令和7年度　インターネット上の偽・誤情報等への対策技術の開発・実証事業

電話音声フェイク検知および自治体向け偽・誤情報総合対策の開発・実証

成果報告書

2026/3/19

技13_NABLAS株式会社

目次

目次

1-1. 開発・実証のサマリ

アプローチする課題・目指す姿	**[アプローチする課題]** **電話音声フェイク検知** • 高精度な音声生成AIの急速な進化により、人間の耳では本物と区別できないなりすまし音声が増加し、電話詐欺被害が拡大している。 **自治体向け偽・誤情報総合対策** • SNS上でフェイクニュースやディープフェイク投稿が急増し、特に災害等の有事において、虚偽情報と真に重要な情報の判別が困難になっている。 **[目指す姿]** • フェイク検知・真正性担保・発信者証明が社会インフラとして普及し、デジタルコミュニケーション全体の信頼性が確保される社会 **電話音声フェイク検知** • 電話・通話環境にフェイク音声の自動検知が組み込まれ、利用者が意識せずともなりすまし・詐欺から守られるシステムの構築 **自治体向け偽・誤情報総合対策** • 自治体・公的機関がSNS上の情報の真偽を即座に判断でき、市民が有事でも信頼できる情報にアクセスできる情報環境の構築
技術区分	コンテンツの真偽判別支援技術、改ざん検知技術 真正性保証・信頼性判断支援
対象とするモーダル種	文章、画像、音声、動画
実施体制（下線：技術開発主体）	NABLAS株式会社 再委託先：NTT東日本株式会社（２社間でコンソーシアムを組成） 再々委託先：ageet株式会社 , NTTテクノクロス株式会社 再々々委託先：株式会社HBA

技術開発の取組・成果

[電話音声フェイク検知]
- 実際の電話環境でも精度を維持モデルを開発し、フェイク音声を検知を実現
- 最新生成モデルに対し迅速に検知性能を評価し改善できる体制を整備
- 電話サービス組込へリアルタイム処理可能な軽量かつ高精度検知エンジン設計

[自治体向け偽・誤情報総合対策]
- 総合的フェイク検知技術および電子透かしおよびDID/VC機能の新規開発
- システムに組み込みにより、「検知」と「証明」を統合した、自治体が実用可能な総合対策システムを構築

社会実装に係る取組・成果

[電話音声フェイク検知]
- NTT東日本との共同開発により電話着信時のフェイク音声リアルタイム判定アプリを開発、ひかり電話・携帯電話・IP電話の実網で検証を実施
- 検知アプリにより、利用者がリアルタイムで判定結果を確認可能にした

[自治体向け偽・誤情報総合対策]
- 伊那市と２回の虚偽投稿の発生を想定した実証を実施
- 実務フローに即した形で想定通り証明書（VC）が埋込まれることを確認し、SNSで拡散後でも真正性が把握できることを確認から実用レベルへ改善

技術開発及び社会実装にあたっての課題・展望

[電話音声フェイク検知]
- 音声処理の全環境を網羅的な対応は一つのモデルでは非常に難易度が高く、最新モデルの検知の難易度が急速に上昇している。
- 利用環境に応じ最適なモデルを自動選択し、環境に左右されず、最新のAIにも対応可能な検知機能をシステムを構築し、あらゆる音声コミュニケーションの場へ機能を拡張していく。

[自治体向け偽・誤情報総合対策]
- 多様な課題を担当する人員や業務フローが未整備、ないしは発足段階のため、業務が停滞することもあり、危機や課題に直面し初めて対策を講じる形になり後手の対応となっている場合が多い。
- 偽・誤情報対策体制ごと代行可能なサービス・インフラを提供することで、自治体の負担なく対策可能なサービスを提供していく。

代表者コメント

NABLAS社 取締役
鈴木都生

電話・自治体向けどちらも実環境・実ニーズを想定した技術開発と実証実験を行いました。
様々な課題や改善点も見えてきておりますが、実環境で提供できる価値も確認することができ、社会実装と社会課題解決に向けて大きな一歩になったと感じています。

目次

2-1. 開発技術によりアプローチする課題

電話音声フェイク検知

**米国ではAI生成音声詐欺被害が1,100万ドルに達し(1)、
日本国内の特殊詐欺被害額年間1,414.2億円(2)の約8割は
電話による当初接触でありAI生成音声の悪用リスクが急速に高まっている。**

社会的課題

上司を装う送金詐欺

- AI生成による音声やディープフェイクを使用したなりすまし電話・投資詐欺の被害が拡大している
- 米国: AI生成音声によるなりすまし電話の詐欺被害が1,100万ドル（約15億円）に達した(1)
- 香港: ディープフェイクビデオ会議でCFOを装い2億香港ドル（約38億円）を詐取(3)
- 日本: 2024年11月に国内メーカーで社長の声を学習したAI音声による送金詐欺未遂が発生(4)

技術的課題

**AIによる声の完全模倣
に追いついていない検知技術**

- 新しい生成技術が次々と登場し、検知が困難になり続けている
- 電話特有のCODEC変換・エコーキャンセル処理・サンプリングレート変化により検出精度が低下
- 検出モデルの大半は電話音声での利用が想定されておらず、実環境で機能しない
- 電話はリアルタイムでの検知が求められるが、高精度な検知モデルは処理負荷が大きく、通話品質を維持しながら検知する点に課題がある

(1) McAfee "Beware the Artificial Impostor" | https://prtimes.jp/main/html/rd/p/000000032.000033447.html
(2) 警察庁「令和7年における特殊詐欺及びＳＮＳ型投資・ロマンス詐欺の認知・検挙状況等について（暫定値）」| https://www.npa.go.jp/bureau/safetylife/sos47/assets/img/new-topics/detail/260213/03/r7_sagi_data_02.pdf
(3) CNN | https://www.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk
(4) 共同通信 | https://nordot.app/1274886991967011553?c=302675738515047521

2-1. 開発技術によりアプローチする課題

自治体向け偽・誤情報総合対策

能登半島地震で104件以上のデマと推定される投稿が確認されている(1)。
総務省調査で過去に流通した偽・誤情報を47.7%が「正しい」または「おそらく正しい」と認識(2)。

社会的課題

災害時のデマ拡散　**クマ情報のデマ拡散**　**選挙時のデマ拡散**

- △ SNS上で大量の偽・誤情報が拡散され、社会に深刻な混乱と被害を引き起こしつつある
- △ 能登半島地震では救助要請投稿1,091件中104件（約10%）がデマと推定され(1)、避難所に関するデマは130万回以上閲覧された(3)
- △ 生成AIの進化により真偽判断が困難化し、閲覧数に応じた収益構造が偽情報の大量拡散を加速
- △ SNSで情報収集する官公庁・企業が偽情報に惑わされ、公式情報の信頼性まで低下する事態が生じている

技術的課題

偽・誤情報の多様化に追いついていない対策技術

- △ 既存の検知技術はスクリーンショット等の部分加工に弱く、コンテンツのみでは信頼性が判定できない
- △ 偽情報の氾濫により真実の情報までもが疑われ、正確な情報が正確なものとして届かなくなっている
- △ 個別技術（検知・ウォーターマーク・発信者証明）は存在するが、それらを統合した総合対策システムが不在
- △ 自治体には偽情報対応の専門人員・運用フローが整備されておらず、技術があっても現場で活用する体制に課題がある

(1)　総務省 令和6年版情報通信白書 | https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r06/html/nd122c00.html
(2)　総務省「ICTリテラシーに係る実態調査」 | https://www.soumu.go.jp/menu_news/s-news/01ryutsu05_02000176.html
(3)　日本ファクトチェックセンター | https://www.factcheckcenter.jp/fact-check/disasters/false-kanazawa-1-5-evacuation-shelter-disaster-certificate/

2-2. 開発技術により目指す姿・ゴール

電話音声フェイク検知

全体ビジョン

フェイク検知・真正性担保・発信者証明が社会インフラとして普及し、デジタルコミュニケーション全体の信頼性が確保される社会

ゴール1

電話・通話環境にフェイク音声の自動検知が組み込まれ、利用者が意識せずともなりすまし・詐欺から守られるシステムの構築

- ひかり電話・携帯電話・IP電話のいずれの通話環境でも動作
- 通話品質を維持しながらリアルタイムでバックグラウンド検知
- フェイク検出時に即座にアラートで利用者に通知

ゴール2

新たな生成技術にも即座に対応でき、電話の信頼性が持続的に確保される基盤の構築

- 新たな生成モデルの出現に対し、検知モデルを迅速に更新できる追随の仕組み
- 電話特有のCODEC変換・ノイズ環境でも高精度を維持するEnd-to-Endモデルの活用
- 通信事業者・Web会議プラットフォーマーが組み込み可能な汎用的な基盤

電話環境下でのフェイク音声をAIが検知しアラートでお知らせ

2-2. 開発技術により目指す姿・ゴール

自治体向け偽・誤情報総合対策

全体ビジョン

フェイク検知・真正性担保・発信者証明が社会インフラとして普及し、デジタルコミュニケーション全体の信頼性が確保される社会

ゴール1

自治体・公的機関がSNS上の情報の真偽を即座に判断でき、市民が有事でも信頼できる情報にアクセス可能な情報環境の構築

- 信頼できる発信者が正しい情報を投稿していることが確認できる
- 悪意のある投稿者がフェイクコンテンツを投稿した時にフェイクと見破れる
- 転載・拡散された際にもその情報が正しいか否かが判定できる

ゴール2

情報発信者の真正性が技術的に証明され、正確な情報が正確なものとして届く仕組みの実現

- 「検知」と「証明」の両面から情報の信頼性を確保
- 偽情報対策が自治体の実務フローに組み込まれ、専門人員に依存せず運用できる体制
- 有事・平時を問わず、市民が信頼できる情報を判断できる仕組みの社会実装

技術 ① 総合的フェイク検知技術

SNS上の画像・動画・テキスト情報を
複合的に解析しフェイクを見破る

SNS上のフェイク
情報を高精度に検知

一部加工や
改変も検知

技術 ② 真正性証明（電子透かし）

正式発信者からの情報にウォーターマークを
埋め込み「正しい情報」を保証

「本物の証」である
電子透かしを付与

市民が信頼できる情報を
判断できる仕組みを構築

2-3. 開発技術により対処可能なユースケース

電話音声フェイク検知

本開発・実証のユースケースとして、音声フェイクによるなりすまし・詐欺の看破を想定する。
音声フェイク検知エンジンは単体で様々なユースケースをカバー可能である。

想定ユースケース 1	想定ユースケース 2	想定ユースケース 3
オンライン会議を使用する企業 機密情報漏洩・なりすまし防止	**証券会社等の機密情報を扱う企業 VIP取引の安全性保証**	**高齢者等の特殊詐欺対策（toC） なりすまし電話の被害防止**
実際の被害事例 1	**実際の被害事例 2**	**実際の被害事例 3**
企業幹部なりすまし Arup社（香港）：CFO含む幹部全員のDF映像+音声でビデオ会議を偽装、約37億円詐取(1)	**金融機関への 音声クローン詐欺** UAE銀行: 取引先ディレクターの音声クローンで支店長を欺き約50億円送金(2)	**知人の声クローン詐欺** 日本企業：社長のAI音声+発信番号偽装で送金詐欺未遂(3)。わずか3秒で85%精度のクローンが可能(4)
本技術による対処 1	**本技術による対処 2**	**本技術による対処 3**
通話中のリアルタイム検知+ アラート通知	**AI生成音声を検知し、 不正な指示を未然に防止**	**着信時のバックグラウンド検知で 利用者に即時警告**

(1) CNN | https://www.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk
(2) Forbes / Dark Reading | https://www.darkreading.com/cyberattacks-data-breaches/deepfake-audio-scores-35-million-in-corporate-heist
(3) Proofpoint Japan | https://www.proofpoint.com/jp/email-and-cloud-threats/voice-phishing-with-AI-learning-ceos
(4) McAfee "Beware the Artificial Impostor" | https://prtimes.jp/main/html/rd/p/000000032.000033447.html

2-3. 開発技術により対処可能なユースケース

自治体向け偽・誤情報総合対策

本開発・実証のユースケースとして、災害や犯罪発生時の偽・誤情報蔓延防止と正しい情報の信頼性確保を想定する。真贋判定技術は単体で様々なユースケースをカバー可能である。

想定ユースケース 1	想定ユースケース 2	想定ユースケース 3
有事にSNS上の情報の真偽を判定し、自治体の初動判断を支援	**選挙期間中のディープフェイク・偽情報の真偽判定**	**自治体の公式発信に真正性を付与し、改ざん・なりすましを防止**
実際の被害事例 1	**実際の被害事例 2**	**実際の被害事例 3**
災害時の偽情報混乱 能登半島地震で救助要請の約1割がデマ(1) 避難所デマは130万回以上閲覧(2)	**選挙時のディープフェイク拡散** 2024年米大統領選でバイデン氏のなりすまし自動音声が有権者に拡散(3)。国内でも岸田首相の顔入替動画が拡散(4)	**公式情報の信頼性低下** 偽情報の氾濫により47.7%が偽・誤情報を正しいと認識(5)。 自治体の打消し投稿が機能せず
本技術による対処 1	**本技術による対処 2**	**本技術による対処 3**
SNS上の画像・動画・テキストを複合的に解析し、真偽を判定	**AI生成コンテンツの判定+ファクトチェックエージェントによる真偽分析**	**電子透かし+DID/VCで公式発信者を証明し、転載後も真正性を検証可能に**

(1) 総務省 令和6年版情報通信白書 | https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r06/html/nd122c00.html
(2) 日本ファクトチェックセンター | https://www.factcheckcenter.jp/fact-check/disasters/false-kanazawa-1-5-evacuation-shelter-disaster-certificate/
(3) BBC | https://www.bbc.com/japanese/68065455
(4) 読売新聞オンライン | https://www.yomiuri.co.jp/national/20231103-OYT1T50260/
(5) 総務省「ICTリテラシーに係る実態調査」 | https://www.soumu.go.jp/menu_news/s-news/01ryutsu05_02000176.html

2-3. 開発技術により対処可能なユースケース

各種技術の横展開ユースケース

すでに示した主要ユースケースに加え、本事業で開発した各技術は単体で幅広い領域に適用可能である。
以下に各技術の横展開ユースケースを示す。

	音声フェイク検知	画像・映像フェイク検知	ファクトチェックエージェント
検知対象	AI生成音声	AI生成・加工された画像・映像 （部分加工含む）	テキスト情報の事実関係
技術的特長	CODEC変換・ノイズ下でも 高精度に判定	Inpainting等の 部分加工にも対応	複数ソースを横断して 自律的に真偽を分析
横展開先	選挙の世論操作対策、 コンタクトセンターの 音声認証強化	軍事・安全保障の情報操作対策、 個人への脅迫・名誉毀損対策	企業の風評被害対策、 フィッシング詐欺対策
想定顧客	通信事業者 金融機関 選挙管理機関	メディア 法執行機関 プラットフォーマー	企業（法務・広報） 金融機関 報道機関

