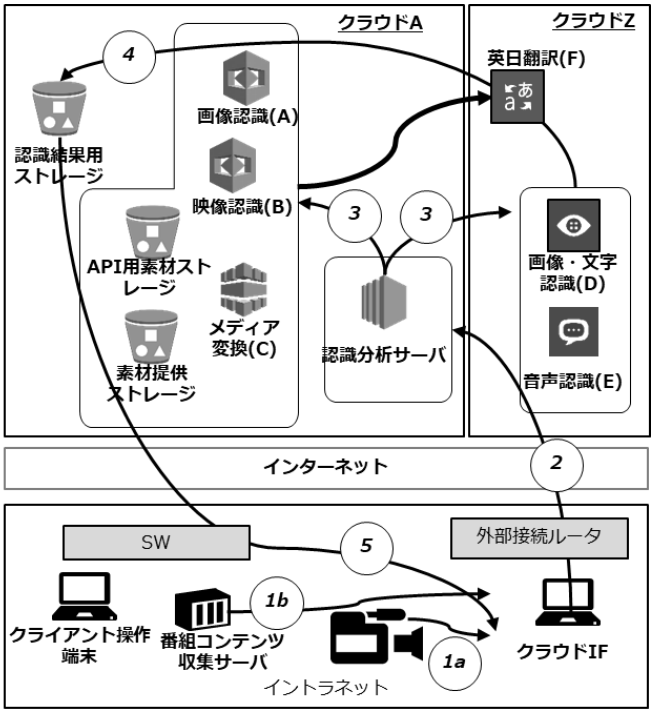

図 2　クラウド IF・クラウドでの
認識処理フローの概要

- Step1a:ロボットがテレビを視聴している状態を想定し，カメラで収録した映像音声をクラウド IF に送信
- Step1b:(もしくは)ロボットがテレビチューナーを内蔵している状態を想定し，番組コンテンツを収集しているサーバから映像音声をクラウド IF に送信
- Step2.:クラウド IF からサンプリングされた画像音声データをクラウドの認識分析サーバに送信．認識分析サーバでは受信した画像音声データを，画像分析・音声分析をそれぞれ担う画像・音声認識 API に送信．
- Step3:クラウドの画像・音声認識 API を利用し画像・音声解析を実施
- Step4:画像・音声データそれぞれの解析結果をクラウド上の認識結果用ストレージに格納
- Step5:クラウド上の認識結果用ストレージから解析結果をクラウド IF に送信

なお本研究でのクラウド IF は，ロボットとクラウドとの中間に位置する PC で実装した．

2.2 安全・安心対策のクラウド IF への適用

本稿ではロボットに安全・安心対策を講じるにあたり，国内外の主なガイドラインを「IoT セキュリティ関連」と「AI 開発・利用関連」の 2 つに分けてまとめた[6]．これはコミュニケーションロボットがネットワーク（インターネット）につながるデバイス（すなわち IoT デバイス）としての要素と，学習機能を搭載した AI デバイスとしての要素を兼ね備えており，安心・安全対策は 2 方面から検討することが必要と考えたからである．表 2 に主なガイドラインを示す．

表 2　コミュニケーションロボットに関わる
開発ガイドライン

		ガイドライン名	策定元
IoTセキュリティ関連	1	IoT セキュリティガイドライン ver1.0	IoT 推進コンソーシアム，総務省，経済産業省
	2	IoT 開発におけるセキュリティ設計の手引き	独立行政法人情報処理推進機構(IPA)
	3	GSMA IoT セキュリティ・ガイドライン	GSM Association (GSMA)
	4	CCDS 製品分野別セキュリティガイドライン v2.0	重要生活構器連携セキュリティ協議会
	5	IoT Security Guidance	Open Web Application Security Project (OWASP)
AI開発・利用関連	6	AI 利活用ガイドライン	AI ネットワーク社会推進会議(総務省)
	7	国際的な議論のための AI 開発ガイドライン案	AI ネットワーク社会推進会議(総務省)
	8	「人間中心の AI 社会原則」	統合イノベーション戦略推進会議(人間中心の AI 社会原則会議)
	9	人工知能学会「倫理指針」	人工知能学会
	10	AI・データの利用に関する契約ガイドライン 1.1 版	経済産業省
	11	Ethics Guideline for Trustworthy AI	European Commission (High Level Expert Group on AI(HLEG))
	12	Ethically Aligned Design	IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
	13	Asilomar AI Principles	Future of Life Institute (FLI)
	14	Tenets	Partnership on AI

表 2 をもとに内容を精査し，本稿ではロボットの開発において参考にしてきた「1.IoT セキュリティガイドライン」と「6.AI 利活用ガイドライン」について述べる．

(1) IoT セキュリティガイドライン

IoT セキュリティガイドラインは，経済産業省や総務省が 2016 年 7 月に発行したものである[7]．IoT 機器やシステム・サービスについて Security by Design を基本原則としつつ，求められる基本的な取り組みを明確化することによって，産業界による積極的な開発等の取組を促すことなどを

目的にしている．一方で，一律に具体的なセキュリティ対策の実施を求めるものではなく，守るべきものやリスクの大きさを踏まえ，役割・立場に応じて適切なセキュリティ対策の検討を行うことを期待している．

