2025年02月28日　技術情報

金融包摂実現を支えるFintechのためのイノベーション

音声AIを活用したデジタルシステムとサービスの開発

#イノベーション創生　#AI・生成AI　#IoT・データ利活用　#デジタルソリューション

執筆者

Mayurakshi Mukherji
R&D Centre, Hitachi India Pvt. Ltd.
現在，音声・音響関連テーマの業務に従事
IEEE会員

Shreyas Kulkarni
R&D Centre, Hitachi India Pvt. Ltd.
現在，音声AI，生成AI，最適化に関する業務に従事

Vivek Kumar
R&D Centre, Hitachi India Pvt. Ltd.
現在，生成AI，NLP，クラウドコンピューティング，フルスタック開発関連の業務に従事

Rahul Mishra, Ph.D.
R&D Centre, Hitachi India Pvt. Ltd.
現在，音声・音響関連テーマの業務に従事

Munender Varshney, Ph.D.
R&D Centre, Hitachi India Pvt. Ltd.
現在，音声・音響関連テーマの業務に従事
IEEE会員

Thiruvengadam Samon
R&D Centre, Hitachi India Pvt. Ltd.
現在，Fintechと持続可能なモビリティソリューションに関するAIソリューションの業務に従事

Senthil Raja G, Ph.D.
R&D Centre, Hitachi India Pvt. Ltd.
現在，音声・音響関連テーマの業務に従事
IEEE会員

Kingshuk Banerjee, Ph.D.
R&D Centre, Hitachi India Pvt. Ltd.
Hitachi India Pvt. Ltd.における研究開発のディレクター業務に従事

ハイライト

インドでは社会のさまざまな層にスマートフォンが普及したことを受け，政府主導の「デジタル・インディア計画」の下，金融，医療，農業などの重要な公共サービスの基盤となる開かれたデジタルシステムの構築をめざしている。その実現に向けては，AIの活用が不可欠である。Hitachi India Pvt. Ltd.の研究開発部門では，特に銀行などの重要な公共サービスへのアクセスを容易にする手段として，最も基本的なコミュニケーション手段である音声の研究に取り組んでいる。

本稿では，音声を用いたファイナンシャル・インクルージョン（金融包摂）実現に向けた三つの重要なアプローチとして，音声に基づく認証，現地語の音声認識，自動話者照合機能のエッジデバイスへの実装について解説する。このうちエッジデバイスへの実装は，スマートフォンのアプリで小型の音声認識エンジンを実行可能とすることで，高レイテンシかつ，オフラインも含む狭帯域幅の状況下においても，瞬時の推論を可能にするものである。

また，インドの現地語の認識に焦点を当てたゼロからのAIモデルアーキテクチャの開発，小型化のためのニューラルネットワークモデルの適切な量子化，精度に関する課題についても述べる。最後に，音声ベースの取引では「金額」を正確に表現することがきわめて重要であることから，連続した数字の発話音声認識に焦点を当てた関連データセットの構築アプローチについても紹介する。

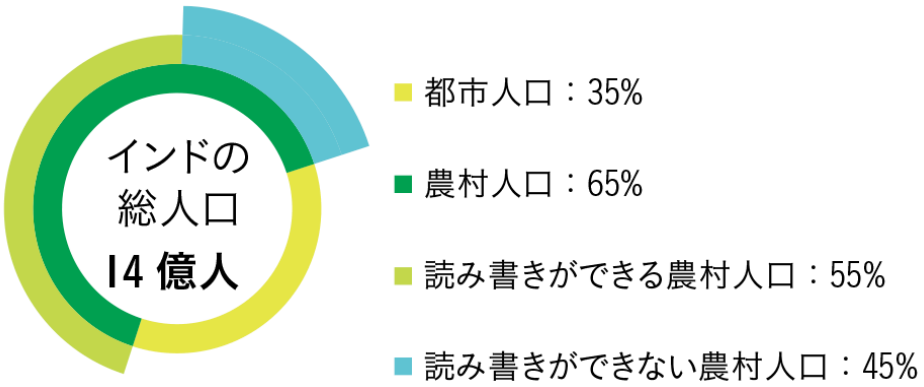
1. はじめに

近年，インドではUPI（Unified Payments Interface）によるQRコード決済※1やeウォレットなどのデジタル決済システムが広く普及しており，スマートフォンアプリを利用することで，銀行に出向いたり現金を使用したりすることなく，送金，請求書の支払い，買い物を簡単に行えるようになっている。しかし，インドは農村部の人口が多く，その多くが文字の読み書きに課題を抱えている（図1参照）。したがって，インド政府が主導する「デジタル・インディア計画」の下，すべての国民に対してデジタル技術を活用した銀行取引および金融サービスを提供するためには，銀行アプリケーションに現地語の音声による支援を導入することが不可欠である。