上記の各技術はNABLASの商用サービス（KeiganAI）として既に提供中であり、顧客のニーズに応じて単体でも組み合わせでも導入可能。本事業の開発・実証を通じて、以下の技術強化により横展開の適用範囲がさらに拡大。

- **音声フェイク検知**：電話環境特有の環境下での検知精度を向上。従来対応できなかった実通話環境への適用が可能に。
- **偽誤情報総合対策**：画像・映像フェイク検知およびファクトチェックエージェントを実務フローに適合させ、実運用可能な形に統合。
- **電子透かし・DID/VCによる真正性証明**：真偽判定に加え、発信者の真正性検証まで一貫して提供可能に。「検知」と「証明」の両面から情報の信頼性を確保

目次

3-1. 技術開発の全体像

電話音声フェイク検知

電話環境で高精度にフェイク音声を検知し、次々と進化する生成AIにも継続的に追随できるシステムを実現するため、以下の3つのテーマで技術開発を実施

	Thema 1 検知モデルの高度化	Thema 2 最新生成AI追随の仕組み	Thema 3 電話サービスへの組込
取組概要	**電話特有の環境下でも精度を維持するEnd-to-Endモデルを開発。 従来の特徴量抽出型では対応困難だった実通話環境への適用を実現**	**ElevenLabs等の最新生成モデルに対して迅速に検知性能を評価し、モデルを改善できる体制を整備**	**電話など多くの人が使う基本的なサービスに組み込み、利用者が意識せず恩恵を受けられる形を目指す。 リアルタイム処理に対応した軽量かつ高精度な検知エンジンを設計**
成果	✓ End-to-Endモデルを導入し、直接フェイク判定する仕組みを実現 ✓ 電話特有のCODEC変換（μ-law / AMR-WB / opus）に対応した検知を実現 ✓ ひかり電話環境においてElevenLabs v2生成音声の検知に成功	✓ ElevenLabs v2を含む音声生成モデルに対して検知性能を評価 ・新たな生成モデルの出現時に検知性能を迅速に評価できる体制を整備 ✓ 生成モデル・電話環境ごとの検知精度差異を分析し、改善方針を策定	✓ NTT東日本ひかり電話と連携したフェイク検知アプリを開発 ✓ 量な検知エンジンを設計し、リアルタイム検知処理を実現 ✓ ひかり電話・携帯電話・IP電話の3パターンの発着信環境で動作を検証

End-to-End モデル: 音声波形から直接フェイク判定を行う深層学習モデル
CODEC変換（μ-law / AMR-WB / opus）：電話回線で音声を圧縮・伝送する際の符号化方式

3-1. 技術開発の全体像

自治体向け偽・誤情報総合対策システム

SNS上の偽・誤情報に対し、「フェイク検知」と「発信者証明」の両面から信頼性を確保するため、基盤技術を自治体が実運用できる総合システムとして統合

技術目標

目標	内容	対応する基盤技術
信頼できる発信者の確認	公的な発信者が正しい情報を投稿していることを確認できる	電子透かし ＋ DID/VC
フェイクコンテンツの検出	悪意ある投稿者によるフェイク投稿を見破る	画像・映像フェイク検知 ＋ ファクトチェックエージェント
転載後の真正性判定	転載・拡散された情報が改ざんされていないか判定できる	電子透かし検証 ＋ DID/VC検証

基盤技術と主な成果

基盤技術	取組概要	主な成果
画像・映像フェイク検知	AI生成・加工された画像・映像（Inpainting（部分加工）含む）を検知	自治体の実務に適合する形でシステムに統合。実証でフェイク画像を全て正しく識別
ファクトチェックエージェント	複数ソースを横断してテキスト情報の事実関係を自律的に分析	複数ソース横断の事実判定機能を実装。文章のみのフェイク投稿も全て判定可能。自治体向けの情報取得を優先的に実施する仕組みを構築
電子透かし（WM）	画像に目に見えない形で発信者情報を埋め込み、改ざん検知を可能にする技術を新規開発	WM埋込・検証機能を開発。ファイルサイズ増加の改善（740%→27%）は実証フェーズで実現
DID/VC	中央管理者なしに発信者の真正性を証明できるデジタル証明書技術を新規開発	認定者機能・証明書埋込機能・証明書検証機能の3機能を開発。伊那市実証で運用検証済み

Inpainting: 画像の一部領域をAIで生成・置換する部分加工技術
DID/VC: 分散型ID（Decentralized Identifier）/ 検証可能な資格情報（Verifiable Credential）。発信者の身元を第三者なしに証明する技術
電子透かし（WM）：コンテンツに不可視の識別情報を埋め込み、改ざんや出所を検証する技術
ファクトチェックエージェント: 複数の情報ソースを自動巡回し、テキストの事実関係を判定するAI技術

3-2. 技術開発の個別詳細

電話音声フェイク検知 - 技術課題とKPI -

電話環境で高精度にフェイク音声を検知し、最新の生成AIにも継続的に追随できるシステムの技術基盤を開発するため、以下の3つの技術課題に取り組んだ。

技術課題	背景・現状の限界	当初目標	達成状況
電話環境に適応した検知モデルの構築	電話回線固有のCODEC（音声圧縮方式）変換・ノイズにより、通常のフェイク音声検知モデルでは精度が大幅に低下する。従来の特徴量抽出型では前処理段階で情報が損失し、電話環境への適用が困難。	電話環境下でのフェイク判定精度90%	End-to-End（一気通貫型）モデルを開発。CODEC変換（μ-law / AMR-WB / opus）を含む電話環境下でのフェイク検知手法を確立。（検証結果は4章に記載）
最新生成AIへの継続的追随	音声生成技術は急速に進化し、ElevenLabs等の高品質な音声生成モデルが次々と登場する。検知モデルの改善は終わりのない取組であり、迅速に対応できる体制が不可欠。	新モデル出現時に迅速に検知性能を評価・改善できる体制の構築	データセントリックAIに基づく改善基盤システムを開発。データ自動生成→学習→評価→デプロイの自動化により、新技術出現時に迅速に対応可能。（評価結果は4章に記載）
電話サービスへの組込	フェイク音声検知を電話サービスに組み込むには、通話中にリアルタイムで処理可能な軽量かつ高精度な検知エンジンが必要。	リアルタイム処理に対応した検知エンジンの設計・実装	リアルタイム処理に対応した軽量検知エンジンを設計。任意の電話サービスに適用可能な汎用アーキテクチャを構築。（アプリ実装は5章に記載）

End-to-End（一気通貫型）モデル: 音声波形から直接フェイク判定を行う深層学習モデル。従来の特徴量抽出型と異なり前処理が不要
CODEC（音声圧縮方式）: 電話回線で音声を圧縮・伝送する際の符号化方式。回線種別により異なり検知精度に影響する

3-2. 技術開発の個別詳細

電話音声フェイク検知 - End-to-Endモデルの設計・開発 -

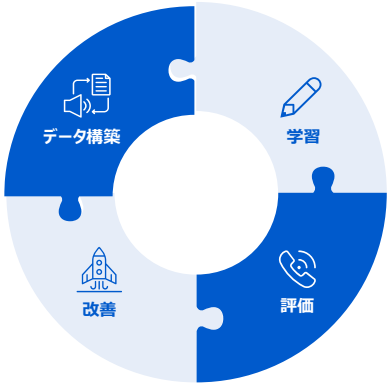
電話回線固有のCODEC変換・ノイズ環境下でもフェイク音声を高精度に検知するため、既存の検知モデルを電話環境のCODEC変換に対応するよう拡張・改良したEnd-to-End（一気通貫型）モデルを開発した。データセントリックAI（データの改善重視）のアプローチを採用し、モデルの評価・学習・実装の自動化を進める基盤システムを開発した。

End-to-Endモデル

- **電話環境への拡張**
  既存の検知モデルを、実際の電話環境で発生するCODEC変換を踏まえて拡張・改良
- **前処理不要化**
  モデル自体の改良により、前処理なしで電話音声を直接入力しフェイク判定が可能
- **データセントリックAI**
  モデルのアーキテクチャ変更ではなくデータの改善を重視。生成AIを活用してフェイクデータを自動生成・学習し、常に最新の検出精度を維持

軽量検知エンジン

- 推論速度と検知精度のトレードオフを最適化した軽量エンジンを設計
- 特定の電話サービスに依存しない汎用アーキテクチャ（任意の電話サービスに適用可能）
- 最も評価スコアが高い検出モデルをクイックにアプリ・サービスに搭載



End-to-Endモデル × 軽量検知エンジン

最新の検知モデルをAPIとしてクイックにアプリケーションに実装

3-2. 技術開発の個別詳細

電話音声フェイク検知 - CODEC変換対応の学習データ構築 -

電話回線の種類により圧縮方式・サンプリングレート・帯域幅が異なり、フェイク検知に有用な音声特徴の劣化パターンが異なる。当初は各ルートのCODEC変換をそのままシミュレーションしたが精度が向上しなかったため、CODEC変換を構成要素に分解し、要素ごとにデータ拡張を行う手法を開発した。

CODEC環境の違いと課題

発着信パターン	CODEC	特性	フェイク検知への影響
ひかり電話→ひかり電話	μ-law → μ-law	8kHz固定・狭帯域（300～3400Hz）・波形ベース圧縮	同一CODECのため情報損失が最小。比較的検知しやすい。
携帯電話→ひかり電話	AMR-WB/NB → μ-law	16kHz/8kHz・フレーム単位圧縮・ビットレート可変	異CODEC間変換で二重の劣化が発生。帯域制限による情報損失。
IP電話→ひかり電話	opus → μ-law	可変サンプリングレート・可変帯域・超低ビットレートあり	異CODEC間変換に加え、環境ごとの変動幅が大きく対応困難。

課題

- 各CODECで圧縮方式・サンプリングレート・帯域幅が異なり、フェイク検知に有用な音声特徴の劣化パターンが異なる。ルート単位でCODEC変換をシミュレーションしても検知精度が向上しなかった。
  → CODEC変換をルート単位ではなく構成要素に分解し、要素ごとに学習データを構築するアプローチに変更

学習データの内容

構成要素	構成要素
周波数	サンプリングレート変換、帯域幅制限（狭帯域～全帯域）
圧縮	量子化・ビットレート制限による情報量削減
時間	Jitter（微小な時間伸縮）、フレーム欠落
ノイズ	付加ノイズ（環境音等）、ノイズキャンセル処理
追加要素	Clipping/Saturation（録音歪み）、Reverb（部屋鳴り）

- 要素ごとに独立してデータ拡張を適用し、全体のデータで網羅的にカバー
- 特定のルートに依存しない汎用的なデータ拡張設計により、未知の電話環境にも対応可能

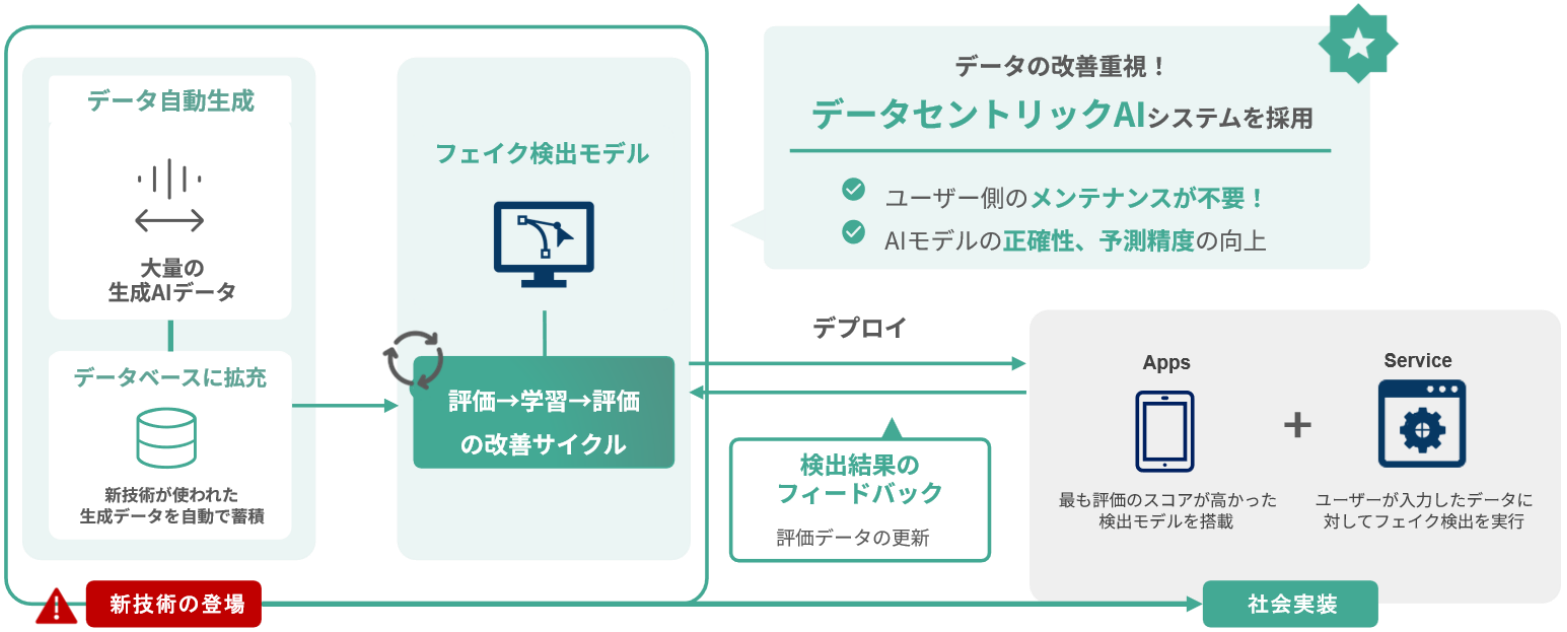
3-2. 技術開発の個別詳細

電話音声フェイク検知 - 継続的改善の基盤システム -

生成AIを活用してフェイクデータを生成・学習し、常に最新の検出精度を維持する。新たな生成技術にも迅速に対応できるよう、モデルの評価・学習・実装の自動化を進める基盤システムを開発した。

データセントリックAIによる改善サイクル

- ✓ **データの改善重視**
  モデルのアーキテクチャ変更ではなく、学習データの質・量の改善によって検出精度を向上させるアプローチ
- ✓ **データ自動生成**
  最新の生成AIを活用してフェイクデータを生成し、データベースに蓄積。新技術が登場すると自動で生成データを蓄積
- ✓ **評価→学習→評価の自動化**
  新データに基づきモデルを更新し、評価を実施する循環プロセスを自動化
- ✓ **検出結果のフィードバック**
  実際の検出結果を評価データの更新に活用



3-2. 技術開発の個別詳細

電話音声フェイク検知 - 技術開発の成果と課題 -

本事業では電話環境でのフェイク音声検知に必要な技術基盤を開発した。同時に、開発過程で今後の改善に向けた技術課題を明確にした。

技術開発の成果

End-to-Endモデルの開発
既存モデルを電話環境のCODEC変換に対応するよう拡張・改良。前処理不要で電話音声を直接入力可能にし、電話環境下でのフェイク検知手法を確立。

