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 概要版

2026/3/19

技13_NABLAS株式会社

目次

目次

1-1. 開発技術によりアプローチする課題・目指す姿

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

- △ 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
- △ 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
- △ 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、電話特有の音声変換等に対応できていない

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あった(3)また、総務省調査で過去に流通した偽・誤情報に対して、47.7%が「正しい」または「おそらく正しい」と認識(4)

- △ SNS上で偽情報（災害・風評・選挙等）が大量に拡散し、社会に深刻な混乱を招いている
- △ 生成AIの進化で人間による真偽判断が困難化し、閲覧数に応じた収益構造が拡散を加速させている
- △ 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

上記課題を踏まえ目指す姿・ゴール

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
- ✓ 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

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可能な情報環境の構築
- ✓ 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

(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)　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
(4)　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

1-2. 技術開発の取組・成果

電話環境に特化したフェイク音声検知モデルの開発

電話環境で音声データをタイムリーに送り込むことで高精度にフェイク音声を検知し、時々刻々と進化する生成AIにも継続的に追随できる検知モデル改善プロセスを構築

取組

□ 電話特有のCODEC変換・ノイズ等環境でも精度を維持するEnd-to-Endモデルを開発

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

□ ElevenLabs等の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

実環境・最新生成モデルに対応するためのモデル開発プロセス

□ 電話サービス組込を見据えリアルタイム処理対応の軽量かつ高精度な検知エンジンを設計

➡ API化

成果

✓ 実電話環境下のフェイク検知精度改善

End-to-Endモデルにより、CODEC変換（μ-law / AMR-WB / opus）を含む電話環境下でのフェイク検知を評価を実施

→ 実環境における検証体制の構築と精度改善を実施

✓ 生成モデルへの検知性能を評価

→ データセントリックAIによる改善サイクルによって最新モデルに対応

✓ 即現場導入可能なアプリケーションを構築

→ 改善したモデルを迅速にアプリケーションに展開可能

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　　**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
データセントリックAI：モデルのアーキテクチャを変えることなく、学習するデータの質向上から性能を最大化するアプローチ

1-2. 技術開発の取組・成果

「フェイク検知」と「発信者証明」を統合した偽・誤情報総合対策システムの開発

「フェイク検知」と「発信者証明」の基盤技術を自治体の実務に適した形で統合し、総合対策システムを構築

取組

技術 ① 総合的フェイク検知技術

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

- 画像・映像フェイク検知技術を、部分加工（Inpainting）対応を含め自治体の実用ケースに適した形で統合
- ファクトチェックエージェント技術を、複数ソース横断の事実判定が可能な形でシステムに統合

技術 ②真正性証明（電子透かし）

「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

- 発信者の真正性を証明する電子透かし（WM）およびDID/VC機能を新規開発し、システムに組み込み

成果

- **フェイク検知機能**
  Inpainting（部分加工）を含む複数の改変手法によるフェイク投稿の見逃し0件*
- **複数ソースの参照機能**
  ファクトチェックエージェントをシステムに統合し、複数ソースから自動検証
- **認定・証明機能**
  DID/VCによる認定・埋込・検証機能を開発・実装し、機能することを確認
- **総合対策システムの構築**
  「検知」と「証明」を統合した、自治体に適した総合対策システムを構築

上記技術を自治体の実務フロー向けに「1つのアプリ」へ統合

Inpainçting: 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
*KPI実績値は伊那市実証（n=66）における達成値

1-2. 技術開発の取組・成果

自治体向け偽・誤情報総合対策システムの主要機能

自治体の実務フローに適した形で基盤技術を統合し、自治体職員が直感的に操作可能なWebアプリケーションとして実装

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

メイン画面から「WM埋込」で証明書埋込画面へ遷移

対象ファイル（画像・動画）を選択「埋込依頼」クリックで証明書埋込開始
埋込完了画面に自動遷移

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

基盤技術

DID/VC 認定投稿者管理
- ✓ 認定投稿者への証明書（VC）発行・管理
- ✓ 所持VCの表示
- ✓ 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