本稿の2章ならびに3章では，音声認識，話者の照合と識別，音声合成技術など，複数のAI（Artificial Intelligence）技術の組み合わせによる効率的でスムーズなバンキング体験の構築について述べる。一方，AIモデルは計算コストが高く，クラウドAPI（Application Programming Interface）で展開する場合は十分なネットワーク帯域幅が必要なため，ユーザーエクスペリエンスが制限される。そこで，エッジ上にAIを展開することで，スマートフォン上のローカルの計算リソースを活用してこの問題に対処し，AIシステムの待ち時間を大幅に短縮した。さらに，4章では，データのセキュリティとプライバシー向上についても解説する。

※1） QRコードは，株式会社デンソーウェーブの登録商標である。

図1｜インドの人口と居住地・識字率の分布

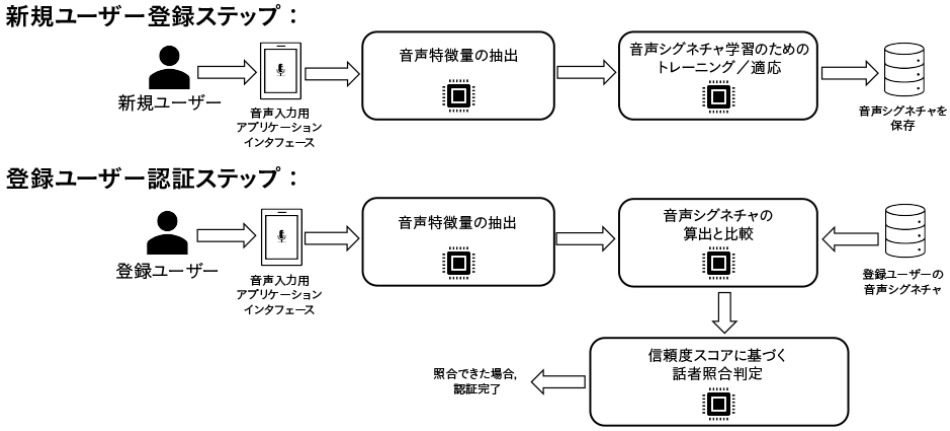


インドの農村部には読み書きに課題を抱える人々も多く存在する。これに対しインクルーシブなFinTechは，現地語の音声という最も自然なコミュニケーション手段を用いたコマンド入力により，そうした人々が銀行のサービスやアプリにアクセスすることを可能にする。さらに，インドでは人口の大半が帯域幅の狭い農村に居住しているため，インターネットでやり取りできる情報量には制限がある。このため，エッジデバイス（スマートフォン）に音声ベースのソリューションを導入する必要がある。日立が開発した現地語のASR（Automatic Speech Recognition：自動音声認識）とエッジベースのVA（Voice Authentication：音声認証）は，インドにおけるファイナンシャル・インクルージョンに向けたFinTechの実現を支援する。

2. 音声ベースの認証

VA（Voice Authentication：音声認証）は通常，登録と認証という二つのステップで構成される（図2参照）。登録ステップでは，ユーザーが音声サンプルをVAシステムに提供して自分の声を登録する。この音声サンプルは，AIモデルによってユーザー固有の音声シグネチャに変換され，システムに保存される。認証ステップでは，ユーザーから提供された音声サンプルをシステムに保存されている音声シグネチャと比較し，両者の音声がどの程度似ているかを定量化した類似度スコアを生成する。類似度スコアが事前に設定された閾値より高い場合，ユーザーは認証される。

図2｜登録ステップと認証ステップを含むVAパイプライン



VAは，登録と認証という二つの主なステップで構成される。登録ステップでは，ユーザーが音声サンプルを提出して自分の声を登録する。この音声サンプルは音声特徴量の抽出に使用される。次に、複数の話者の音声サンプルに基づいて学習したResNet34モデルにより，音声特徴量から音声シグネチャが算出され，システムに保存される。