CODEC要素分解による学習手法の確立
当初のルート単位シミュレーションでは精度が向上しなかった課題に対し、CODEC変換を5つの構成要素（周波数/圧縮/時間/ノイズ/追加要素）に分解した学習データ構築手法を開発。

軽量検知エンジンの設計
リアルタイム処理に対応した軽量エンジンを設計。任意の電話サービスに適用可能な汎用アーキテクチャを構築。

改善基盤システムの構築
データセントリックAIに基づき、データ自動生成→学習→評価→デプロイの自動化基盤を開発。ユーザー側メンテナンス不要。

明らかになった技術課題

△ **シミュレーション vs 実環境の差異:**
CODEC変換の要素分解により精度は改善したが、シミュレーションと実際の電話網を通った音声にはなお差異がある。エコーキャンセル処理等、実環境固有の要素への対応が必要。

△ **CODEC間の精度差**
同一CODEC（μ-law→μ-law）では検知に成功するが、異CODEC変換（AMR-WB/opus→μ-law）では精度が低下。各CODEC特有の信号劣化への対応が必要。

△ **生成モデルの急速な進化**
ElevenLabs等の生成モデルが短期間で高品質化。構築した改善サイクルで継続的に追随する必要がある

- 各課題に対する定量的な検証結果（精度数値、分析データ等）は4章（検証及び調査）で記載する。
- アプリ開発成果は5章（社会実装）で記載する。

3-2. 技術開発の個別詳細

自治体向け偽・誤情報総合対策システム - 技術課題とKPI -

現状の画像・映像の生成AIコンテンツ検出技術、テキスト情報のファクトチェック技術、電子透かしを始めとした発信者情報の信頼性担保技術には以下の課題が残されている。これら6つの課題に取り組んだ。

技術課題	背景・現状の限界	当初目標	達成状況
画像の部分加工（Inpainting）検知	画像全体を生成する手法では高い検出精度を実現できているが、画像内の一部分のみを加工するInpaintingへの対応が進んでいない。	Inpainting含むAI加工画像の検知精度90%以上	Inpainting対応検知モデルを開発しシステムに実装。ベンチマーク評価を実施
最新生成技術への継続的追随	検出技術と生成技術のいたちごっこが発生する。常に最新の生成技術に追随して検出モデルをタイムリーにアップデートする仕組みが必要。	新モデル出現時に迅速に検知性能を評価・更新できる体制の構築	新しいモデルに対して評価・更新できる体制を構築
ファクトチェックにおけるソース信頼性判断	ファクトチェック時に参照する情報源の信頼性を判断した上で総合的にファクトチェックできる技術が存在しない。	ソース信頼性を加味した総合的ファクトチェック。誤情報検出率90%以上	複数ソース横断の事実判定機能を実装。ソース信頼性の加味およびモデル改善は本事業スコープ外
電子透かし（WM）の実用化	質に影響を与えず改ざんが難しく確実に検出できるWMが存在しない。SNSへのアップロード・ファイルの複製・加工・スクリーンショットで情報が失われ機能不全となる。除去技術への対抗も進んでいない。	SNS投稿後も復元可能かつ実用的なファイルサイズのWM技術を開発	WM埋込・検証機能を新規開発。SNS投稿後も復元可能な技術を実装
既存発信者証明技術（OP等）の限界克服	OP技術等はHTMLベースに対象が閉じており、活用環境に制限がある。画像・映像ファイルが掲載元URLなどを明示せずに個別に再編集・拡散されると効力を失する	画像・映像ファイル自体に発信者証明を埋込可能なDID/VC基盤の新規構築	3機能実装完了（認定者・埋込・検証）
SNS上で利用可能な総合システム	真正性を保証・確認できるシステムとして、SNS上でユーザが利用しやすいソリューションが現状存在しない	基盤技術を統合し自治体が実務で利用可能なシステムを構築	基盤技術統合Webシステム構築完了

Inpainting（部分加工）：画像の一部領域をAIで生成・置換する加工技術
OP技術：Originator Profile。Webコンテンツの発信者情報を付与する技術。HTMLベースのため画像・映像単体の拡散時に機能しない
DID/VC：分散型ID / 検証可能な資格情報。OP技術の限界を克服し、画像・映像ファイル自体に発信者証明を埋め込む技術

3-2. 技術開発の個別詳細

自治体向け偽・誤情報総合対策システム - WM + DID/VC（証明・検証の統合） -

公的な発信者が投稿する情報にWMとDID/VCを埋め込み、転載・拡散後も「誰が投稿した原本か、改ざんされていないか」を検証可能にする技術を新規開発した。

証明フロー（情報発信側）

STEP1 **認定者（自治体管理者）**
投稿者をDID/VCで認定

STEP2 **投稿者（認定済み）**
画像をシステムにアップロード

STEP3 **システム**
WM埋込 + VC（証明書）埋込を自動実行

STEP4 **投稿者**
WM+VC入り画像をSNS（LINE/X等）に投稿

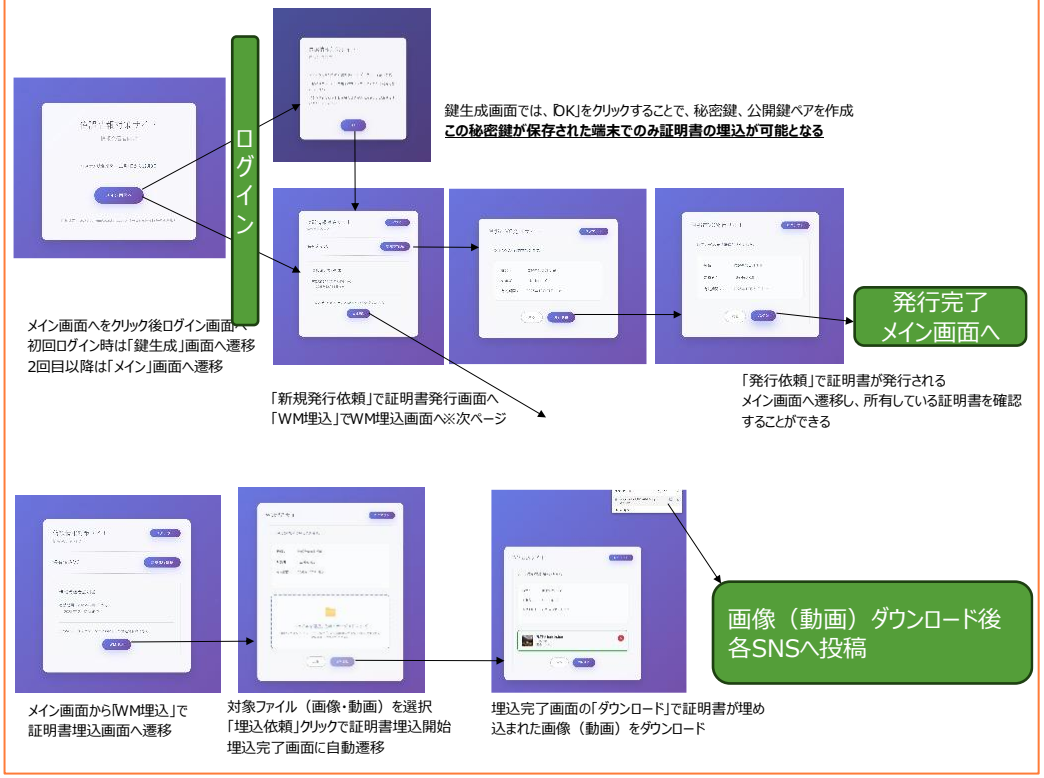
検証フロー（情報検証側）

STEP1 **閲覧者/検証者**
SNS上の画像を取得

STEP2 **システム**
WM抽出・復元 + VC検証を実行

STEP3 **閲覧者/検証者**
信頼度スコアとともに「認定発信者の原本か」を判定

WM + DID/VC埋め込みアプリUI



3-2. 技術開発の個別詳細

自治体向け偽・誤情報総合対策システム - 画像・映像フェイク検知 -

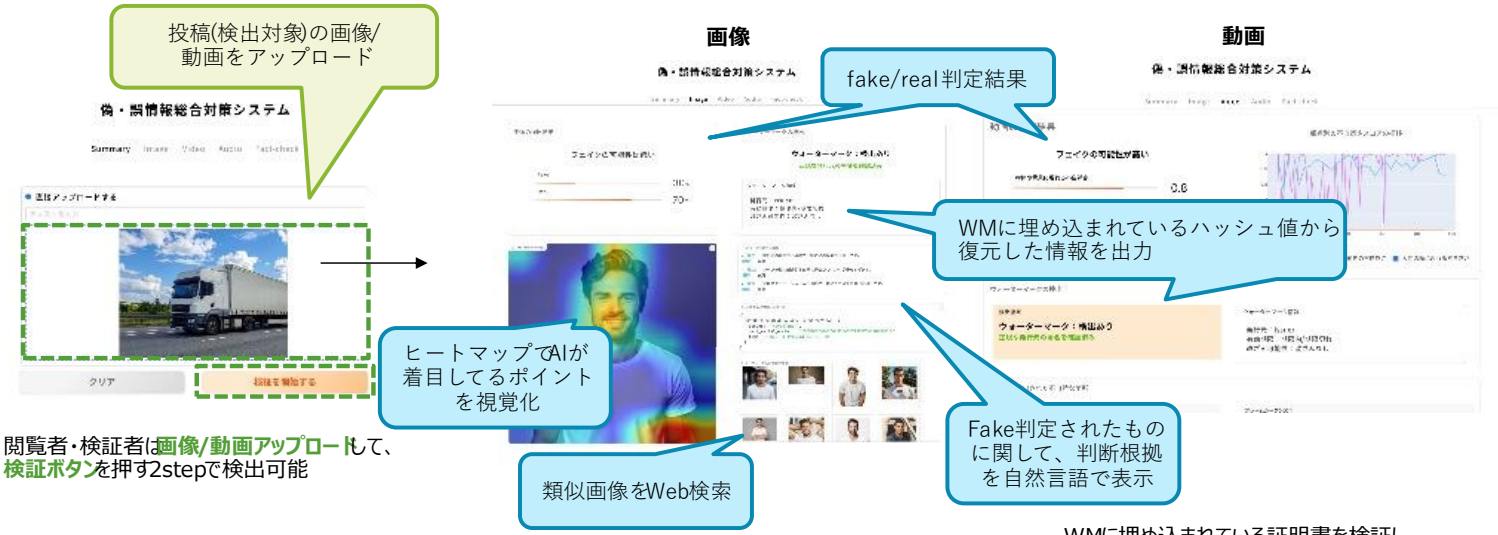
AI生成・加工された画像・映像を検知する技術を開発。全体生成だけでなく、画像の一部をAIで加工するInpainting（部分加工）にも対応し、複数モデルをアンサンブルすることで、検出の網羅性を強化した。

画像フェイク検知

- 画像全体の生成だけでなく、部分加工（Inpainting）にも対応する検知モデルを開発
- WMの改ざんがある場合、改ざん位置を可視化
- 開発した検知モデルをシステムに統合し、自治体が利用可能な形で実装

映像フェイク検知

- 既存検知モデルのアンサンブル（複合活用）による検知性能を評価
- 動画の時間軸ごとにスコア可視化
- 怪しいフレームについて可視化し、自然言語でレポートを表示



3-2. 技術開発の個別詳細

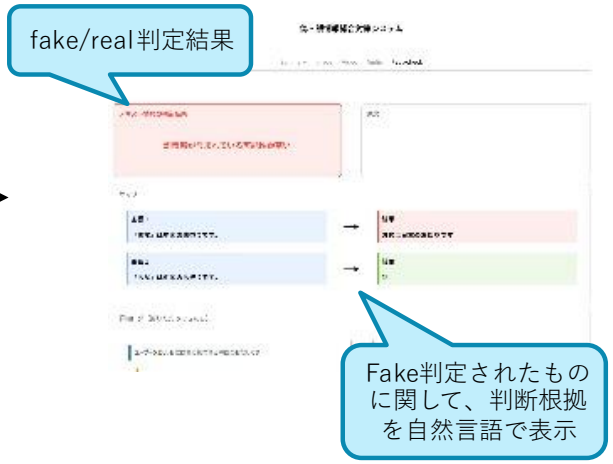
自治体向け偽・誤情報総合対策システム - ファクトチェックエージェント -

画像・映像を含まないテキストのみの偽情報にも対処するため、複数の情報ソースを自動巡回し、事実関係を自律的に判定するファクトチェックエージェントをアプリケーションとして構築した。

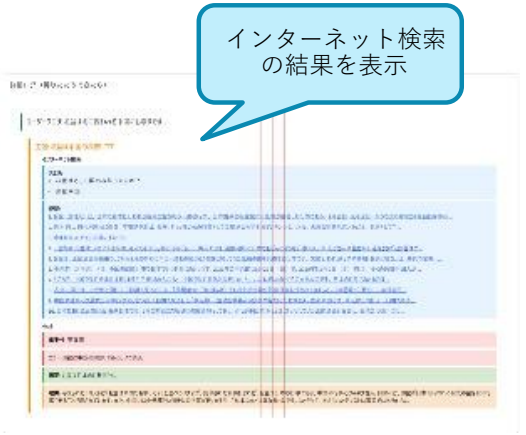
技術構成

- ✓ **Reasoningモデル**：テキストの主張を分解し、各主張に対する根拠を複数ソースから収集・評価
- ✓ **情報ソースのカスタマイズ**：取扱いユースケースに応じて、参照先の情報を制限・除外等のカスタマイズが可能
- ✓ **判定結果の透明性**：根拠となる情報源を提示し、検証プロセスの透明性を確保

閲覧者・検証者は **テキストアップロード**して、
検証ボタンを押す2stepで検出可能



文章のフェイク（AIによって生成・編集されたもの）を検出



文章に対してインターネット検索を行い、真偽を判定。
詳細ログにて判断詳細を表示。

3-2. 技術開発の個別詳細

自治体向け偽・誤情報総合対策システム - 総合システム化 -

画像・映像フェイク検知、ファクトチェックエージェント、電子透かし（WM）、DID/VCの4技術を、自治体職員が一貫して利用できるWebシステムとして統合した。

情報発信フロー（証明）

DID/VC 認定投稿者管理

- ✓ 認定投稿者への証明書（VC）発行・管理
- ✓ ブロックチェーンベースのレジストリで真正性を担保
- ✓ 投稿→閲覧→拡散後も検証可能

電子透かし（WM）埋込・検証

- ✓ 公式画像へのウォーターマーク自動埋込
- ✓ 画像の改ざん有無を検出

情報検証フロー（検知）

画像・映像フェイク検知・ファクトチェック

- ✓ 対象コンテンツをアップロード → 自動で真偽判定
- ✓ Inpainting（部分加工）を含む複数の改変手法に対応
- ✓ 判定結果を信頼度スコアとともに表示

ファクトチェックエージェント

- ✓ テキスト情報を入力 → エージェントAIが複数ソースから自動検証
- ✓ 根拠となる情報源を提示し、検証プロセスの透明性を確保

成果

- 4技術を統一UIで操作可能なWebシステムとして統合完了
- 情報発信（証明）と情報検証（検知）の一貫したフローを実現
- 伊那市と2回の実証実験で実用性を検証