画像・映像フェイク検知
- ✓ 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

ブロックチェーンベースのレジストリ: 中央管理者不要のデータの改ざん耐性を持ったデータベース

目次

2-1. 社会実装に係る取組・成果

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。

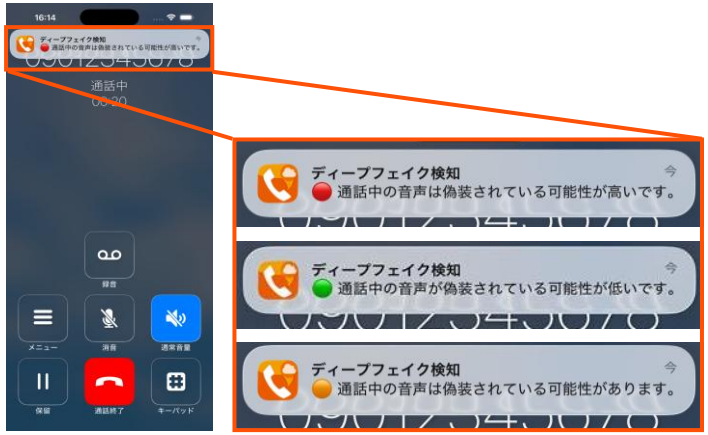
取組

- 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



2-1. 社会実装に係る取組・成果

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、虚偽投稿の発生を仮定した実証を実施

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

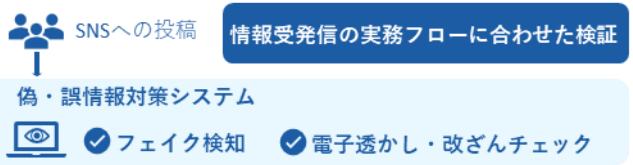
■実証手順の詳細

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

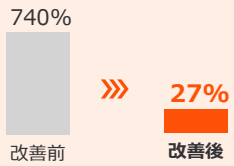


成果：実用性の確立と技術改善

- 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

2-1. 社会実装に係る取組・成果

自治体向け偽・誤情報総合対策システムの全体構成

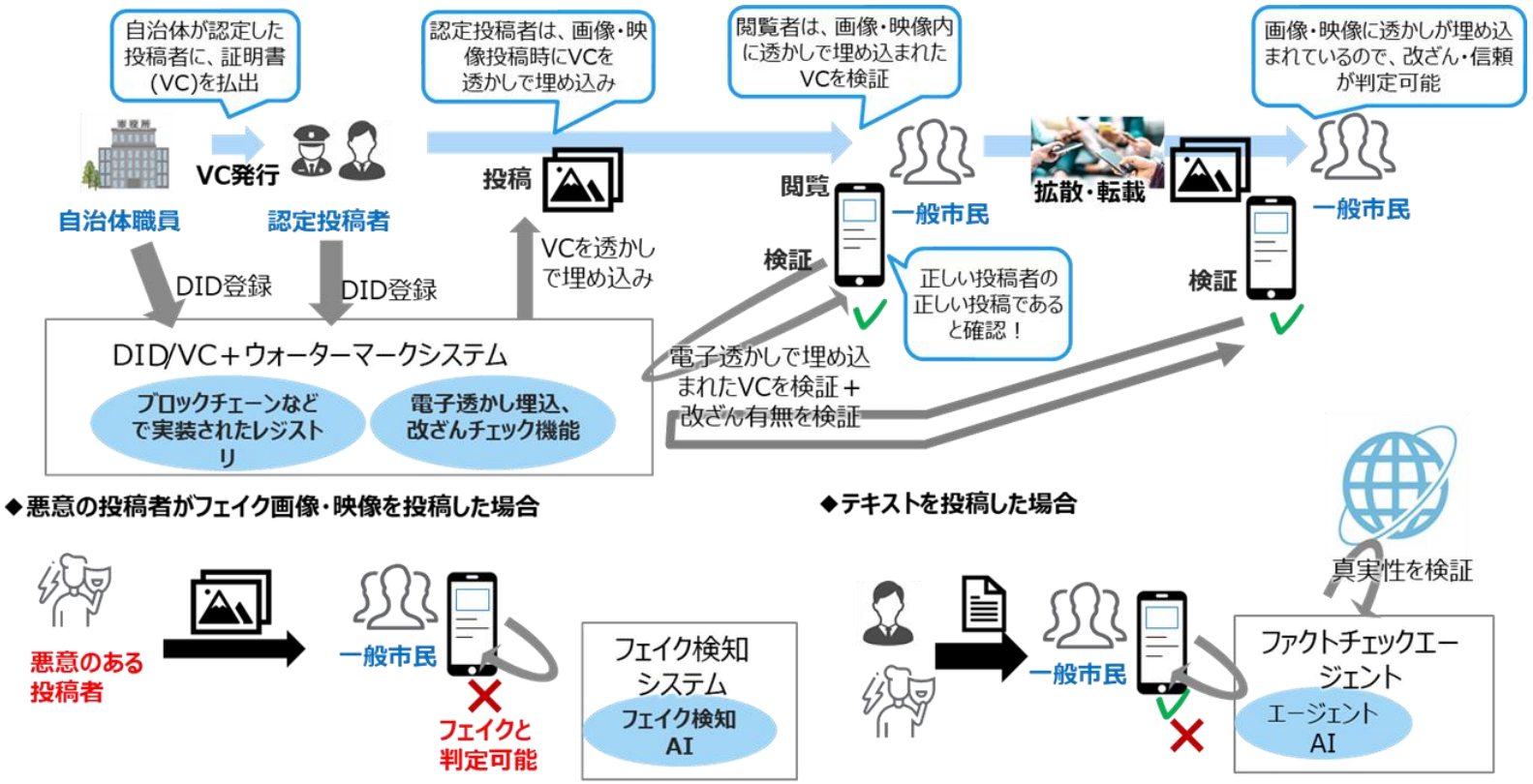
自治体の実務フローに適した形で基盤技術を統合し、Webベースの総合対策システムとして構築