(2) AI 利活用ガイドライン

AI 利活用ガイドラインは，総務省情報通信政策研究所が 2016 年 10 月に「AI ネットワーク社会推進会議」を立ち上げ，2017 年 7 月に公開した AI 開発ガイドライン案[8]をもとに拡張したもので，2019 年 8 月に公開された．このガイドラインでは 6 つの基本理念と 8 つの AI 利活用原則を提唱し，開発者のみならず，AI をビジネスに活用する利用企業や，企業の製品やサービスを通じて AI に触れる一般消費者も対象としている．しかしながら現状は強制力がない「ソフトロー」と位置づけ，AI 開発を制約するものでないと解釈される．

IoT セキュリティガイドラインは，実際に利活用している開発ベンダーも多く，実装に近いものである．そのため当該ガイドラインをもとに IoT セキュリティ設計を進め 3 章に述べるセキュリティ防御機能の実装を行った．特に，当該ガイドラインにおいて，「外部インターフェース」へのセキュリティ対策を奨励していることが記載されていることからクラウドとの出入口へのセキュリティ対策が重要であると考え，クラウド IF におけるセキュリティ対策を中心に実装することとした．

一方，AI 利活用ガイドラインにおいては，個別具体的な手引きになってはいない．ただし当該ガイドラインに含まれる AI 利活用原則にて「安全の原則」と「尊厳・自律の原則」があり，利用者に危害を与えたり，人間の尊厳と個人の自律への尊重を妨げないようにしたりすることが提言されている．そこでロボットの言葉遣いに着目し，利用者への不安感の緩和を目的に，4 章で述べるように，発話生成に使用するために抽出されるキーワードに不快感を与えるものを選定できるような仕組みを検討した．そして不快感を与える用語のブラックリストをデータベース化するとともに，今後学習機能を搭載する際に利用していくことを考えた．

3. クラウド IF の IoT セキュリティ

2 章ではロボットの安全・安心対策として，クラウド IF での「IoT セキュリティ関連」と「AI 開発・利用関連」に分けて述べた．3 章では本稿における IoT セキュリティ対策について説明する．

本研究で試作したロボットは，クラウド内の処理機能以外は，実験室内に構築している．またロボット本体とクラウド IF を除く発話生成などに関わるサーバおよび端末は，実験室内のローカルネットワークで接続されている．現在ロボットを利用した実験は実験室内で実施されており，情報セキュリティの障害が発生したとしても，ネットワーク回線を即座に遮断することにより迅速な対応が可能である．

一方，将来的には，施設や家庭内にロボットを設置して，視聴実験を行うことを検討している．その場合，施設や個人宅などの実験室外で長期間の視聴実験を行うためには，ロボット本体と軽量な可搬型端末のみを持ち込む程度で実験を行えるシステム構成が求められる．

しかしロボットが施設や個人宅で撮影・収録した映像・音声をクラウドへ送信し，クラウド内で画像・音声認識処理を行うためには，個人情報やプライバシーを保護するための対策が必要となる．これは実験室内という閉じた領域でのデータのやり取りと異なり，外部とのネットワーク接続により侵入・攻撃のリスクが増えるためである．

3.1 クラウド IF の IoT セキュリティ設計要件

本研究ではセキュリティ対策の取り組みとして，ロボットに搭載したカメラ・マイクが接続され，クラウドへ直接接続するクラウド IF への対策に絞った．その上で現時点でのクラウド IF およびクラウドへのセキュリティ脆弱性を調べるために，外部調査機関を利用して脆弱性検査を実施した．この診断項目にあたって外部調査機関では，表 2 に挙げた「IoT セキュリティ関連ガイドライン群」に加え，検査基準として情報推進機構（IPA）が公開している「安全なウェブサイトの作り方」[9]および OWASP が公開している「OWASP IoT Top 10 2018」[10]を参照した．

「安全なウェブサイトの作り方」は，「ウェブアプリケーションのセキュリティー実装」に関して，11 種類の脆弱性を取り上げている。そして，それぞれの脆弱性で発生しうる脅威や，特に注意が必要なウェブサイトの特徴等を解説し，脆弱性の原因そのものをなくす根本的な解決策，攻撃による影響の低減を期待できる対策を示している．

「OWASP IoT Top 10 2018」には，開発者，製造業者，企業，消費者が IoT システムの作成や利用に関してより良い判断をするために，避けるべきセキュリティ上の注意点が Top10 形式で示されている．

以上この 2 つの検査基準は多くの脆弱性検査機関で一般的に採用されており，デファクトスタンダードとなっている．今回行った外部調査機関によるクラウド IF の脆弱性検査においても，この 2 つの検査基準をクラウド IF に構築したプログラムのソースコードの診断項目に採用した．またソースコード診断といった静的診断に加え，攻撃フレームワーク（Pacu）を利用したペネトレーションテスト（動的診断）も合わせて実施した．

3.2 クラウド IF のセキュリティ防御機能

脆弱性検査の結果から，クラウド IF には主にデータの改ざん・搾取防止，出入口対策，そしてログ収集管理のセキュリティ対策が必要であることがわかった．そこで，公開されている IoT セキュリティガイドラインを参考に，データの改ざん・搾取防止には「データの秘匿化」「データの識別化」，出入口対策には「入力デバイスの特定・検出」「データ入出力部の強化」，ログ収集管理には「セキュリティログの収集と分析」を考案した．