検出結果をサマリーレポートとして一覧で表示

目次

4-1. 検証及び調査の全体像

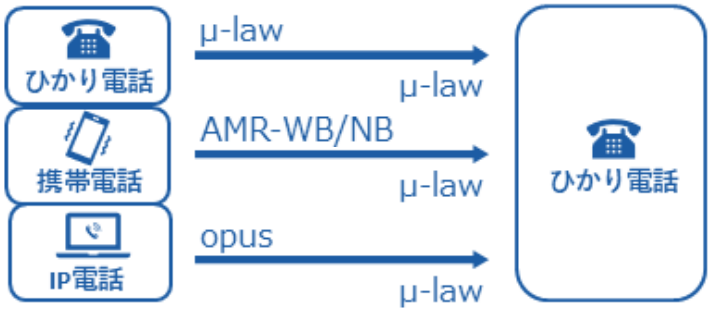
電話音声フェイク検知

3章で開発したEnd-to-Endモデル・CODEC要素分解による学習手法・改善基盤システムの有効性を検証するため、以下の検証を実施した。

検証概要

検証項目	検証内容
発着信パターン別の検知精度	3パターン（ひかり/携帯/IP）の電話環境で、フェイク音声の検知精度を測定
CODEC環境ごとの精度差異分析	同一CODEC vs 異CODEC変換での精度差を分析
シミュレーション vs 実環境の差異分析	シミュレーションで構築した学習データと、実際の電話網を通った音声での精度差を検証
要素分解アプローチの改善効果	ルート単位シミュレーション（ベースモデル）→ 要素分解（再学習モデル）での精度改善を確認
オープンソースモデルとの精度比較	他のオープンソースモデルと自社開発モデルの検知精度を比較分析

NTT東日本の実際の光回線ルートを使用し、実インフラ環境での検証を実施



コーデック変換・エコーキャンセル処理を含む実環境での検証

↓

End-to-Endモデルを開発

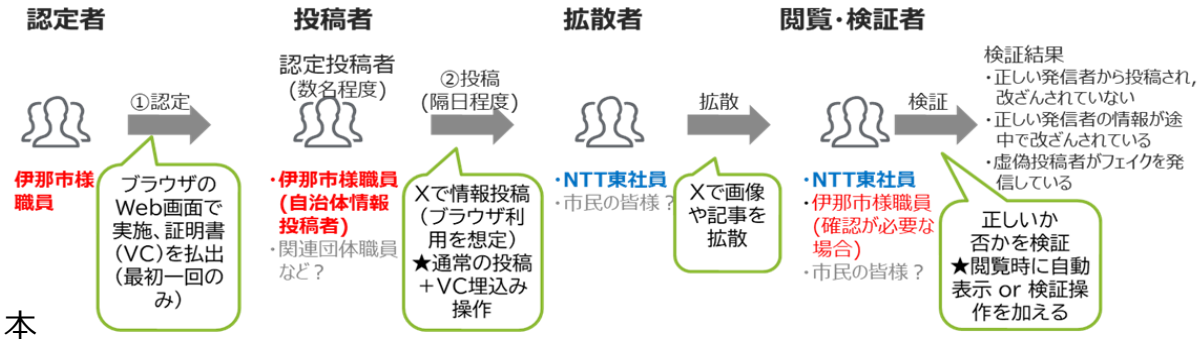
4-1. 検証及び調査の全体像

自治体向け偽・誤情報総合対策

開発した画像・映像フェイク検知、ファクトチェックエージェント、電子透かし（WM）、DID/VCの各技術について、単体技術検証と伊那市での実証実験を実施した。

検証概要

検証項目	検証内容	目標KPI
画像フェイク検知	AI生成・加工画像（Inpainting含む）の検知精度評価	検知精度90%以上
ファクトチェックエージェント	複数ソース横断による事実判定の精度評価	判定精度指標
電子透かし（WM）	SNS（LINE/X）経由での復元率、目視判別率、ファイルサイズ影響	復元率・目視判別率
DID/VC	認定者機能・証明書埋込・検証の動作確認	機能動作確認



実証実験の概要

実証先：長野県伊那市, NTT東日本
実証期間：実証1（10/4-11/4）、実証2（1/21-1/27）
実証方法：伊那市公式アカウントからWM入り画像をSNS投稿 → NTT東検証アカウントで虚偽/一般投稿 → 真偽判定
実証結果概要：フェイク検知率100%*（目標90%）

*KPI実績値は伊那市実証（n=66）における達成値

4-2. 検証及び調査の個別詳細

電話音声フェイク検知 - 検証概要と評価指標 -

電話環境における音声フェイク検知の有効性を検証するため、NTT東日本の実電話網を使用した検証を実施した。
評価指標にはF1スコアを採用し、フェイク音声の見逃しと誤検知の両面からバランスの取れた評価を行った。

検証の全体設計

項目	内容
検証目的	電話環境下でのフェイク音声検知精度の評価。シミュレーションではなく実電話網での検証
当初目標	最新生成AI（ElevenLabs等）による電話環境下フェイク判定精度90%
検証環境	NTT東日本のひかり電話実網。3パターンの発着信環境で検証
検証対象モデル	自社End-to-Endモデル + OSSモデル（同一データセットで比較評価）

評価指標: F1スコア, BAcc（均衡精度）

指標	定義	意味
Precision（適合率）	「フェイク」と判定した音声のうち、実際にフェイクである割合	高いほど誤検知（正常な通話をフェイクと判定）が少ない
Recall（検出率）	実際のフェイク音声のうち、正しく「フェイク」と判定できた割合	高いほどフェイク音声の見逃しが少ない
F1スコア	PrecisionとRecallの調和平均: 2×P×R/(P+R)	両指標のバランスを取った総合評価
BAcc（均衡精度）	Real正解率とFake正解率の平均: (Recall_Real + Recall_Fake) / 2	両クラスの検知性能を均等に評価

F1スコアとBAccを併用する理由

- ✓ 電話フェイク検知では、フェイクの見逃し（詐欺被害）と正常通話の誤検知（不要な警告）の両方が問題となる
- ✓ **F1スコア**: Precision/Recallの両面を1つの指標で評価でき、フェイク検出の総合性能を示す
- ✓ **BAcc**: F1はFakeが多数派の場合に全件Fake判定でも高い値が出るため、Real/Fake両クラスの正解率を均等に評価するBAccで補完する
  BAcc=50%はランダム判定と同等であり、モデルが実質機能していない状態を明確に示せる

4-2. 検証及び調査の個別詳細

電話音声フェイク検知 - 実験条件とデータセット -

実際の詐欺シナリオを想定したデータセットを構築し、NTT東日本の実電話網を通した3パターンの発着信環境で検証を実施した。

データセット

項目	内容
生成モデル	ElevenLabs v2（高品質音声生成AI）
評価対象	Real音声 + Fake音声(v2)
発話内容	実際の詐欺内容を想定し、架空請求・詐欺の内容かつ男女の声で5種類を生成
回線別内訳	①ひかり→ひかり/ ②携帯→ひかり/ ③IP→ひかり
音声形式	wav（着信側: μ-law PCM, 8kHz）

3パターンの発着信環境

着信側はすべてひかり電話（μ-law, 8kHz）。発信側のコーデックが異なる3パターンで検証した。コーデック変換の詳細と検知難度への影響は5章に記載する。

発着信パターン	発信側	着信側
ひかり電話→ひかり電話	μ-law	μ-law
携帯電話(docomo)→ひかり電話	AMR-WB/NB	
IP電話→ひかり電話	opus	

実験条件

- **実電話網:** 実際の電話交換機・回線を経由。エコーキャンセル・回線ノイズ等の実環境要因が含まれる
- **最新生成AI:** ElevenLabs v2による高品質な合成音声

4-2. 検証及び調査の個別詳細

電話音声フェイク検知 - 音声検知モデル検証結果 -

自社開発のEnd-to-Endモデルを3パターンの発着信環境で評価した。電話回線全体でのF1スコアは0.727となり、当初目標の90%には未到達だが、回線種別により検知性能に大きな差異が確認された。

総合評価指標（電話3回線）

指標	スコア	意味
F1スコア	0.727	Precision/Recallのバランス指標
Precision（適合率）	0.750	Fake判定のうち実際にFakeだった割合
Recall（検出率）	0.706	実際のFake音声を正しく検出できた割合
BAcc（均衡精度）	0.686	Real/Fakeの各クラス正解率の平均

回線別 BAcc（均衡精度）

発着信パターン	BAcc	特徴
ひかり→ひかり	0.650	CODEC変換なしだが目標未達
携帯→ひかり	0.667	携帯CODECの帯域制限が影響
IP→ひかり	**0.750**	3回線中で最も高い性能

結果の分析

- **全回線で目標未達:** 当初目標90%に対し、最良のIP回線でもBAcc=75.0%にとどまる
- **Recall=70.6%:** Fake音声の約3割を見逃している状態
- **回線間の差異:** 同一CODEC（ひかり→ひかり）が最良とは限らず、IP回線（IP→ひかり）が最も高い性能を示した。CODEC変換の種類によって影響が異なることが示唆される
- **Precision=75.0%:** Fake判定の4件に1件はReal音声の誤検知（正常通話への不要な警告が発生）

4-2. 検証及び調査の個別詳細

電話音声フェイク検知 - 精度低下の原因分析と改善 -

検知精度低下の原因分析から、CODEC変換に伴うサンプリング周波数の変化により、分類モデルが依拠する音声特徴量が不可逆的に欠損することが主因と示唆された。学習データの最適化により、携帯→ひかり回線では本評価データセットにおいてはBAcc=100%を達成した。一方、単一モデルですべてのコーデック変換に対応することには限界があり、回線ごとに最適化したモデルの構築が必要となる可能性がある。

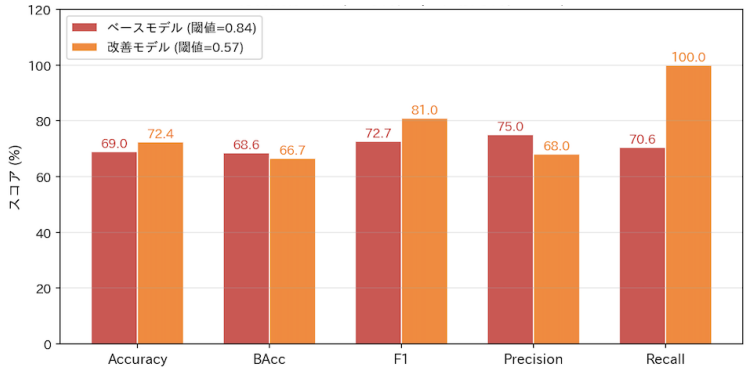
精度低下の根本原因 - CODEC変換による特徴量欠損 -

要因	メカニズム	影響
サンプリング周波数の変化	異CODEC変換時にサンプリング周波数が変化（例: opus 48kHz → μ-law 8kHz）。高周波帯域の情報が不可逆的に失われる	フェイク検知モデルが依拠する高周波特徴量が欠損しReal/Fakeの判別が困難に
圧縮方式の違い	各CODECの圧縮アルゴリズムが異なり、音声の微細な特徴が変換ごとに変質する	シミュレーションで学習した特徴と実環境の特徴にギャップが生じる
エコーキャンセル等の付加処理	実電話網固有の信号処理が介在	シミュレーションでは再現できない音声劣化が発生

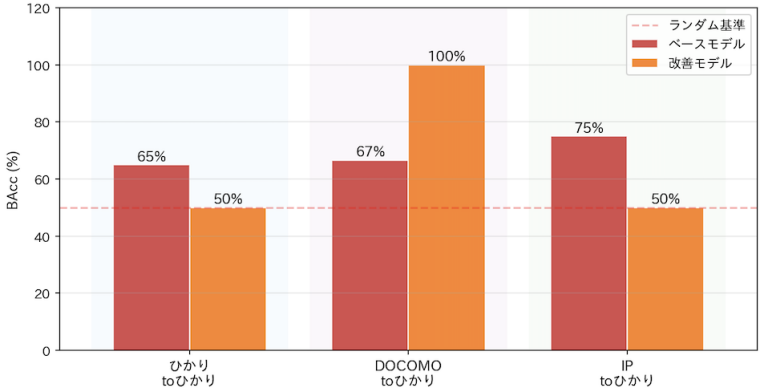
学習データ最適化による改善（ベースモデル → 改善モデル）

改善手法: CODEC変換の影響による要素を分解し、それぞれの要素ごとにAugmentationしたデータで再学習を実施

総合評価指標



回線別の改善効果（BAcc）



改善結果の解釈:

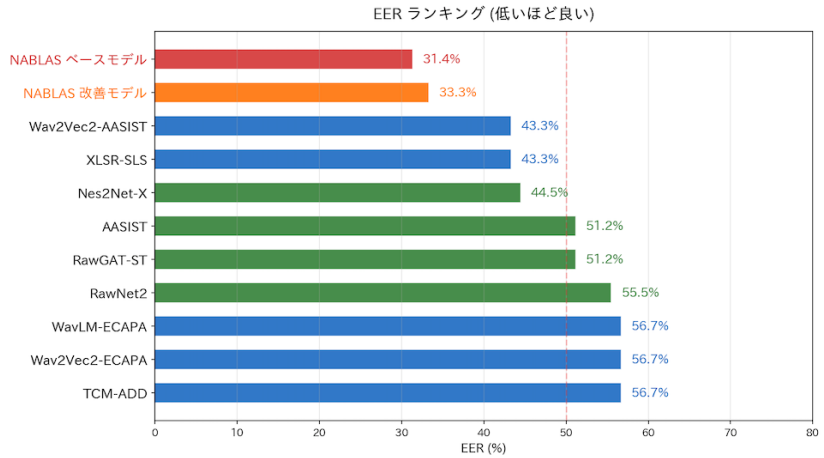
- **携帯→ひかり回線への適応に成功:** 学習データ最適化により、携帯→ひかり回線ではBAcc=100%を達成
- **他2回線ではBAcc=50%（ランダム同等）に低下:** Recall=1.000はFake検出力の向上ではなく、モデルがFake判定に偏った結果。Real音声を過剰にFakeと判定している
- **単一モデルの限界:** 1つのモデルですべてのコーデック変換に対応するのは困難であり、回線ごとに最適化したモデルの構築が必要となる可能性がある