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）

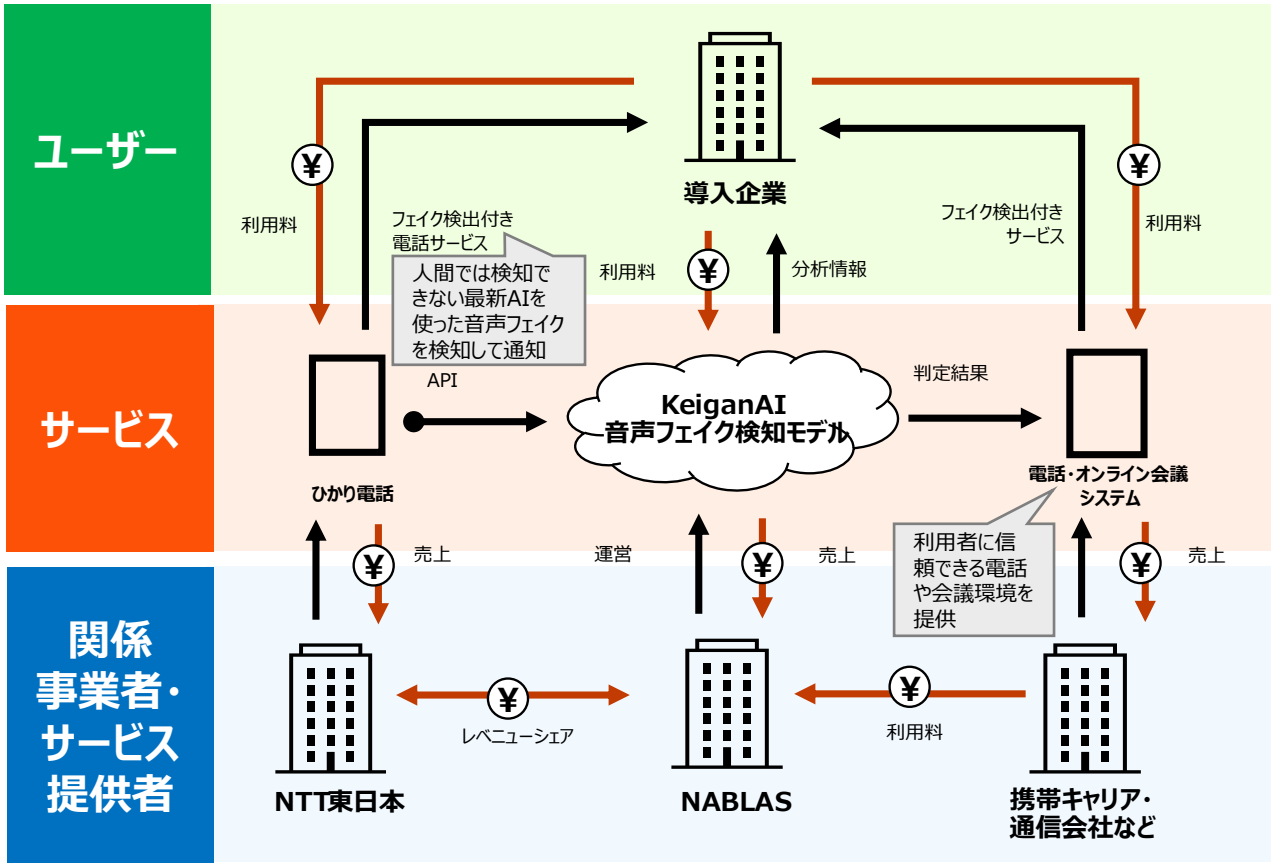
情報検証フロー（検知）

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



2-2. 社会実装時のビジネスモデル等

社会実装時のビジネスモデル - 電話音声フェイク検知 -



導入企業
- 信頼できる通話・会議環境を確保するため、本サービスを導入・利用