IoT セキュリティ設計要件としてまとめたものを表 3 に示す．さらに表 3 をもとに，クライド IF において実装したセキュリティ防御機能の概要を図 3 に示す．図中の濃い箇所が本章で述べている IoT セキュリティ防御機能を実装している箇所である．なお網線箇所は 4 章で紹介する不快用語用マルチエージェントシステムである．

表 3　IoT セキュリティ設計要件

	要件	説明
1	IoT 機器（ロボット・カメラ・マイク・エッジ PC 等）の特定・検出	IoT 機器が持つ識別符号等をクラウド IF 側であらかじめ登録し，照合することにより，映像・音声の受信の判断を行えるようにする．

2	映像音声データの特定識別化	映像・音声データに識別信号などを付加し，クラウド IF もしくはクラウド側であらかじめ登録していた信号と照合することで，データ処理を実施するかの判断ができるようにする．
3	映像音声データ・処理データの秘匿化（暗号化）	映像・音声データおよび処理結果データ（キーワード）が第三者に閲覧されないように，暗号化や匿名加工化などの秘匿性を維持する．
4	入出力部（クラウド IF・クラウド）の強化	映像・音声データおよび処理結果データ（キーワード）が第三者に詐取されたり，攻撃されたりしないように入出力部の制限をかける．
5	セキュリティログの収集と分析	不具合を迅速に判断するために，必要なセキュリティログを収集し分析できるようにする．

3.2.1 IoT 機器（ロボット・カメラ・マイク・エッジ PC 等）の特定・検出

IoT 機器の特定・検出要件はロボットに固有の識別符号を付加する実装により対策した．IoT 機器としてのロボットが識別符号等を持ち，予めクラウド IF 側に当該識別符号を登録しておくことで，使用時に識別符号を照合することにより，映像音声の受信の判断を行えるようにする．

本実装ではインデックスキャプションとして，読み取り困難な濃淡パターンを実装した．また，クラウドへの送信時の漏洩防止，IoT 機器の特定を高速化するため，IoT 機器の特定をクラウド IF 内で行い，クラウドへは判断結果のログのみを送付する方式とした．

これにより次項「3.2.2 映像音声データの特定識別化」が機能するよう，クラウド側で指定のクラウド IF からのデータであるかどうか検知し，正しい経路でデータが伝送されたかを認証できるようになった．

3.2.2 映像音声データの特定識別化

映像音声データの特定識別化要件は，ロボット側において付加された映像・音声データの識別信号を，クラウド IF 側で予め登録していた信号と照合することで，指定されたロボットからのデータとして処理を実施するかどうかを検知・判断する実装により対策した．これにより万一異なるロボットからの入力データを検知し異常が発生した場合でもアラートで警告を出し，「3.2.5 セキュリティログの収集と分析」にて停止判断ができる．

さらにクラウド IF 自身を特定するため，当該クラウド IF の MAC アドレス，IP アドレスを利用した特定識別を実装した．これにより指定されていないクラウド IF（即ち不正なユーザ）を検知し，システムを自動停止するための警告を出力する．

3.2.3 映像音声データ・処理データの秘匿化（暗号化）

映像音声データ・処理データの秘匿化要件は，映像音声データおよび処理結果データ（キーワード）が第三者に閲覧されないように，暗号化による秘匿性を維持する実装により対策した．具体的には 3.2.2 項の「映像音声データの特定識別化」で特定した映像音声データに暗号化を施し，クラウド内で送受信・保存および演算できるようにした．

本実装では送受信手段に TLS（Transport Layer Security）を用い，保存手段については，RSA 暗号方式を採用した（秘密鍵はシステム管理者が所有し，クラウド IF に公開鍵を配布して暗号化する）．なおクラウド内ストレージ（AWS S3）の暗号化は，AWS の商用サービスを使用している[1]．本実装では AWS S3 バケットに保存する際および認識結果の json 形式ファイルを暗号化した．

3.2.4 入出力部（クラウド IF・クラウド）の強化

入出力部の強化要件は，映像音声データおよび処理結果データ（キーワード）が第三者に搾取されたり，攻撃されたりしないように入出力部の制限をかける．

本実装ではデータ処理異常検出通知から警告表示を受信し，クラウド IF の動作を一時停止させる自動停止モードを図 3 のセキュリティ監視制御に設定できるようにした．また警告表示を目視で確認し，手動で一時停止できるよう手動停止モードも設定できるようにした．

さらに停止後，手動で動作復帰できるモードを設定し，復帰後に自動停止モードあるいは手動停止モードを選択できるようにした．これにより異常を検知した場合に障害の発生を未然に防止することができセキュリティを確保できる．自動モードでエラーを検出した場合には，自動停止する．また，警告を検出した場合，手動モードでシステムを停止できる．

3.2.5 セキュリティログの収集と分析

不具合を迅速に判断するために必要なセキュリティログを収集できるようにした．具体的には「3.2.4 入出力部の強化」にてクラウド内処理を停止する制御をする．これによりクラウド IF にて，システム全体の管理及びデータの改ざんや紛失などによるセキュリティ障害を未然に防止でき，