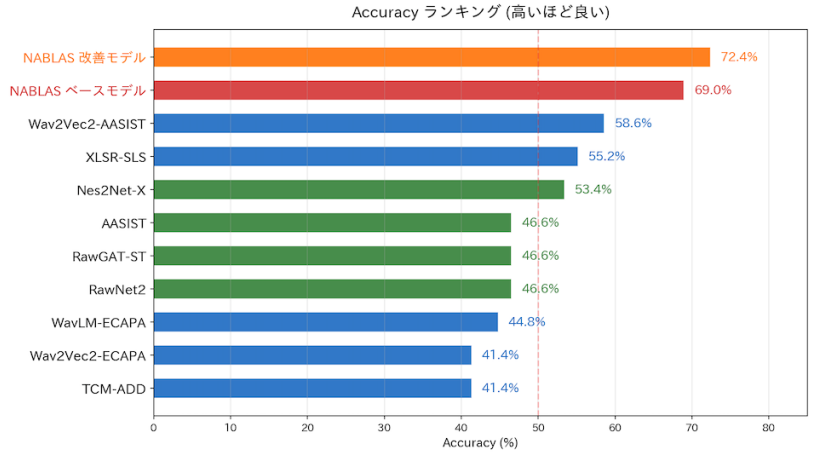
4-2. 検証及び調査の個別詳細

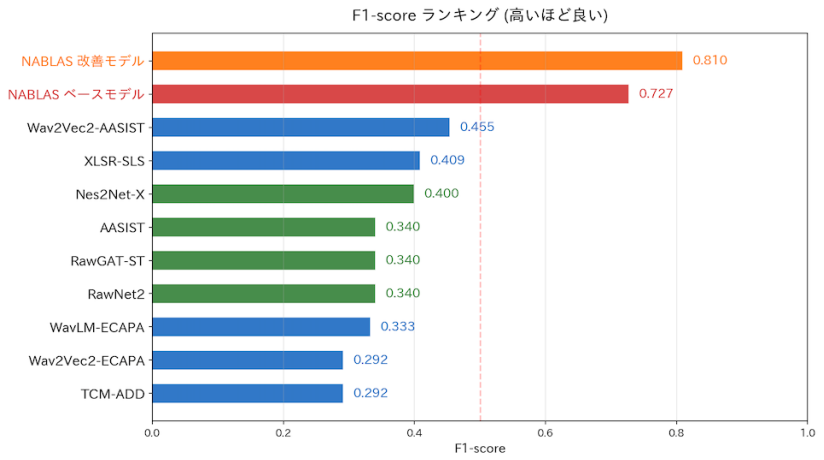
電話音声フェイク検知 - オープンソースモデルとの比較 -

自社モデルの検証結果を客観的に評価するため、Speech DF Arena（音声ディープフェイク検出の国際ベンチマーク）に掲載されている9種のオープンソースモデルを、同一のNTTデータセットで評価した。

比較結果







比較分析

- F1, Accuracy, EERのいずれにおいてもNNABLASモデルが最良
- F1でNABLAS 0.727~0.810 vs OSS最良0.455
- OSSモデルの多くがEER≥50%: ランダム判定以下の性能。電話環境でのフェイク検知がモデルの種類を問わず困難であることを示す

今後の課題

- 学習データ最適化により携帯→ひかり回線でBAcc=100%に改善
  - ただし改善効果は回線依存であり、回線ごとの個別最適化が必要
  - 根本原因はCODEC変換時の特徴量欠損。音声波形のみでは原理的に解けない問題の可能性がある
- 電話環境でのフェイク音声検知は世界的にも未解決の難問であり、OSSモデル9種との比較でこれを客観的に実証。本事業で実環境での検証を行い課題を明確化したことに技術的価値があると考える。

4-2. 検証及び調査の個別詳細

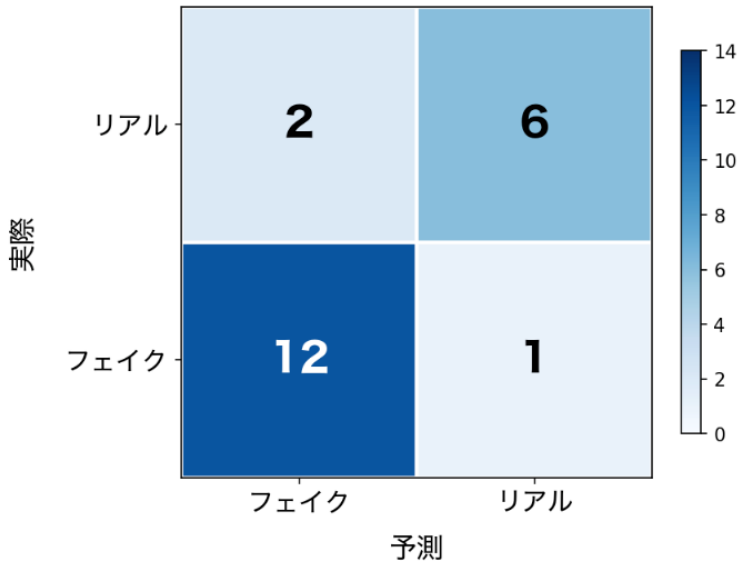
自治体向け偽・誤情報総合対策 - 画像フェイク検知 -

AI生成・加工画像（Inpainting含む）および最新映像生成モデルで作成されたフェイクコンテンツに対する検知精度を評価した。今回自治体と議論を重ね、直近のユースケースでは画像の投稿頻度が高いことがわかったため、画像中心に実ユースケースを想定したテストデータセットを作成し、検出を行った。

検証条件: 2シナリオ（熊/川）× 多様な生成手法 + 女川町(SNSから取得)、計21件

生成方法	正解率
AI完全生成（Gemini）	4/4
実写背景+AI合成（nano banana）	3/3
動画スクショ（Sora / Sora2 / Veo3）	3/3
生成手法不明画像(フェイク)	2/3
リアル画像（比較対象用）	6/8

Accuracy：85.7%
Recall：92.3%
Precision：85.7%



今後の課題

- **新規生成モデルへの迅速な追従**:新モデル出現時の学習データ追加・再学習パイプラインの自動化
- **リアル画像の誤検知（偽陽性）低減**:検知閾値の最適化、リアル画像の学習データ拡充、段階的判定の導入
- **映像生成モデルに対する詳細検証**:映像生成モデル別のテストセット構築と検知精度の定量評価
- **未知のドメインへの汎化性能**:多様性のある学習データの拡充、ドメイン適応技術の導入

4-2. 検証及び調査の個別詳細

自治体向け偽・誤情報総合対策 - ファクトチェックエージェント -

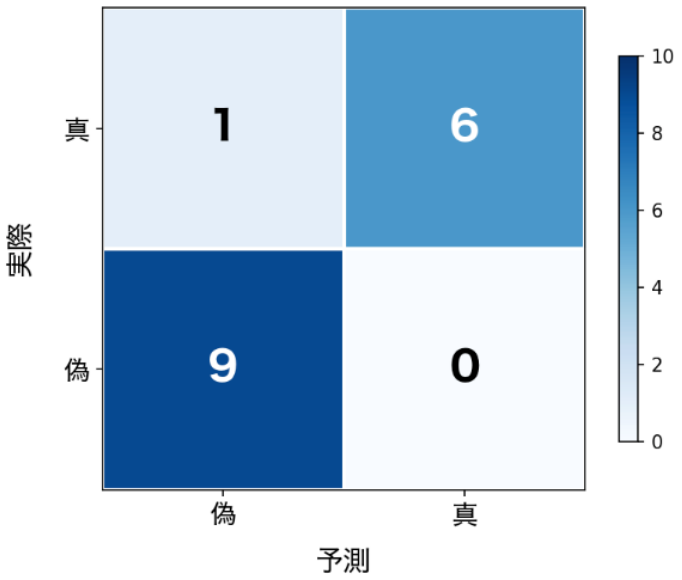
伊那市の実データ（熊目撃情報・気象警報等）を用いて、ファクトチェックエージェントの真偽判定精度を評価した。
3回の検証を通じてプロンプト・参照元の調整による改善を実施した。

検証条件: 伊那市の実データに基づく16件（偽の情報9件 + 真の情報7件）

- 熊目撃情報（場所・日時・動物種の真偽）
- 気象警報（濃霧注意報・洪水警報・大雨警報の真偽）
- 災害情報（河川氾濫の真偽）

検証項目	正解率
誤情報の検出（偽→偽）	9/9
正情報の判定（真→真）	6/7

Accuracy：93.8%
Recall：100%
Precision：90%



今後の課題

- **正情報の誤判定（偽陽性）低減**:参照元データの拡充、入力表現の揺れに対する判定のブレの強化
- **参照元の網羅性確保**:地域情報データベースとの連携、自治体公式情報の優先参照
- **情報鮮度への対応**:時系列を考慮した検索アルゴリズムの導入や信頼性の高い情報ソースの優先参照

4-2. 検証及び調査の個別詳細

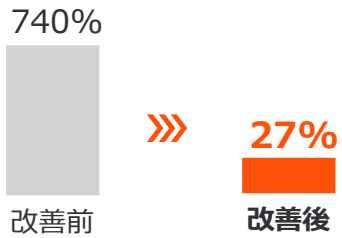
自治体向け偽･誤情報総合対策 - 電子透かし（WM）･DID/VC -

WMの埋込･検証機能およびDID/VCの3機能（認定者/証明書埋込/証明書検証）について、SNS（LINE/X）での技術検証を実施した。

検証条件:
3種ブラウザ環境 × 18種WM埋込画像 × 6種WM埋込動画

検証項目	正解率
SNS投稿後のWM検出･VC検証（画像）	18/18
SNS投稿後のWM検出･VC検証（動画）	6/6
伊那市X投稿のWM検出･VC検証	8/8
伊那市LINE投稿のWM検出･VC検証	5/9

ファイルサイズへの影響



WM埋め込み後のファイルサイズ増加740%→27%と実用可能なレベルに改善

今後の課題

- **SNS圧縮に対するWMのロバスト性向上**:LINE等の圧縮アルゴリズムに耐性のあるWM埋込方式の改善
- **小領域加工に対するWM検出のロバスト性強化**:局所的な改ざんを検出できるブロック単位のWM埋込･検証方式の導入
- **ファイルサイズ増加のさらなる抑制**:圧縮アルゴリズムの継続改善、画像･動画フォーマット別の最適化

4-2. 検証及び調査の個別詳細

自治体向け偽･誤情報総合対策 - 伊那市実証実験 -

長野県伊那市と2回の実証実験を実施し、開発した4技術（画像･映像フェイク検知、ファクトチェック、WM、DID/VC）を統合したシステムの実環境での有効性を検証した。伊那市との実証実験では、伊那市が実際に投稿している画像や文章をベースにデータセット構築を行い検証した。

実証実験概要:

項目	実証1	実証2
期間	10/4～11/4	1/21～1/27
実証先	長野県伊那市	
参加者	伊那市職員 NTT東日本	
投稿プラットフォーム	LINE / X	

総合KPI達成状況:

KPI	計画値	達成値	達成状況
投稿数	20件以上	36件	達成
偽情報投稿数	30件以上	66件	達成
フェイク検知率（Recall）	90%以上	100%	達成

達成できたこと

- ✓ フェイク検知率100%（Recall）を達成
- ✓ 4技術の統合システムとして実環境で動作を確認
- ✓ 2回の実証を通じた改善サイクルの実践
- ✓ WMファイルサイズ問題の解消

今後の課題

- **ロ リアル画像の正判定精度向上**:閾値の最適化、リアル画像の学習データ拡充
- **ロ 加工済み画像（ポスター等）への対応**:フェイク検知モデルにおける「デザイン加工」と「AI生成/改ざん」の識別能力向上

目次

5-1. 社会実装に向けた取組の全体像

NTT東日本ひかり電話環境での検知アプリ開発と実網検証

NTT東日本との共同開発により電話着信時のフェイク音声リアルタイム判定アプリを開発。ひかり電話・携帯電話・IP電話の実網で検証を実施。本技術は任意の電話サービスに適用可能。

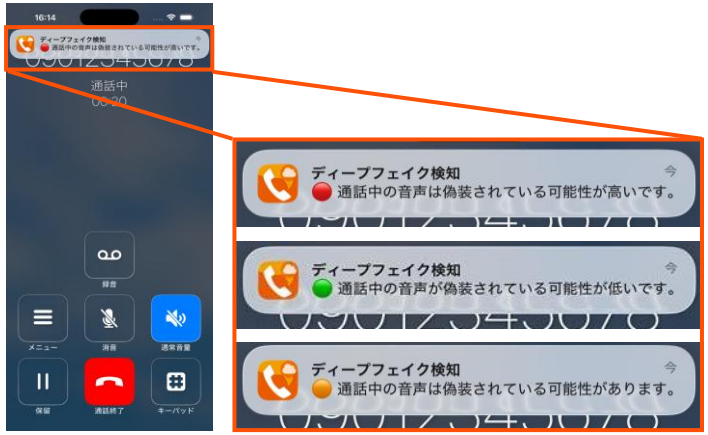
取組

- NTT東日本ひかり電話と連携し、着信時にリアルタイムでフェイク判定するアプリを開発
- 着信→リアルタイム検知→結果通知の機能を実装
- NTT東日本の実際のひかり回線ルートを使用し、3パターンの発着信環境で検証（男女・会話内容など複数パターンの音声で実施）
- CODEC変換・エコーキャンセル処理を含む実環境での検知性能を測定

成果

- ✓ 実際の電話サービスに音声フェイク検知エンジンを適用する技術的ポイントが明らかになり、社会実装に向けて進展
- ✓ 実電話回線環境ではCODEC変換による精度低下を確認
  → 実環境音声での再学習により改善見込み
- ✓ 検知アプリにより、利用者がリアルタイムで判定結果を確認可能

発着信パターン	CODEC	F1スコア
①ひかり電話→ひかり電話	μ-law → μ-law	0.727
②携帯電話→ひかり電話	AMR-WB/NB → μ-law	0.769
③IP電話→ひかり電話	Opus → μ-law	0.700



5-1. 社会実装に向けた取組の全体像

長野県伊那市との実証実験と総合対策システムの社会実装

実際の自治体による情報発信および、その発信が拡散された場合を想定するとともに、虚偽投稿の発生を想定した実証を実施

取組：実証環境と具体的なステップ

■システム構築と2度の実証実験

- 基盤技術を統合したWebベースの総合対策システムを構築し、自治体が利用できる形で提供
- 伊那市を実証フィールドとし、実務フローに合わせた検証を実施
- **実施期間**：(2025/10-11月) + (2026/1月) の計2回実施

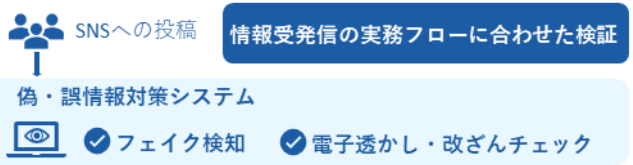
■実証手順の詳細

❶ VC（証明書）発行
伊那市担当者に認定投稿者としてのVC（証明書）発行

❷ WM埋め込みとSNS投稿
VCをウォーターマークとして投稿画像に埋込み、SNSへ投稿

❸ ブラインド混合投稿
虚偽投稿後者（NTT東日本）が本物とフェイクを投稿

❹ 検証・真偽判定
双方の投稿を検証し、真偽判定を実施

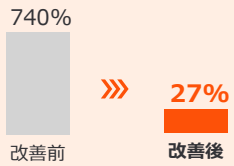


成果：実用性の確立と技術改善

- 実務フローに即した形で想定通り証明書（VC）が埋込まれることを確認
- SNSで拡散後でも真正性が把握できることを確認
- フェイク投稿に対しては検知エンジンで対処可能なことを確認
- 実証1の伊那市フィードバック5項目に全て対応し、実用レベルへ改善