- 電話サービスや会議システムに対し「利用料」を支払い

NABLAS
- 電話フェイク検出サービスを含むKeiganAIのモデル開発およびサービス基盤やAPIの提供・運営
- サービス利用料/API利用料

NTT東日本
- 「ひかり電話」などの通信インフラを提供し、フェイク検知付き電話サービスとして展開
- 利用料をNABLAS社とレベニューシェア

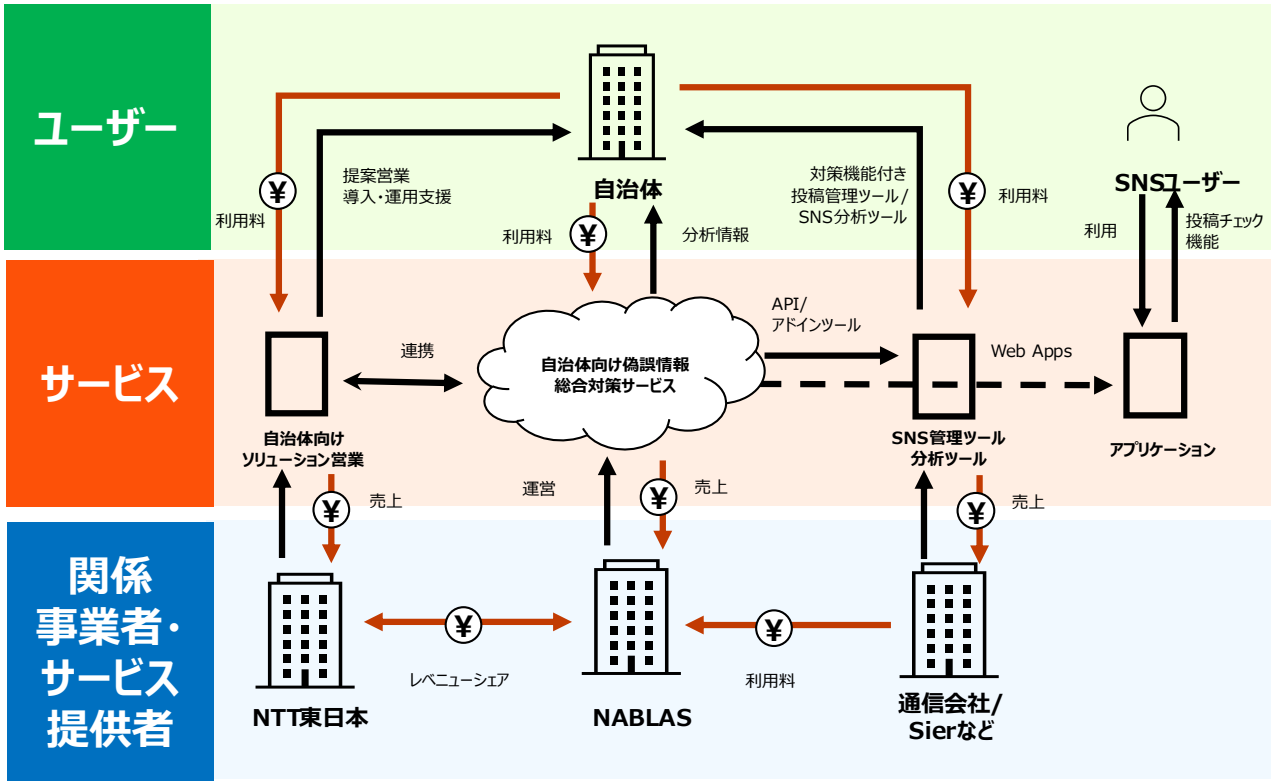
携帯キャリア・通信会社など
- 電話やオンライン会議システムにフェイク検知機能を組み込み、利用者に提供