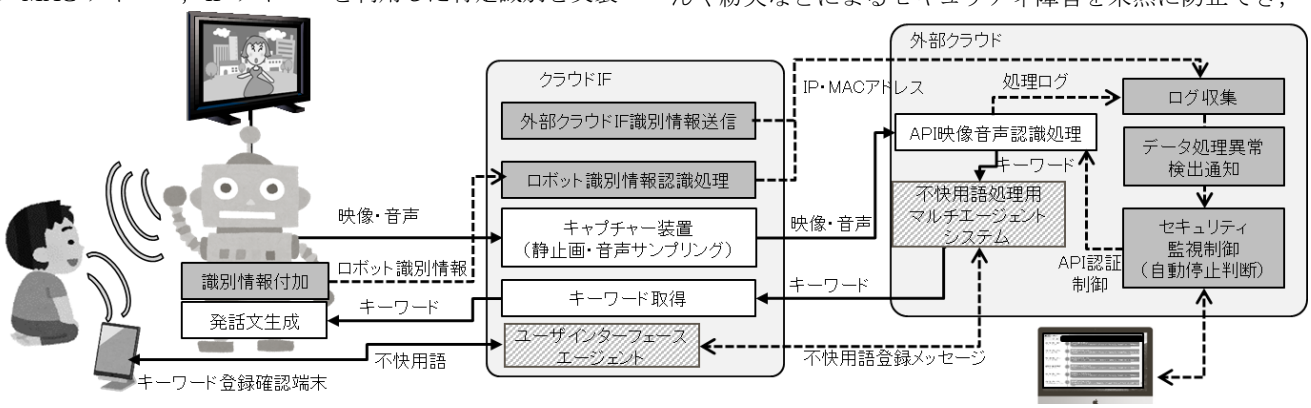

図 3　セキュリティ防御機能の実装
（図中の濃い箇所が IoT セキュリティ防御機能．網線箇所は不快用語用マルチエージェントシステム（4 章参照））

1) Amazon S3 のデフォルトバケット暗号化の有効化
（https://docs.aws.amazon.com/ja_jp/AmazonS3/latest/userguide/default-bucket-encryption.html）

「ログ異常検知機能」で検出された警告ログによるシステムの自動停止もしくはユーザによる手動停止が可能となった．

3.2.2 映像音声データの特定識別化およびログ異常検知機能の異常検出の結果通知をグラフィック表示できるようにした（図 4 参照）．実装においてはログ異常を示すテスト用のデータを用意し，異常を検知し，結果として 3.2.4 入出力部の強化によりクラウド IF の動作を一時停止できること，および一時停止したことを示す表示を確認した．

図 4　情報セキュリティ監視制御画面

なお，従来 AWS が管理用サービスとして有償提供している CloudTrail は 15 分以内にログを配信するとしているが，この時間は保証しないとしている2)．一方，実装したログ異常通知機能を使った検知はクラウド IF で取り扱うログ収集に限定して機能の高速化を図ったことにより，1 分程度で異常検知処理が可能となった．

3.2.6 セキュリティ防御機能実装時の自己検証

以上，本研究では 3.2.1 から 3.2.5 の各要件における実装を行った．これにより異なる「ロボット」や「クラウドインターフェース」が接続されたり，「認証許可のないもの」から「クラウド」にアクセスされたりした場合に，自動でシステムを停止することができ，万一データが搾取されても，暗号化で防ぐ効果が期待できる．

なお，自己検証として，JPCERT が提供しているチェックリストから本稿の実装内容に該当する項目を情報処理安全確保支援士の監修のもと取捨選択・修正して行った[11]．表 B（文末付録）に検証結果の一部を示す．自己検証の結果，暗号化やシステム設定，通知，セキュリティ管理などでは，自己検証チェックリストの該当項目を満足することが確認できた．一方，セッション管理など現時点では未使用のため未検証の項目については今後の機能拡張を検討する際の確認項目と考える．

4. ユーザに与える「不快感の緩和」への対応

次にユーザに与える不快感の緩和への対応として「不快用語の管理と検出」について述べる．本研究では「不快や不安」を与える「発話やしぐさ」，例えば差別に関わる発話や，普段とは違う突拍子のない仕草，気分を害する行動などの排除を目的とした．その上で最初の取り組みとして，発話生成に使用するキーワードから，不快用語を排除することを検討した．これはロボットが，キーワードをもとに発話生成する仕組みであることから，まずはキーワードそのものへの対策が必要であると考えたからである．

4.1 不快用語の定義

まず不快を与えるキーワード「不快用語」を定義するにあたり，差別用語に関する先行文献調査[12][13][14][15][16][17]と「NHK 放送ガイドライン」[18]の関連部門にヒアリングを実施し，不快用語をリストアップした．また，「同じキーワードでも不快と感じるかどうかには個人差があり，不必要な制限をかける」のは避けるべきとの助言を受けた．

そこで不快用語を 3 種類に定義して分類し，一部を個別管理できるように設計した．本稿で定義した不快用語I類，II類およびIII類を表 4 に示す．

I類は「無条件に差別を示す」として，使用が制限されている用語．II類は使用する環境により差別を示すとして，使用を留意されている差別的用語．そしてIII類はユーザが個人的に，差別・不快と感じる用語とした．

表 4　不快用語の定義

不快用語Ⅰ類	放送ガイドライン（解説編）[18]，市販されている複数の差別用語に関する専門書により，無条件に差別を示すとして使用制限されている差別用語（335 語）
不快用語Ⅱ類	前述の専門書により，使用する環境により，差別を示すとして使用を留意されている差別的用語（85 語）
不快用語Ⅲ類	前述の専門書には記載されていないが，ユーザ個人が差別・不快と感じる用語