評価指標（KPI）	目標	実績	達成度
自治体担当者 投稿数	20件以上	36件	180%
偽誤情報投稿数（検証用）	30件以上	66件	220%
フェイク検知率	90%以上	100%	perfect

- WM埋め込み後のファイルサイズ増加率740%→27%と実用可能なレベルに改善



*KPI実績値は伊那市実証（n=66）における達成値

5-1. 社会実装に向けた取組の全体像

自治体向け偽・誤情報総合対策システムの全体構成

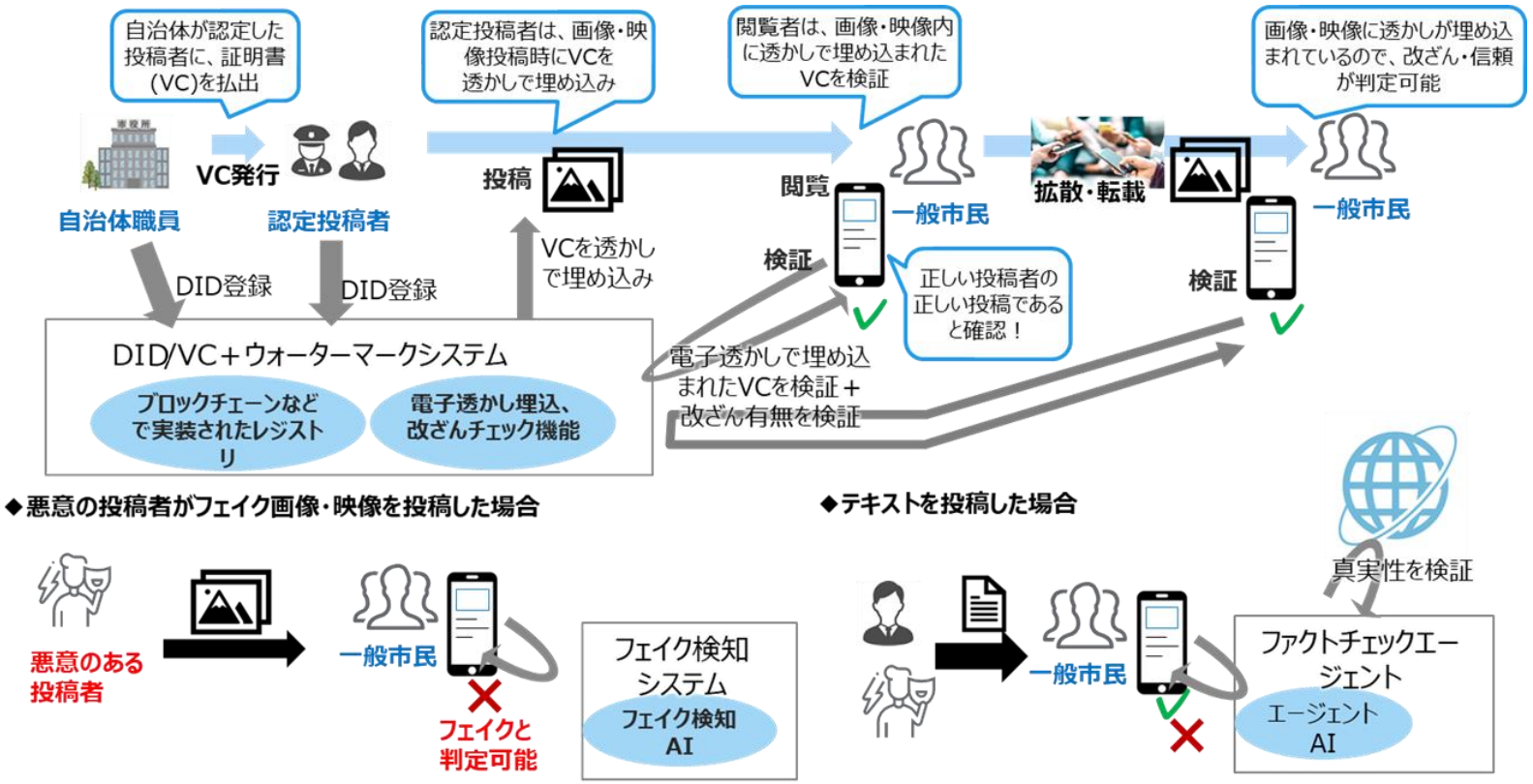
自治体の実務フローに適した形で基盤技術を統合し、Webベースの総合対策システムとして構築

情報発信フロー（証明）

自治体職員 → VC発行 → WM埋込 → SNS投稿（真正性付与）

情報検証フロー（検知）

SNS投稿の収集 → フェイク検知 ＋ WM検証 ＋ ファクトチェック →
→ 統合判定 → 結果表示



5-2. 社会実装に向けた取組の個別詳細

電話音声フェイク検知 - 検知アプリの全体像 -

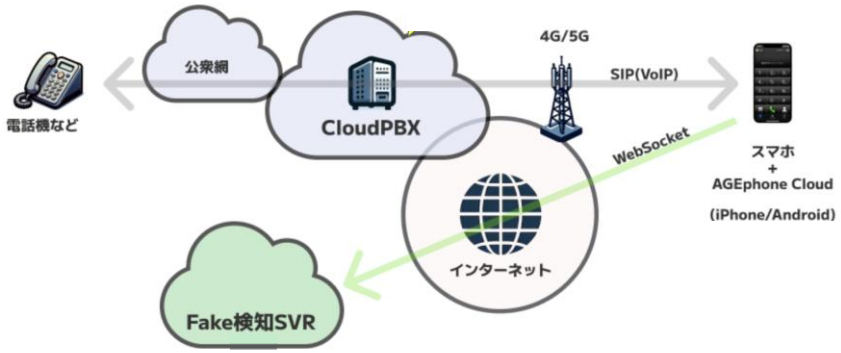
NTT東日本およびageetとの共同開発により、電話着信時にリアルタイムでフェイク音声を判定するアプリケーションを開発した。スマートフォンアプリとNABLASのディープフェイク検知サーバ（DFS）をWebSocketで接続し、通話中の相手音声をリアルタイムで解析する構成を実現した。

システム構成

構成要素	内容
AGEphone Cloud	スマートフォンアプリ。通話中の相手音声を取得しDFSに送信。判定結果をリアルタイムでユーザーに通知
ディープフェイク検知サーバ（音声検出モデル）	End-to-Endモデルによるリアルタイム音声解析。解析区間ごとにフェイク確率を算出

設計思想

- ✓ **リアルタイム性**
  解析区間で逐次判定を行い、通話中に被害の未然防止を可能にする
- ✓ **汎用性**
  特定の電話サービスに依存しないアーキテクチャ。ひかり電話に限らず任意の電話サービスに適用可能
- ✓ **堅牢性**
  切断時の自動再接続。通話コンテキスト保持による再接続対応
- ✓ **ユーザビリティ**
  3段階の直感的な判定表示。技術に詳しくない一般ユーザーでも即座に理解可能



5-2. 社会実装に向けた取組の個別詳細

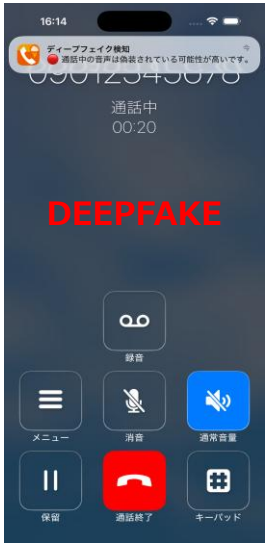
電話音声フェイク検知 - 検知アプリUI -

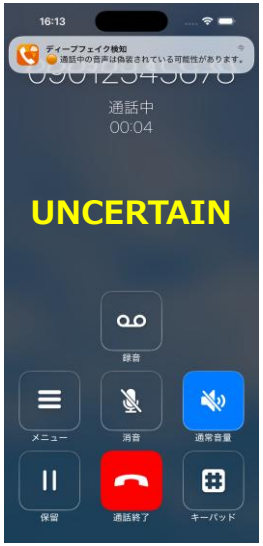
通話中にリアルタイムでフェイク判定結果を表示するUIを開発。フェイク確率に基づく3段階判定を視覚・聴覚の両方で通知し、利用者が通話中に適切な判断を行えるよう設計した。

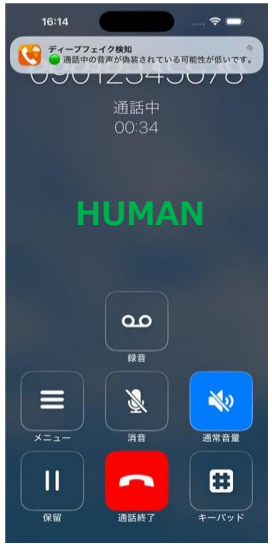
3段階判定と通知仕様

判定	アイコン	サウンド	通知テキスト
HUMAN	緑チェックマーク	なし	「この通話はディープフェイクの可能性が低いです」
UNCERTAIN	黄色感嘆符	ビープ音（単発）	「この通話はディープフェイクの可能性があります」
DEEPFAKE	赤色警告灯	クラクション音	「この通話はディープフェイクの可能性が高いです」

検知アプリUI







5-2. 社会実装に向けた取組の個別詳細

電話音声フェイク検知 - 実網検証環境の構築 -

フェイク音声検知の実用性を評価するため、NTT東日本のひかり電話網を使用した実環境検証基盤を構築した。
3パターンの発着信環境を用意し、コーデック変換が検知精度に与える影響を実際の電話網で検証した。

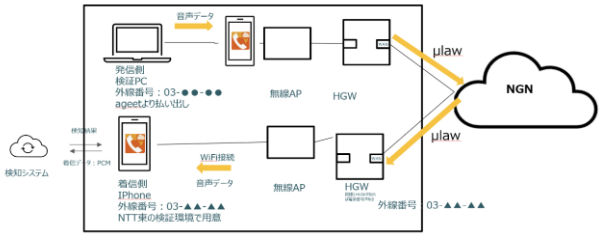
各コーデックの特性

コーデック	サンプリング周波数	特性
μ-law	8kHz	固定電話の標準CODEC。対数圧伸による非圧縮PCM（64kbps）。帯域は300-3400Hz
AMR-WB	16kHz	VoLTE/4G通話で使用。広帯域（50-7000Hz）。可変ビットレート（6.6-23.85kbps）
AMR-NB	8kHz	3G通話で使用。狭帯域（300-3400Hz）。可変ビットレート（4.75-12.2kbps）
opus	8-48kHz	VoIP/WebRTCで使用。超広帯域対応。可変ビットレート（6-510kbps）

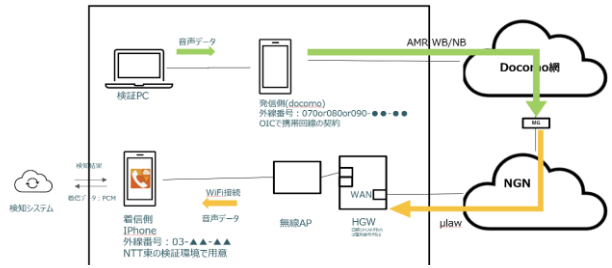
二重コーデック変換と検知難度

- 実電話網では、音声は発信側でエンコード → ネットワーク伝送 → 着信側でデコード＋μ-lawへ再エンコードという二重変換を経る
- この過程で、フェイク検知モデルが依拠する音声特徴量が不可逆的に変質・欠損する
- 特にAMR-WB（16kHz）やopus（最大48kHz）からμ-law（8kHz）への変換では、高周波帯域の情報が完全に失われる
- シミュレーション環境（コーデック変換なし）とは根本的に異なる条件であり、実環境特有の検知困難性がある

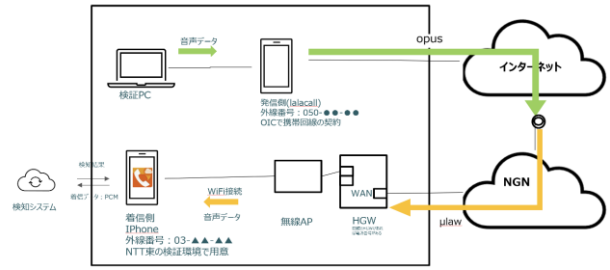
・発信側：ひかり電話、着信側：ひかり電話
・コーデック：μlaw→μlaw



・発信側：携帯電話(docomo)、着信側：ひかり電話
・コーデック：AMR WB/NB→μlaw



・発信側：IP電話、着信側：ひかり電話
・コーデック：CS ACELP/opus→μlaw



5-2. 社会実装に向けた取組の個別詳細

電話音声フェイク検知 - 実環境検証で得られた知見 -

実電話網での検証を通じて、シミュレーションでは得られない実運用上の重要な知見を獲得した。これらの知見は今後の技術改善と社会実装の方向性を定める基盤となる。

得られた知見

知見	詳細	社会実装への示唆
実環境とシミュレーションの乖離	シミュレーションで学習したモデルは実環境で精度が低下する。エコーキャンセル・回線ノイズ等が影響	シミュレーションのみでの開発には限界。実環境データの活用が不可欠
CODEC変換の影響の大きさ	CODEC変換がフェイク検知精度に最も大きく影響	電話サービスごとにCODECが異なるため、展開先に応じた最適化が必要
リアルタイム処理の実現性	通話中の音声を逐次送信・解析し、リアルタイム判定が技術的に実現可能であることを確認。WSS通信による安定接続と再接続機構も動作を検証	検知精度の課題は残るが、パイプラインとしての実装は完成
安全側に倒した通知設計の有効性	危険度が上がった場合は即座に警告し、下がった場合は慎重に判定を緩和する設計が、利用者の安全を優先する通知体験に有効であることを確認	「見逃し」を防ぐことを最優先とした設計原則は、他のフェイク検知UIにも適用可能

5-2. 社会実装に向けた取組の個別詳細

電話音声フェイク検知 - 実装の成果 -

本事業では検知精度に課題を残しつつも、実際の電話網での検証基盤構築とアプリ実装を実現し、社会実装に向けた重要な基盤を開発した。

実装の成果

成果	内容
検知アプリの開発	リアルタイム検知アプリケーションを実装。WSS通信・3段階判定UIを搭載
実電話網での検証基盤	NTT東日本のひかり電話網で3パターンの発着信環境を構築。実環境でのフェイク検知評価基盤を確立
汎用アーキテクチャ	ひかり電話に限定されず任意の電話サービスに適用可能な設計。WSS接続による柔軟なサーバ連携
実運用上の知見獲得	CODEC変換の影響、シミュレーションと実環境の乖離等、今後の改善に不可欠な知見を獲得

社会的意義

- **先駆的取組:** 電話環境でのフェイク音声検知を実際の電話網で検証した事例は希少。実環境での課題を明確化したことに技術的価値がある
- **詐欺対策の基盤:** フェイク音声による特殊詐欺（オレオレ詐欺等）の被害は深刻化しており、検知技術の社会実装は喫緊の課題。本事業で実装基盤と検証環境を確立した
- **技術的ポイントの明確化:** 電話サービスへの検知技術組み込みに必要な技術要件（リアルタイム性、CODEC対応、UI設計等）を明らかにした
- **今後の改善の方向性:** CODEC変換による特徴量欠損に対する解決策の検討に着手済み