認証ステップでは，登録ユーザーが認証時に新しい音声サンプルを入力すると，登録時と同じResNet34モデルによって音声シグネチャが算出され，登録ステップで登録された音声シグネチャと比較される。比較スコア（信頼度スコア）が閾値より高い場合，ユーザーは認証される。

VAはさまざまな分野に適用することが可能であり，ここでは銀行アプリへの応用に焦点を当てる。リテールバンキングアプリでは，VAはログインパスワードの代わりとして，あるいは少額決済の送金，取引履歴の確認時にも利用できる。QRコードを利用したUPI決済が普及しているインドでは，ほとんどの店舗が決済時に銀行口座のQRコードを顧客に提示して決済を促す。QRコードには，（1）受取人の銀行口座情報のみを含む静的QRコード（顧客が支払金額を入力する必要がある）と，（2）受取人の銀行口座情報と支払金額を含む動的QRコード（顧客が支払金額を入力する必要がないもの）の2種類がある。VAとASR（Automatic Speech Recognition：自動音声認識）を組み合わせることで，加盟店は「300ルピーのQRコードを生成して」といった音声コマンドを使用して音声入力と本人認証を行うことで，動的QRコードを生成して顧客から支払いを受けることができ，迅速な決済が可能になる。これらは音声コマンドのみのワンステップで完了するため，プロセスが簡素化される。

支払いを迅速化してユーザーの利便性を向上させるには，認証時に必要な音声サンプルを短くすることが求められる。しかし，音声サンプルを短くすれば，ユーザー固有の音声情報が減り，精度が低下する。これに対し，登録ステップにおいて対象ドメインに特化した音声サンプルを使用し，保存された音声シグネチャと認証音声サンプルが重なる部分を大きくすることで問題の軽減を図った。日立は，WeSpeaker[1)]（ResNet34アーキテクチャ[2)]をベースとしたオープンソースの自動話者照合モデル）を使用して，登録音声サンプルのテキストコンテンツを慎重に設計することで，音声認証の結果を改善している。さらに，利便性や安全性の高さといったユースケースの要件に基づいて類似度スコアの閾値を調整することで，精度を向上させることができる。また，WeSpeakerの基本モデルをファインチューニングすることで，特定のユースケースに特化した音声サンプルで構成されるテストデータセットでもパフォーマンスを向上できることを確認した。ファインチューニングは，ヒンディー語で複数桁の数字を発音した2～3秒間の音声サンプルを用いて実行する。2秒間のヒンディー語音声サンプルのユースケースに特化したデータセットでテストした結果，EER（Equal Error Rate：等価エラー率）はWeSpeakerの基本モデルが4.2%であったのに対し，ファインチューニングしたモデルを使用した場合は1.96%に低減された。

現在，銀行業務におけるVAの利用はテレフォンバンキングやIVR（Interactive Voice Response）システムに限られている。これらのシステムでは，ユーザーの音声シグネチャを生成するために長い音声サンプルを複数セット収集する必要があるうえ，認証時にも長い音声サンプルの入力が求められる。これによって利便性が妨げられ，音声技術を利用したデジタルバンキングの導入が進まない要因となっている。そのため，利便性に優れた音声AIによるFinTechソリューションを実現するには，短い音声サンプルによる高精度のVAシステムが求められている。

3. 現地語による連続数字音声認識

CNR（Connected Number Recognition）とは，ASRの1分野であり，連続する数字列の発話音声の認識に特化したものである。例えば，*Toḷāyirattu aintu*というタミル語のフレーズは数字の「905」に相当する。インドの銀行業務では，ユーザーとのやり取りを簡素化するためにCNRの重要性が増している。複雑なメニューシステムを操作する代わりに，音声によって取引や問い合わせを行うことができるため，顧客の利便性が向上する。このシステムをインドのより広い地域で利用できるようにするためには，各地で用いられる現地語で機能させる必要がある。ユーザーが自身の現地語で対話することが可能になれば，アクセシビリティの向上だけでなく，文化の保護にも貢献でき，テクノロジーをよりインクルーシブなものにすることができる。また，より多くの人々がデジタルサービスを利用できるようにすることで，教育面や経済面の支援にもつながり，デジタル時代の進展に伴うコミュニケーションギャップを埋めることが可能である。金融など，音声認識の精度が重要な分野では，連続数字列の発話音声を認識することが頻繁に求められており，本研究はこうした課題に対応することを目的としている。