※文献 18 参照．なお（解説編）は非公開．

4.2 不快用語管理マルチエージェントシステム

本研究では「不快用語の管理と検出」を行うために，4.1 節で定義した不快用語にもとづき，次の要件を設定した．

- I類は確実に削除もしくは置換する．
- II類とIII類はユーザごとに個別管理できるようにする．
- 似たようなテレビ視聴傾向（4.2.3 項参照）のあるユーザ同士で不快用語を共有できるようにする．

これを実装するために，ユーザおよびユーザモデルそれぞれにエージェントを配備し，エージェント同士が協調してユーザの不快用語の蓄積・更新できる，マルチエージェントシステム（Multi-Agent System．以下，MAS）を設計した．図 5 にその概要図を示す．

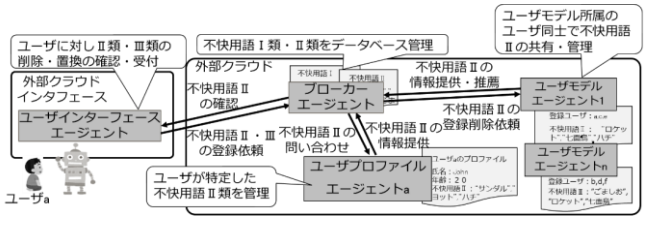

図 5　不快用語処理用マルチエージェントシステム概要

4.2.1 不快用語データベース

不快用語データベースでは 4.1 節で分類定義した不快用語I類，II類およびIII類それぞれのデータベース（リスト）を作成し，それぞれ更新できるように設計している．図 3

2) CloudTrail の詳細
（https://docs.aws.amazon.com/ja_jp/awscloudtrail/latest/userguide/how-cloudtrail-works.html）

中の「API 映像音声認識処理」部から抽出されたキーワードが不快用語に合致した場合，不快用語I類であれば予め作成した不快用語I類のデータベースとの照合により削除し，不快用語II類であれば削除を希望するかどうかユーザに確認できるようにした．以降，当該用語が再度出現した場合は4.2.2項に示すユーザプロファイルを照合したうえで削除する．不快用語III類については，ユーザが自発的に不快を示した場合において削除し，当該用語が再度出現した場合は不快用語II類と同等の処理を行う．

4.2.2 ユーザプロファイルデータベース

各ユーザのユーザプロファイルを作成し，不快用語に接触した日時を保存する．ユーザプロファイルはそのほかに，初期設定として氏名（ニックネーム可），性別，年代，居住経験地域，職業，興味を含み，追加で登録できるよう予備項目を準備している．当該ユーザプロファイルはユーザが属するユーザモデルデータベース（4.2.3 項参照）と照合できるようにする．これにより不快用語への取捨選択ができ，その結果を保存する．また 4.2.4 項の MAS からの通信によりユーザの回答により追加・置換された不快用語を格納する．

4.2.3 ユーザモデルデータベース

ユーザモデルはユーザが属するグループ（クラスター）を指す．ユーザモデルデータベースにはユーザのユーザプロファイルが，ユーザモデルのプロファイルと近似している場合に格納する．初期値として居住経験地域，テレビ視聴傾向を含み，追加で登録できるよう予備項目を準備しておく．

なおテレビ視聴傾向については4.2.4マルチエージェントプラットフォームによりユーザに予めアンケートをとり，初期設定として割り振れるようにする．本研究では例として，８つのテレビ視聴傾向（笑い刺激型，熱中トレンド型，ロマンフィクション型，息抜きザッピング型，気楽おトク型，健全実用型，報道教養型，関係希薄型）を用いた[19][20].

4.2.1，4.2.2 および 4.2.3 の各データベースに格納するデータは暗号化を施す．またデータベースの構築には solid（https://solidproject.org）を利用した．

4.2.4 マルチエージェントシステム（MAS）

上記 4.2.1 項，4.2.2 項および 4.2.3 項の各データベース間の通信連携を管理・運用するための基盤として MAS を作成した．当該 MAS では，例えば，不快用語II類に関わる用語が検出されたときに，ユーザに対して当該用語を自発的に提示し，不快に感じるかどうか尋ねる．その結果を 4.2.2 項のユーザプロファイルに保存する．一方ユーザから不快用語III類に関わる用語を不快に感じると伝えた場合は，当該 MAS を通じてユーザプロファイルに保存する．

MAS の構造とし，ユーザプロファイルとユーザモデルそれぞれにエージェントを配置・連携させることで，双方がダイナミックにデータの生成・蓄積・更新・削除が可能となる．なお実際には不快用語データベースに配置させたブローカーエージェントを仲介として不快用語をやり取りする仕組みとしている．さらにエージェント通信言語の暗号化により，ユーザプロファイルとユーザモデル間の指示命令・メッセージ内容・データの秘匿性を確保することができる．なお MAS として本仕様書では JADE[21]のアルゴリズムを採用し，Python 仕様のマルチエージェントシミュレーションである Mesa(https://mesa.readthedocs.io/en/master/)で実装した．