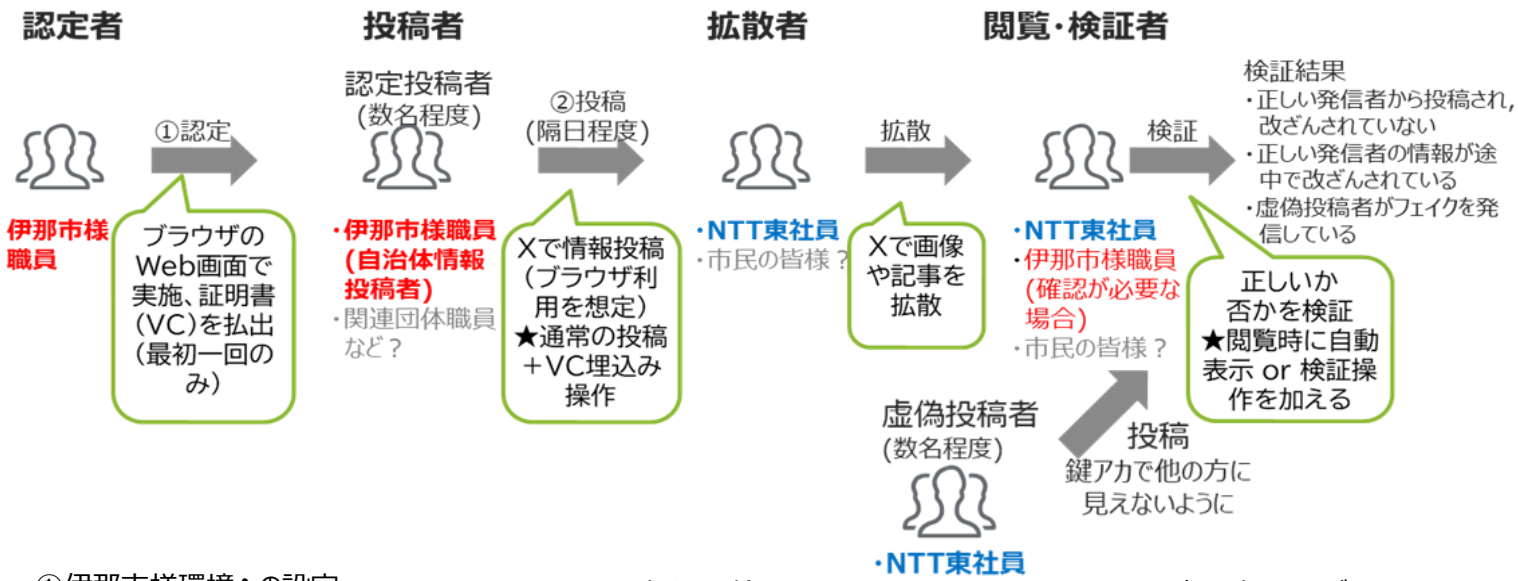
5-2. 社会実装に向けた取組の個別詳細

自治体向け偽・誤情報総合対策 - 実証実験の運用フローと実施体制 -

伊那市の通常業務フローにシステムの操作を組み込み、実運用に近い形で2回の実証実験を実施

検証で判定できる内容

- ✓ 正しい発信者から投稿され、改ざんされていない証明
- ✓ 正しい発信者の情報が途中で改ざんされている検出
- ✓ 虚偽投稿者のフェイク発信の検出



①伊那市様環境への設定
- WM埋込システムにログインします
- WM埋込処理に必要になる、鍵生成を行います。
- SNS投稿されるアカウントにVCを発行します。

②WM埋込～検証
- 画像にWMを埋めこみます
- WMを埋め込んだ画像、AIフェイク画像、文章ファクトチェックの手順確認を行います。

③NTT東日本によるデモ
- 動画にWMを埋めこみます
- WMを埋め込んだ動画、AIフェイク動画の検証を行います。

5-2. 社会実装に向けた取組の個別詳細

自治体向け偽・誤情報総合対策 - 実証実験中の役割 -

実証実験では、伊那市側が「投稿者」および「閲覧・検証者」として、NTT東日本側が「検証投稿者」および「技術個別技術検証」として、それぞれの役割を分担して実施

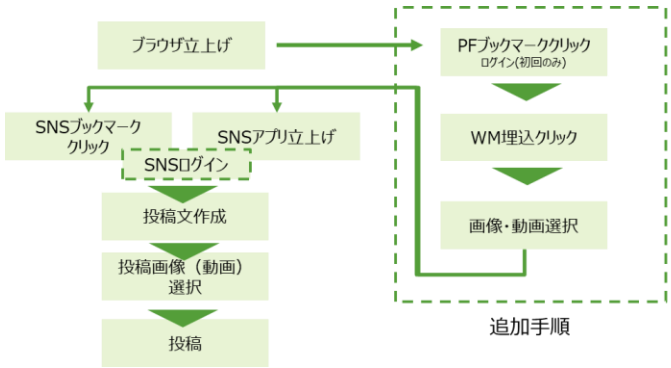
伊那市側の役割 :合計36件の検証投稿

役割	内容
投稿者 (通常業務)	通常業務のSNS投稿時に、画像へのWM埋込を追加して投稿 投稿内容:地域イベント情報、熊目撃情報、火災想定訓練模様等
閲覧・検証者	NTT東からの検証投稿を本システムで検証し、判定結果を確認
フィードバック	ディープフェイクの精巧さや怖さを経験し、人目ではわからないことをシステムで検知できることを体験

NTT東日本側の役割 :合計66件の検証投稿

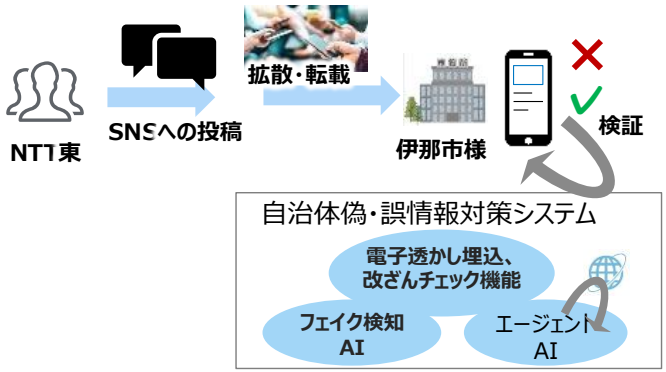
役割	内容
虚偽投稿者役	フェイク画像（AI生成）を使用した投稿をLINE・Xの検証用鍵アカウントから実施。市民には非公開
一般ユーザ役	フェイクでない画像を使用した投稿を同アカウントから実施
検証操作デモ	動画へのWM埋込、WMを埋め込んだ動画・AIフェイク画像の検証操作をデモンストレーション

①通常業務



投稿する画像にWMを埋込を実施し、埋め込まれた画像を投稿

②閲覧側



NTT東が投稿する、正しい情報及び偽・誤情報（を想定したもの）を本システムで検証いただき、どのような投稿が判定できるかの検証

5-2. 社会実装に向けた取組の個別詳細

自治体向け偽･誤情報総合対策 - 実際の投稿例 -

実証実験時の伊那市の通常業務によるSNS投稿（WM埋込）と、NTT東日本による検証用投稿の実例

投稿方法

1. 証明書（VC）をWMとして埋め込んだ画像を、伊那市公式アカウントから投稿
2. 虚偽投稿者（フェイク画像を使用）及び、一般ユーザ（フェイクではない画像を使用）をNTT東の検証用アカウントで投稿

伊那市公式アカウントの投稿

- 地域イベント情報
- 熊目撃情報
- 火災想定訓練模様など

NTT東の検証用虚偽投稿／一般投稿

- 熊目撃情報
- 川の氾濫情報

5-2. 社会実装に向けた取組の個別詳細

自治体向け偽・誤情報総合対策 - ファクトチェックエージェントの検証 -

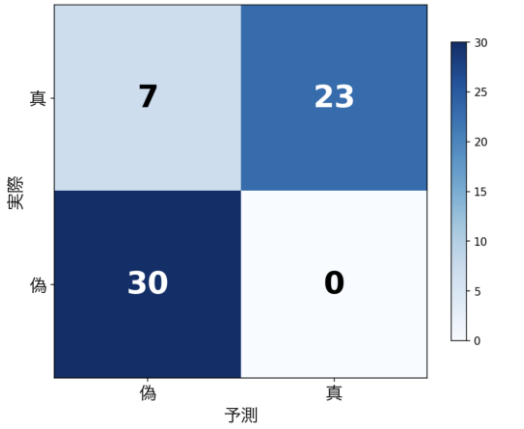
技術検証で精度を確認したファクトチェックエージェントについて、伊那市Webサイトの実際の情報を用いた社会実装観点での最終評価を実施

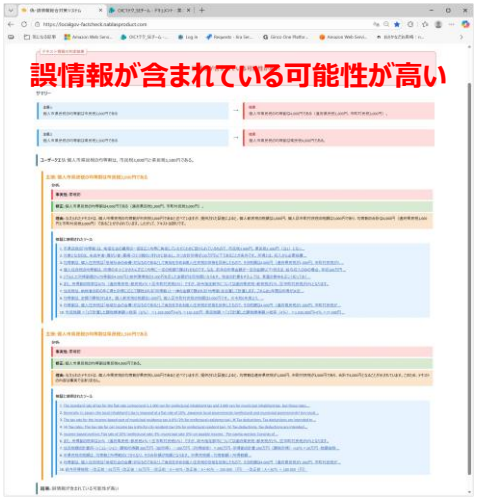
検証内容

データセット：伊那市Webサイトの実データに基づく60件
（正しい情報30件 ＋ 誤情報30件）

検証項目	正解率
誤情報の検出（偽→偽）	30/30
正情報の判定（真→真）	23/30

Accuracy：88.3%
Recall：100%
Precision：81.1%





正情報の誤判定原因と改善方針

- 伊那市ではない場所の個人市民県民税を参照元としている可能性がある
- 令和6年度から税率更新されているので「令和6年度の」を指定しないと正しい参照元を検索できない可能性がある

→参照元を限定する等の対応が必要

5-2. 社会実装に向けた取組の個別詳細

自治体向け偽･誤情報総合対策 - フィードバック対応と今後の課題 -

実証1で伊那市から寄せられた5項目の改善要望に全て対応し、実証2で改善効果を検証した。今後の社会実装に向けた課題を示す。

伊那市フィードバック対応（5項目 → 全対応済み）

フィードバック項目	対応結果
編集ソフトによる加工への検出精度向上	生成AI以外の加工も検出可能に改良。注意喚起表示を追加
判定結果の表示方法改善	「Real/Fake」二択 → **3段階表示**に改善（徴候なし/徴候あり/可能性が高い）
WM埋込画像からの編集検出	WM検証済みでも編集が検出された場合に注意喚起する仕組みを追加
WM埋込によるファイルサイズ増大	圧縮方式に変更。**740% → 27%**に大幅改善
ファクトチェックの検出精度向上	参照元限定・プロンプト調整。誤情報検出率**100%**達成

今後の課題

課題	対応方針
加工済み画像（ポスター等）のUI表現	WM検証結果とフェイク検知結果の統合表示ロジック設計
WM埋込作業の負担軽減	投稿フローへの自動化・一括処理機能追加
利用者別の結果表示最適化	職員向け（詳細）/ 市民向け（簡潔）の表示UI設計
ブラウザからの直接検証UI	ブラウザ拡張機能 or 簡易検証ページ開発
市民通報（不法投棄等）への応用	通報システムとの連携インタフェース設計

目次

6-1. 普及啓発活動の全体像

普及啓発活動に係る取組・成果の全体像

本事業で開発した偽・誤情報対策技術の社会的認知向上と情報リテラシー強化のため、メディア露出・展示会出展・情報発信の3つの柱で普及啓発活動を推進した。

メディア露出

テレビ取材協力を通じた社会的認知の向上 ディープフェイクの脅威と検出技術を一般視聴者に訴求

主な実績

- ✓ テレビ取材7件（地上波5局）
- ✓ 雑誌・Webメディア掲載5件

展示会・業界連携

業界関係者・自治体向けのデモンストレーションと技術連携

主な実績

- ✓ 展示会出展3件
- ✓ GMOブランドセキュリティとの技術協力

情報発信・リテラシー向上

偽情報への対処法やフェイク検知の知見を一般向けに発信

主な実績

- ✓ 記事公開
- ✓ インタビュー掲載

定量的インパクト

- ✓ テレビ放映: 地上波5局・7番組（NTV / フジ / テレ朝 / STV / TeNY / NHK）
- ✓ SNSリーチ: ライブドアニュース記事（NTT東日本×NABLAS）が13万インプレッション、1,130いいね、563リポスト
- ✓ 雑誌: 日経トレンディ「フェイク AIバスター」として掲載

6-2. 普及啓発活動の個別詳細

メディア取材・掲載実績

フェイク検知技術の社会的認知向上のため、テレビ取材協力および雑誌・Webメディアへの情報提供を積極的に実施した。*

テレビ取材協力（7件）

日付	番組名（局名）	内容
2025/10/19	日本テレビ「スクール革命！」	ディープフェイクの素材提供と検出技術紹介
2025/10/24	フジテレビ「Live news イット！」	ディープフェイク動画の検出技術（KeiganAI）取材
2025/12/11	STV札幌テレビ「どさんこワイド179」	生成AIによるディープフェイク動画の検出
2025/12/14	テレビ朝日「有働 Times」	クマに関するディープフェイク動画の検出とKeiganAI紹介
2025/12	NHK	音声フェイク検出アプリについて
2026/01/29	TeNYテレビ新潟「TeNY新潟一番ニュース」	生成AI×選挙の特集でフェイク動画の検出・情報提供
2026/02/04	フジテレビ「Live news イット！」（2回目）	選挙関連ディープフェイク動画の検出技術取材

雑誌・Webメディア掲載（5件）

日付	メディア	内容
2025/12	日経トレンディ 2026年1月号	代表中山インタビュー（「フェイク AIバスター」として掲載）
2026/01/27	NABLAS公式サイト	KeiganAIプロジェクトメンバーインタビュー
2026/01/29	ライブドアニュース	「”生成AIを悪用”フェイク音声など見破る技術公開」（**13万imp**）
2026/02/18	GMOブランドセキュリティ ホワイトペーパー	2026年2月衆院選におけるSNS投稿・生成AI動画の流通実態分析
2026/02/24	NABLAS公式サイト	「フェイク情報とどう向き合うか ― 情報に振り回されないために必要なことは？」

*テレビ取材はNABLASのフェイク検知プロダクト「KeiganAI」としての取材対応だが、弊社フェイク検知技術が基盤となっているとめ普及啓発活動の一環として記載

6-2. 普及啓発活動の個別詳細

メディア取材・掲載実績 – 詳細 -

メディア取材や掲載時のX投稿による詳細例を示す。

フジテレビ「Live news イット！」

NABLAS株式会社 @nablas... · 2025/10/27
📺取材協力のお知らせ
10/24に放送されたフジテレビ「Live news イット！（@livenews_it）」でNABLASが取材協力をしました。