ユーザ・導入先の詳細とそのペインポイント
- 日本国内での特殊詐欺被害は年間1,414.2億円規模であり、そのうちの約8割の当初接触ツールが電話であることから、AI生成音声の悪用リスクが急速に高まっている一方で、最新の生成AIを活用したなりすましや詐欺が増加しているが検知する手段がない

ユーザ・導入先への提供形態
- リアルタイムで自動で音声に判定がかかり、怪しい場合はアラートを通知するアプリケーションと基盤の提供
- 既存の通話や会議システムに対しても機能するアプリケーションとして提供も可能にする

2-2. 社会実装時のビジネスモデル等

社会実装時のビジネスモデル -自治体向け偽・誤情報総合対策-



自治体

- サービスを利用し、SNS上の分析情報の取得や、対策機能付きの投稿管理ツールを使用し、NABLAS/NTT/(提携SIer)に利用料を支払う

SNSユーザー

- 市民に提供されたアプリケーションを通じて、投稿チェック機能などを利用

NABLAS

- 「自治体向け偽・誤情報総合対策サービス」のサービス基盤やAPIを運営・提供し、サービス利用料/API利用料を収益とする

NTT東日本（販売・支援パートナー）

- 自治体に対し、ソリューション営業や導入・運用の支援

通信会社／SIerなど（ツール提供パートナー）

- SNS管理・分析ツールやWebアプリ、APIなどを介して、実務的なツールを自治体や一般ユーザーへ提供

ユーザ・導入先の詳細とそのペインポイント

- 自治体として有事の際に偽・誤情報が充満する中で、自治体に関連する投稿・情報の真偽や信頼性を判断する手段がない
- 自治体に属す市民が判断するすべがなく、困りごととなり問い合わせ対応が増加している
- 正しい情報を正しいと示していくことが難しくなってきている
- 日々問い合わせも増加し、課題意識も高まっているが知見やリソース不足を認識

ユーザ・導入先への提供形態

- さまざまな投稿や記事・動画をチェックすることや分析することができる総合分析プラットフォームの提供
- SNS投稿時に真正性を付与できる投稿サポートツール ＋ 市民向け検証ツールの提供

2-3. 技術開発及び社会実装にあたっての課題・展望

技術開発及び社会実装にあたっての今後の課題

電話音声フェイク検知

実用環境の多様性

着信・発信双方の音声圧縮方式やノイズキャンセル等の音声処理環境が多様に存在し、それらがデータに与える影響が大きい。全環境を網羅的にカバーすることは非常に難易度が高いことが判明。

最新生成モデルへの追随

最新モデルのElevenLabs v3等は、商用検出モデルを含む既存手法ではほぼ検出困難なレベルに到達しており、検知の難易度が急速に上昇。

自治体向け偽・誤情報総合対策

対策体制・業務フローの未整備

今後様々な課題やユースケースが存在しているが、現状それらを担当と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。