4.3 動作確認

不快用語は頻繁に発生するものではないため，不快用語のダミーデータを準備し，動作確認を行った．図 6 に不快用語の登録状況を管理する監視画面を示す．右側のユーザごとにプロファイル登録された不快用語は，中央のブローカーエージェントを介して，ユーザが属する左側のユーザモデル内で，他のユーザと不快用語を共有する．互いに不快用語の情報提供や推薦することで，自身のプロファイル更新に役立てることができるようになっている．

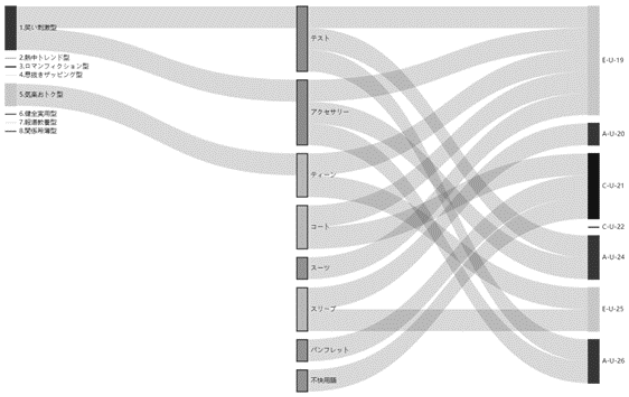

図 6　MAS 内不快用語管理（監視画面）

5. 残課題と今後に向けて

以上，本稿ではロボットの安全・安心対策として，クラウド IF での「セキュリティ防御（3 章）」と「不快感の緩和（4 章）」それぞれについて，要件設定および設計・実装，ならびに動作確認結果について述べた．本章では，これを受けてこれらの残課題と今後に向けて述べる．

5.1 セキュリティ防御

3.2.1 項に述べた識別情報付加のインデックスキャプションについては，固定の文字列をインデックスキャプションとし，読み取り困難な濃淡パターンに変換して実装した．したがって当該手法では，変更されないインデックスキャプションが第三者に漏えいした場合に複製・なりすましされる恐れがある．そのためインデックスキャプション方式としてはランダムな文字列化を行い，さらにステガノブラフィー化（例：電子透かし）を施したり，あるいはハッシュ化（ハッシュ値に変換して伝送）を利用したりするなど，より秘匿性の高い方式が必要と考える．また個人情報・プライバシー保護への対策についても，現状では映像コンテンツはそのままクラウド API の認識処理を施している．すなわちクラウド IF を経由してクラウドに送ってしまっている．そのためクラウドへは，撮影された映像をそのまま送信するのではなく，テレビ画面などキーワード抽出に必要な最小限の映像領域を切り出して送るなど，プライバシー保護を強化することが求められる．

そこで番組解析に直接不要なユーザの顔画像や部屋の映像は，クラウド IF もしくはロボット本体でのローカル処理で識別処理する．たとえば，部屋を撮影した場合，テレビ

受像機のエリア（テレビ番組画面）とそれ以外のエリア（人や家具などの家庭内を映した映像）とを切り分け，それぞれ別系統で認識処理を施すことが求められると考える．さらには今回実装しなかった音声認識においても同様の施策が考えられる．

また 3.2.2 項に述べたクラウド IF へのアクセス制限について，指定時間帯の設定や指定 IP アドレスについても，それぞれ改ざんされた場合に，なりすましで警告を回避される可能性は残されている．そのため他の認証方法との組み合わせや暗号化への対応等を検討することが考えられる．

さらに本研究におけるセキュリティ対策はクラウド IF に限定してクラウドの異常検知に特化した．3 章で述べたように今後実験場所を特定しない室外での視聴実験を想定するならば，セキュリティ防御機能をクラウド IF からテレビ視聴ロボット全体に拡張し，包括的な異常検知機能を検討する必要がある

5.2 不快用語をもとにした不快感の緩和

人に不快を与える発話を制限するために，クラウドの認識 API によって抽出されたキーワードから不快用語を制限する仕組みを MAS で実装した．これはロボットがキーワードをもとに発話生成する仕組みを応用しており，あくまで基礎的手法であると捉えている．

即ち，人が不快に感じる発話は，単語のみで不快な場合だけでなく，単語を組み合わせることで不快を与えることもあるため[22]，不快用語だけでは十分とはいえない．そのため不快用語を削除・置換したキーワードから生成した発話文を，さらに言語解析により不快を感じる恐れのある文の削除・置換することも求められる．

さらには発話だけでなく，ロボットの振る舞いにおいても，急に騒いだり，ユーザの意図しないタイミングで話しかけたりするなどして，不快感を与える可能性がある．これらについては，ロボットを使った視聴実験を通じて実験参加者からの評価データを分析し，動作の許容範囲や発話タイミングについて調査を進めることが求められる[23].