ディープフェイク動画に対する検出技術として弊社のKeiganAIを取材いただきました！

内容はYoutbeからもご視聴いただけます。
ぜひご覧ください。

テレビ朝日「有働 Times」

NABLAS株式会社 @nablas... · 2025/12/13
＼📺取材協力のお知らせ／
12月14日（日）放送の、テレビ朝日「有働Times」でNABLASが取材協力をしました。

急増するクマに関するディープフェイク動画の検出を通し、弊社プロダクト「KeiganAI」をご紹介しています。

12/14(日) よる8:56～
※有事の際は変更の可能性あり

場所: tv-asahi.co.jp

日経トレンディ 2026年1月号

NABLAS株式会社 @nabla... · 2025/12/09
『日経トレンディ』掲載のお知らせ

日経トレンディ 2026年1月号に、弊社代表中山のインタビュー記事が掲載されました！
生成AIによるディープフェイクを検出する技術についてご紹介しています。

6-2. 普及啓発活動の個別詳細

展示会・業界連携・リテラシー向上活動

業界関係者・自治体向けの展示会出展と、情報リテラシー向上に向けた取り組みを実施した。

展示会・フォーラム出展

日付	イベント名	内容
2026/01	地域ミライ共創フォーラム	出展（テレビ取材あり）
2026/02	保安通信協会展示会	出展
2026/02	ITmedia Apex Innovations 2026冬	出展

業界連携・標準化活動

活動	内容
GMOブランドセキュリティとの技術協力	2026年2月衆院選のSNS投稿・生成AIコンテンツの流通実態をNABLASの技術で分析。ホワイトペーパーとして公開
フェイク情報とどう向き合うかについての情報発信	「フェイク情報とどう向き合うか ー 情報に振り回されないために必要なことは？」

目次

7-1. 技術開発及び社会実装における課題・展望

電話音声フェイク検知 - 技術課題 -

4章の精度検証および5章の社会実装を通じて、以下の技術課題が明らかになった。

課題1: CODEC変換による検知特徴量の欠損

- 実電話網では二重コーデック変換により、フェイク検知に有効な音声特徴量が不可逆的に欠損する
- 特に高サンプリング周波数（AMR-WB、opus）からμ-law（8kHz）への変換で高周波帯域の情報が完全に失われる
- 回線パターンによって精度が大きく変動し、単一モデルでの全環境カバーが困難

課題2: 実環境とシミュレーションの乖離

- シミュレーション環境で学習したモデルは、実電話網で精度が低下する
- エコーキャンセル・回線ノイズ等、シミュレーションでは再現しきれない要因が影響
- 実環境データの蓄積と活用が不可欠だが、データ収集コストが高い

課題3: 音声波形のみによる検知の原理的限界

- CODEC変換後の音声では、フェイク/リアルの差異を示す情報が物理的に復元不可能な場合がある
- 音声波形の特徴量のみに依拠する検知手法には原理的な限界が存在する可能性がある

課題4: 最新音声生成技術への追随

- 音声合成技術は急速に進化しており、新たな生成モデルが次々と登場する（例: ElevenLabs V3等）
- 最新の生成モデルは、商用検出モデルを含む既存の検知手法ではほぼ検出困難なレベルに到達しつつある
- 検知モデルの学習データに含まれない生成手法で作られた音声に対しては、汎化性能が大幅に低下する

7-1. 技術開発及び社会実装における課題・展望

電話音声フェイク検知 - 今後の展望 -

課題に対し、以下の4つのアプローチで検知精度の向上と社会実装の拡大を目指す。

展望1: 環境適応型モデルの実装

- 回線種別（ひかり/携帯/IP）ごとに最適化したモデルを構築し、推論時に自動ルーティングする仕組みを実装
- CODEC変換の影響要素を分解したAugmentationデータでの学習を拡充
- 実電話網での学習・検証データを継続的に蓄積し、モデルの実環境適応を強化

展望2: アンサンブル型検知体制の構築

- End-to-Endモデルと特定の生成手法に特化したモデルを組み合わせたアンサンブル構成で、未知の生成手法にも対応可能な検知体制を構築
- 新たな生成モデルの登場に対し、特化型モデルの追加で迅速に対応できる拡張性を確保
- OSSモデルとの比較評価で確立した評価基盤を活用し、継続的にベンチマークを実施

展望3: マルチモーダル検知への拡張

- 音声波形の特徴量に加え、会話内容・文脈からの詐欺検出を組み合わせたハイブリッドアプローチ
- CODEC変換の影響を受けない会話内容ベースの検知は、波形検知の原理的限界を補完する
- End-to-Endモデルと特化型モデルのアンサンブルで最新生成モデルへの対応力を強化

展望4: 検知基盤の社会展開

- ひかり電話に限定されない汎用アーキテクチャを活かし、他の電話サービスへの展開を推進
- 3段階判定UI・安全側に倒した通知設計を標準的な検知UIとして確立
- 実電話網での検証基盤を継続運用し、新たな回線環境・生成モデルへの対応を検証

7-1. 技術開発及び社会実装における課題・展望

自治体向け偽・誤情報総合対策 – 運用・実装課題 -

伊那市との実証検証を通じて、運用方法および検証結果表示に関して以下の課題が確認された。

課題1: WM埋込および検証システムの利便性

- 正しい情報発信者としてVC（証明書）を埋め込む作業が利用者側の負担となり、導入・運用のハードルが高い
- 投稿内容の検証作業においても、通常のSNS閲覧に加えて追加操作が必要となり、利用につながりにくい
- 可能な限り少ない操作で利用できる運用フローの設計・改善が必要

課題2: 利用ケースごとの適切な検証結果表示

- 市民からの問い合わせに対応する自治体の立場では、誤解されることのない慎重な判定結果の提示が求められる
- 一方、市民が自ら投稿内容を確認する場合には、判断が明確にわかるシンプルな結果表示が望まれる
- 利用者の立場によって求められる情報量や表現が異なるため、検証結果の表示方法については引き続き議論・検討が必要

課題3: 対策体制・業務フローの未整備

- 偽・誤情報対策の専任人員や業務フローが未整備の自治体が多い
- 有事に直面して初めて対策を講じる後手の状況であり、平時からの体制構築が必要

課題4:評価基準・指標の未確立

- 偽・誤情報検知技術の標準的な評価データや指標・基準が明確化・標準化されていない
- 技術導入の判断基準が不透明なため、自治体が安心して導入できる環境が整っていない

7-1. 技術開発及び社会実装における課題・展望

自治体向け偽・誤情報総合対策 - 今後の展望 -

課題に対し、以下のアプローチで運用の簡易化と社会実装の拡大を目指す。

展望1: 運用の簡易化・自動化

- **WM埋込みの自動化**: SNS投稿時に利用者の追加操作を必要とせず、自動的にWMを埋め込む仕組みを実装。投稿者の負担を軽減し、運用ハードルを低減
- **検証システムの簡易化**: SNS上の投稿を閲覧した際に、ワンクリックで真偽判定を実施できる機能を提供。利用ケースに応じて可能な限り少ない操作で検証が完了する仕組みを整備

展望2: 利用者に応じた検証結果表示の最適化

- **サマリ表示と詳細表示の切替**: 利用者の立場に応じて2種類の表示方式を提供
  - **サマリ表示**: 一目で判定結果が分かる簡潔な結果表示
  - **詳細表示**: 判定に至った根拠や分析内容を確認できる詳細情報の提示
- 自治体・市民など利用者に応じて必要な情報量が異なることから、柔軟な表示方法を選択可能にする

展望3: ツール提供から一気通貫支援サービスへ

- ツール提供にとどまらず、注目事象の自動抽出→調査→レポート判定→方針検討まで、全体をサポートできる支援サービスを構築
- 平時からの体制構築を支援し、有事に即座に対応可能な運用基盤を提供

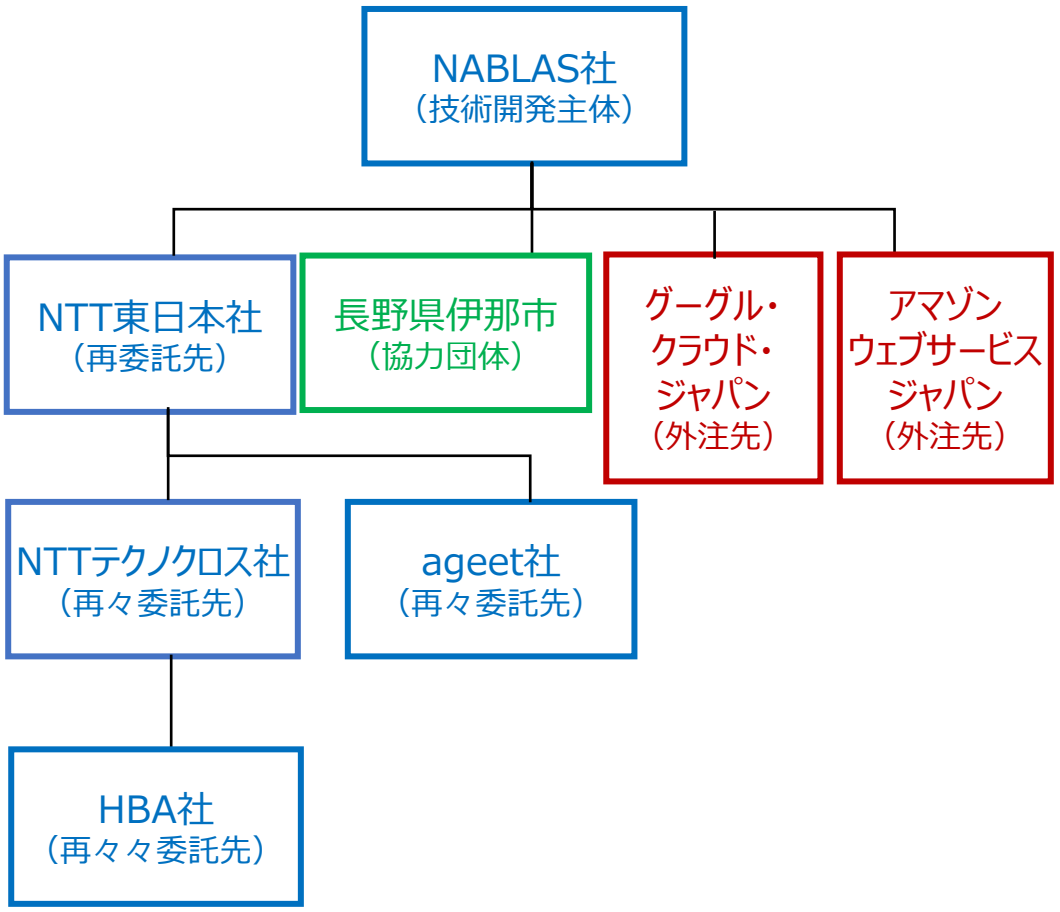
展望4: 評価基準の標準化推進

- 評価データセットを公開し、ベンチマーク結果やリーダーボードを提供
- ユーザー側が技術を評価・比較できるアプリケーションを提供
- 基準や規格の標準化を推進し、自治体が安心して導入できる環境を整備

目次

8-1. 実施体制及び役割分担

本事業の実施体制図



各団体の役割・業務範囲

- **NABLAS株式会社**
  - ディープフェイク検出モデル開発
  - ファクトチェックAIエージェント開発
  - 学習・評価用データセット整備
  - システム開発用インフラストラクチャー構築
  - 統合アプリケーション開発
  - ビジネスモデル・事業化企画
- **NTT東日本株式会社**
  - フェイク検知付き電話サービスへの組み込み開発・実証
  - DID/VC技術開発
  - 自治体協業・実証
- **NTTテクノクロス株式会社**
  - DID/VC技術の技術開発
- **株式会社HBA**
  - DID/VCシステムの製造及び試験
- **株式会社ageet**
  - 電話サービス組み込みアプリの開発
- **長野県伊那市**
  - 情報発信及びその発信が拡散された場合を想定するとともに、虚偽投稿の発生を想定した実証
- **グーグル・クラウド・ジャパン合同会社**
  - AIモデルの継続的な学習・効率的なデータ管理及び実証時の堅牢なサービス運用の安定したインフラ・コンピューティングリソース確保
- **アマゾンウェブサービスジャパン合同会社**
  - 堅牢なシステム運用のためのインフラ確保

8-2. 全体スケジュール

主な実施事項（電話音声フェイク検知）	令和7年 8月	9月	10月	11月	12月	令和8年 1月	2月	3月
(1)インターネット上の偽・誤情報等への対策技術の開発	→	→	→					
1.音声フェイク検出モデルの開発	→	→						
2.電話サービス組み込みアプリ（スマホアプリ）の開発	→	→						
3.システムの統合			→	→				
4.最新生成技術に追随する仕組みの構築	→	→	→	→				
(2)インターネット上の偽・誤情報等への対策技術の有効性等に関する検証及び調査				→	→	→	→	
1.技術的な有効性検証・モデルのアップデート				→	→	→	→	
2.社会実装に向けた実証実験（NTT東日本社内）				→	→	→		
(3)インターネット上の偽・誤情報等への対策技術の社会実装に向けた取組				→	→	→	→	
1.ビジネスモデル・マネタイズ方法の検討				→	→			
2.具体的な企業・団体との提携の検討				→	→	→	→	
3.社会実装に向けた想定ターゲットへのヒアリング				→	→	→	→	
(4)成果報告書及び社会実装実施計画書の作成							→	→
1.成果報告書の作成							→	→
2.社会実装実施計画書の作成							→	→
(5)普及啓発活動への協力				→	→	→	→	→
1.出展先やワークショップ・勉強会等実施計画の検討				→	→	→	→	→

8-2. 全体スケジュール

主な実施事項 （自治体向け偽・誤情報総合対策）	令和7年					令和8年		
	8月	9月	10月	11月	12月	1月	2月	3月
(1)インターネット上の偽・誤情報等への対策技術の開発	→	→	→					
1.DID/VCシステムの開発	→	→						
2.ウォーターマーク、動画像・音声ディープフェイク検出技術、ファクトチェックエージェントの実装	→	→						
3.システムの統合		→	→					
4.最新生成技術に追随する仕組みの構築	→	→	→					
(2)インターネット上の偽・誤情報等への対策技術の有効性等に関する検証及び調査			→	→	→	→	→	
1.各種モデルの技術的な有効性検証・システムアップデート			→	→	→	→	→	
2.社会実装に向けた実証実験（自治体）			→	→	→	→		
(3)インターネット上の偽・誤情報等への対策技術の社会実装に向けた取組				→	→	→	→	
1.ビジネスモデル・マネタイズ方法の検討				→	→	→		
2.具体的な企業・団体との提携の検討				→	→	→	→	
3.社会実装に向けた想定ターゲットへのヒアリング				→	→	→	→	
(4)成果報告書及び社会実装実施計画書の作成						→	→	→
1.成果報告書の作成						→	→	→
2.社会実装実施計画書の作成						→	→	→
(5)普及啓発活動への協力				→	→	→	→	→
1.出展先やワークショップ・勉強会等実施計画の検討				→	→	→	→	→