WM運用負荷と評価基準の未確立

電子透かし（WM）埋込・検証の操作ステップが多く導入障壁になりうる。また、偽・誤情報検知技術の標準的な評価基準・指標が未確立。

上記課題を踏まえた今後の展望

電話音声フェイク検知

環境適応型モデルの実装

推論時に各環境に合わせたモデルを自動ルーティングする仕組みを実装。環境変化に強い共通特徴を捉えるモデル・特徴量抽出手法の開発を推進。

アンサンブル型検知体制の構築

End-to-End（一気通貫型）学習モデルと特化型モデルを組み合わせた複合活用で、最新生成モデルにも対応可能な検知体制を構築。

自治体向け偽・誤情報総合対策

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、役割ごと提供・全体サポートできる支援サービスを構築し実証を推進。

運用の簡易化と標準化推進

WM埋込の自動化・ワンクリック検証の実現。評価データセット公開・ベンチマーク提供による業界標準化。KeiganAIプラットフォームへの速やかな機能統合による社会実装の加速。

自動ルーティング: 利用状況を判定し、最適なモデルを選択する仕組み
アンサンブル: 1つのモデルに頼らず、複合的に活用することで、最適な結果を選択・統合する手法

2-4. 事業の拡大に向けた中長期的な計画

電話音声フェイク検知ビジネスの展開

特定企業への先行導入で実績を構築し、通信インフラへの標準搭載を経て、あらゆる音声コミュニケーションの場にフェイク検知を提供するプラットフォームビジネスへ段階的に展開する

フェーズ1：FY2026
通信インフラへの標準搭載

- **主要通信網への標準オプション化**
  NTT東日本の「ひかり電話」や主要携帯キャリアの通信網に、フェイク検知機能を標準的なオプションとして組み込む
- **環境変化に強いモデルの構築**
  利用環境に応じ最適なモデルを自動選択する仕組みにより、環境に左右されない共通の特徴を捉え、最新のAIにも対応可能な検知機能をシステムに組み込む

フェーズ2：FY2027
領域の拡大と体制の確立

- **レベニューシェア体制の確立**
  レベニューシェアモデルを確立し、通信会社が安心・安全な通話環境を付加価値として提供する体制と実運用・改善のインフラを構築
- **電話音声検出の実績を活かし別領域に拡大**
  実環境で検知性能が確認された領域を中心に拡大展開し、実績とともに市場としての立場を確立する

フェーズ3：FY2028以降
多様な会議・対話プラットフォームへの拡張

- **全音声領域へ展開**
  固定電話だけでなく、オンライン会議システムやカスタマーサポート用の対話AIなど、あらゆる音声コミュニケーションの場へ機能を拡張する
- **業界標準化と評価透明性**
  「信頼できる音声環境」をグローバルな通信基準の一つとし、NABLAS独自のAIモデルを軸としながらも、業界全体の技術促進やインフラ基盤を推進していく

2-4. 事業の拡大に向けた中長期的な計画

自治体向け偽・誤情報総合対策サービスの展開

先進自治体での実証導入から全国展開・標準化を経て、偽・誤情報対策を社会インフラとして定着させ、AI時代の情報信頼性基盤を構築する

フェーズ1：FY2026 基盤構築と先行導入

- **先進自治体への実証導入**
  主要な都道府県や政令指定都市の中でも積極性の高い・フェイクリスクの高い自治体を中心に、実証実験を兼ねた導入を進める
- **NTT東日本と連携し体制の確立**
  地方自治体や警察への導入支援体制を確立し、分析・追跡機能を含めサービス提供する
- **API連携によるSNS管理の強化**
  APIやアドインツールの拡充により、既存SNSやその管理ツールとの連携を深め、運用の効率化を図る

フェーズ2：FY2027 全国展開と標準化

- **分析プラットフォームの全国展開**
  導入自治体や警察を全国規模に拡大し、自治体間での継続的な分析プラットフォームとしての機能と運用基盤を強化する
- **ベンチマーク・規格の標準化推進**
  自治体間での連携やベンチマーク・規格などの標準化も進める

フェーズ3：FY2028以降 社会インフラ化と一般普及

- **一般ユーザーへの投稿チェック機能の普及**
  一般のSNSユーザー向けの「投稿チェック機能」を普及させ、個人利用プランやそのカスタマイズ機能なども広く提供し、社会全体のAI時代のメディアリテラシー向上に寄与する
- **社会的防壁としての地位確立**
  自治体内での周辺機能の強化や、公共性の高い民間企業へも展開し、偽情報に対する社会的な防壁としての地位を確立する