また本研究では不快用語をユーザに提示して直接取捨選択させる手法をとっている．本手法は，本人の意志により正確に不快用語を判別できる利点がある．しかし不快を与えかねないキーワードを本人に直接提示することへの指摘も考えられる．不快用語を提示したときでもユーザに不快感を与えるえる恐れがあるためである．

本研究では，MAS を用いてユーザプロファイルとユーザモデルとを協調させる仕組みを構成した．あるユーザプロファイルで登録された不快用語は，所属するユーザモデルがそれを参照し，別のユーザのユーザプロファイルに推薦する．推薦されたユーザは，自分で不快用語の登録を判断することになる．

ここで不快用語を推薦されたユーザが，当該ユーザが信頼するユーザの不快用語であれば本人に承諾なしで自動登録できる仕組みも考えられる．例えば，親が小学生以下の子のユーザプロファイル内の不快用語を自動更新できる仕組みである．これにより本人に不快感を与えることなくユーザプロファイルの不快用語を更新することができる．

さらには，学習・推論機能を導入することにより，ユーザそれぞれに合わせた不快用語の自動登録ができる仕組みも考えられる．例えば本研究では未実装である推論エージェントを構築し，他のエージェントと協調しながらユーザプロファイルを更新する仕組みが考えられる．こうした学習・推論機能も取り入れることにより，2.2 節に示した AI 開発・利用面での機能追加に取組むことが必要と考えられる．

今後ユーザプロファイルに登録するデータの種類や量を増やしていくに従い，個人情報・プライバシー保護，倫理面での考慮が必要である．3 章のセキュリティ防御によって保護対策を施し，その上で安全・安心にユーザプロファイルを維持管理することが期待できる．

6. まとめ

本研究ではテレビ視聴ロボットの安全・安心のための，「ユーザの個人情報保護に基づいた情報セキュリティ対策」と「ロボットが起こす不適切な行動により，ユーザに与える不快感の緩和への対応」について，先行調査とシステム試作および動作確認を行った．情報セキュリティの確保においては，ロボットに関するガイドラインの調査，セキュリティ設計要件の策定，およびセキュリティ機能の試作・自己検証を行った．一方不快用語の管理と検出については，先行文献調査により不快用語を定義および分類し，ユーザプロファイルによる不快用語の管理システムの設計要件を策定，および試作と動作確認を行った．

現状の実験室で動作するロボットを使った視聴実験における IoT セキュリティ対策を実装により自己検証できたことは，今後実験室外での視聴実験を検討するにあたり，個人情報・プライバシーの保護対策への取り組みに役立てることができる．また不適切な行動への防止に向けて，まずはロボットの発話に使用するキーワードから不快用語を削除・置換する仕組みを MAS で実装したことは，今後，学習・推論機能の導入や，ユーザプロファイルとユーザモデルとの協調に関わる研究開発への貢献に加え，実験運用中におけるセキュリティ対応時での人的判断への負担軽減・自動化に応用できる．

参考文献

[1] 村﨑康博，星祐太，萩尾勇太，上村真利奈，金子豊，山本正男，“テレビ視聴ロボットに求められる形態と機能”，NHK 技研 R&D 2021 年 冬号 報告 03 (2021)
[2] 金子豊, 星祐太, 上原道宏，“人と一緒にテレビを視聴するロボットの機能検討と試作”，RSJ2017 (2017)
[3] 萩尾 勇太, 金子 豊, 星 祐太, 村崎 康博, 上原 道宏, “人とロボットの共時視聴実験に向けたコミュニケーションロボットの設計と試作”, 映像情報メディア学会年次大会, 33B-31,(2019)
[4] “セキュリティバイデザインとは？ 必要性やメリット，注意点について徹底解説”，サイバーセキュリティ.com，URL:https://cybersecurity-jp.com/security-measures/29134
[5] 村﨑康博，星祐太，萩尾勇太，上村真利奈，金子豊，山本正男，“テレビ視聴ロボット用外部クラウドインターフェースにおけるセキュリティ対策”，情報処理学会研究報告，Vol.2020-CSEC-88，No.36，pp.1-8（2020）
[6] 村﨑康博，金子豊，星祐太，上原道宏：“テレビを一緒に視聴するロボットの開発ガイドライン策定に向けての一考察”，情報処理学会研究報告，Vol.2017-EIP-78，No.14，pp.1-7（2017）
[7] “IoT セキュリティガイドライン ver1.0”，IoT 推進コンソーシアム，総務省，経済産業省 20l6 年 7 月 5 日 URL: https://www.meti.go.jp/press/2016/07/20160705002/20160705002.html (2016)
[8] “国際的な議論のための AI 開発ガイドライン案”，AI ネットワーク社会推進会議ホームページ，URL:http://www.soumu.go.jp/main_content/000499625.pdf (2017)

[9] 情報処理推進機構：“安全なウェブサイトの作り方”，URL:https://www.ipa.go.jp/files/000017316.pdf (2021)
[10] OWASP：“Internet of Things (IoT) Top 10 2018”，URL:https://owasp.org/www-pdf-archive/OWASP-IoT-Top-10-2018-final.pdf (2018)
[11] JPCERT/CC，IoT セキュリティチェックリスト，URL:https://www.jpcert.or.jp/research/IoT-SecurityCheckList.html (2020)
[12] 用語と差別を考えるシンポジウム実行委員会編，“ 差別用語－ゆたかな日本語をめざして－”，汐文社 (1978)
[13] 山中央，“新・差別用語”，汐文社 (1992)
[14] “記者ハンドブック 第 13 版 新聞用字用語集”，共同通信社 (2016)
[15] 小林健治，“最新差別語不快語”，にんげん出版 (2016)
[16] 上原善広，“私家版差別語辞典”，新潮選書 (2011)
[17] 高木正幸，“差別用語の基礎知識'99”，土曜美術社出版 (1999)
[18] NHK 放送ガイドライン 2020，URL: https://www.nhk.or.jp/info/pr/bc-guideline/
[19] “８つの「テレビ視聴型」とステーションイメージ”，NHK 放送文化調査研究年報 45(平成１２年度版) (2000)
[20] “視聴者の視聴タイプを利用した番組選択システム”，情報科学技術フォーラム 2002 Oe2-4 (2002)
[21] JADE，URL:https://www.infoq.com/jp/articles/JADE_20110915/ (2011)
[22] 東中竜一郎，“AI の雑談力”，角川新書 (2021)
[23] 星 祐太, 金子 豊, 萩尾 勇太, 村﨑 康博, 上原 道宏, "ロボット発話に向けたテレビ視聴時の人同士の対話解析," 信学技報, CNR2019-1, pp.1-6 (2019)

付録

表 A 認識処理 API の公開ドキュメント（仕様）URL

	サービス名	ドキュメント URL
A	AWS Rekognition Image	https://aws.amazon.com/jp/rekognition/resources/
B	AWS Rekognition Video	
C	AWS Elastic Transcoder	https://docs.aws.amazon.com/elastic-transcoder/index.html
D	Azure Computer Vision	https://docs.microsoft.com/ja-jp/azure/cognitive-services/computer-vision/
E	Azure Speech Service	https://docs.microsoft.com/ja-jp/azure/cognitive-services/speech-service/
F	Azure Translator Text	https://docs.microsoft.com/ja-jp/azure/cognitive-services/translator/

表 B セキュリティ防御機能実装時の自己検証用チェックリスト（一部．表中の※は保留もしくは未実証事項）

大項目	小項目	開発する際に確認する項目	確認	回答の補足
セキュリティ管理	ログ管理機能	ログ情報が見られることを確認	○	クラウド API 利用時のログを保存
	セッション管理 (Cookie 設定)	Cookie の適切な値に secure 属性，HttpOnly 属性が設定されていることを確認	※	機能未使用
	セッション管理 (URL リライティング)	URL にセッション ID が埋め込まれていないか確認	※	機能未使用
	セッション管理 (ログイン時や重要な確定処理の時のセッション ID の払い出し)	ログイン時や重要な確定処理の前後でセッション ID が変わっていることを確認	※	機能未使用
	クライアントデータの操作のセキュリティ対策	他のアカウントのデータが操作・閲覧できないことを確認	○	閲覧権限の分離を確認
	システムデータの操作のセキュリティ対策	特定のシステム管理者以外でシステムデータが操作・閲覧できないことを確認	○	ログインには各クラウドのアカウントが必要
	クラウドインターフェースやネットワークの脆弱性（API インターフェースやクラウドベースの Web インターフェース等）	公開情報を元に脆弱性情報を確認	○	利用している外部ソフトウェアをドキュメントに明示し，脆弱性情報が公開された場合はパッチ対処する
	XSS, SQLi, および CSRF の脆弱性	公開情報を元に脆弱性情報を確認	○	外部調査機関にて指摘・対処済
	Web アプリケーションの SSL 証明書	利用している証明書を確認	※	現在は試験フェーズのためホスト確認画面は証明書を利用せずに運用．本格運用の際には要検討．
アクセス制御	管理されていない物理手段によるアクセス	管理されていない物理的手段によるアクセスに対して制限ができているか確認	○	クラウド IF にてアカウント，パスワードによる認証
	リモートアクセス用ポートのデフォルトポート	デフォルトポートの変更を行えるか確認	※	基本的に標準ポートから変更できない
	無線通信におけるセキュリティ(暗号化方式)	接続時にセキュアな暗号化方式が選択されていることを確認	○	ログの保存，閲覧には TLS を使用
	無線通信におけるセキュリティ(WPS)	WPS が動作するか確認	○	ただし利用していない
不正な接続	ネットワークポートの制限	ポートの制御が設定したとおりに開閉されていることを確認	○	Web 側の仕様を確認
	UPnP	デバイスを接続したときに，設定した通りの挙動になっていることを確認	※	未使用
暗号化	データの暗号化機能	データを暗号化する機能があることを確認	○	SSL/TLS を利用
	通信の暗号化機能	暗号化通信が利用できるようことを確認	○	SSL/TLS を利用
	暗号化方式	利用している暗号化方式を確認	○	ログの閲覧には SSL/TLS を利用．クラウド側 REST API 呼び出し時は SSL/TLS 接続を利用
	証明書更新機能	証明書が有効であることを確認	○	証明書を利用していない
システム設定	センサーの動作状況確認機能	センサーの動作状況を確認	○	カメラ側から機器識別画像の通知機能がある
	ログのセキュリティ管理	閲覧権限のないユーザーでログが見えないことを確認，閲覧可能なユーザーでログが書き換えられないか確認	○	ログの閲覧と作成とで権限を変更していることを運用時に確認
通知	セキュリティイベントのアラートと通知機能（状態異常等）	仕様通りに動作するか確認	○	メール，クライアントソフトに通知する
	セキュリティイベントのアラートと通知機能（認証失敗，証明書の期限切れ等）	仕様通りに動作するか確認	○	メール，クライアントソフトに通知する
セキュリティ管理	ログ管理機能	ログ情報が見られることを確認	○	クラウド API 利用時のログを保存
	セッション管理 (Cookie 設定)	Cookie の適切な値に secure 属性，HttpOnly 属性が設定されていることを確認	○	機能未使用