タミル語とヒンディー語でのCNRについて，既存の最先端のASRモデルの性能を分析した結果，これらの言語で連続数字を扱うと，すべてのモデルで著しい性能低下が見られることが判明した。関連データセットを作成するため，タミル・ナードゥ州内の複数地区からタミル語の音声サンプルを，インド北部の複数の州からヒンディー語の音声サンプルを収集した。サンプル提供者にはランダムに数字を提示して日常的な状況で話すのと同じようにその数字を読み上げてもらい，録音した音声をモデルのトレーニングとテストに使用した。このようにして，インドの10地域から複数の現地語サンプルを収集した。

メインのモデルであるLSTM-TDNN（LT-Kaldi）は，Kaldi音声ツールキットの一部である。これは，六つの畳み込み層と15の時間遅延ニューラルネットワークで構成され，3,100万個のパラメータを持つ[3)]。Kaldiの標準的なトレーニングレシピに従ってCMN（Cepstral Mean Normalization：ケプストラム平均正規化）を施した高解像度MFCC（音声サンプルから抽出した特徴量）を用い，このモデルの学習を行った。さらに，タミル語とヒンディー語の公開データセットを用いてファインチューニングした二つの事前学習モデル，Wav2Vec2.0[4)]とWhisper[5)]についても評価を行った。

性能面では，ベースラインのLT-Kaldiモデルは，タミル語で15%，ヒンディー語で7%のWER（Word Error Rate）を達成した。スペクトルゲーティング，スペクトルサブトラクション，ダイアリゼーションなどの前処理法も検討したが，わずかな改善（タミル語で2%，ヒンディー語で1%）にとどまった[6)]。

LT-KaldiをWav2Vec2.0およびWhisperと比較したところ，Wav2Vec2.0はタミル語で98%，ヒンディー語で71%のWERを示し，Whisperはタミル語で93%，ヒンディー語で85%のWERを示した。これらの結果は，LT-KaldiがCNRの堅実なベースラインとなること，一方でWav2Vec2.0やWhisperのような最先端のモデルでCNRの性能を向上させるためには，特定のデータセット上でファインチューニングを行う必要があることを示唆している。

4. エッジデバイスにおける音声ベースのAI

音声データをローカルで処理し，速度と安全性，アクセシビリティに優れたサービスを実現できるエッジデバイス上の音声ベースのAIソリューションは，インクルーシブなFinTechのカギである。音声処理をクラウドに頼らずエッジに移行することで，レイテンシを短縮し，帯域幅の使用を最小限に抑えるとともに，プライバシーを強化することができる。これは，迅速かつ正確な応答とデータのセキュリティがきわめて重要な金融サービスにおいて特に価値があるポイントである。また，エッジベースの処理により，インターネット接続に制約のある地域でもサービスへのアクセスが可能となり，デジタルディバイドの解消にも貢献できる。

エッジにおけるAIベースのソリューションには，レイテンシ，精度，モデルサイズなどの指標がある。エッジでの音声ベースのAIソリューションの場合，特に取引やサービスの問い合わせにリアルタイムで応答するためには低遅延が不可欠であり，地域ごとに特色のあるアクセントや，さまざまな環境下で録音された音声でも確実に理解できる高い精度が求められる。また，スマートフォンや組み込みシステムの限られたストレージに収まるようモデルサイズを小さくする必要もある。これらの指標を総合的に評価することで，当該モデルがエッジ展開に実用的かつ効率的かどうかを判断できる。実際に，スマートフォンのバーチャルアシスタント（Googleアシスタント※2やSiri※3）では，クラウドへのオフロードの必要性を減らすために一般的なコマンドをデバイス上で処理している。これらは，エッジ端末での音声AI活用の可能性を示している。

本研究の一環として，日立は第2章で述べたVAモデルに，AIモデルのエッジ実装の概念を適用することに注力している。そして高精度，低レイテンシ，モデルサイズの削減を達成するため，主にモデルの量子化として知られる技術を探求している。

量子化を行うと，モデルの重みのビット幅が少なくなり，それによってモデルのサイズと推論時のレイテンシの両方が削減される（図3参照）[7)]。一方，精度についてはトレードオフの関係が存在する。元の32ビットモデルを8ビットモデルに量子化し，WeSpeakerモデルをエッジデバイス（Raspberry Pi 4※4）で評価した場合，モデルサイズは1/4に，推論レイテンシは1/2に削減された一方で，EERは2%増加した。日立が用いた静的量子化とは，キャリブレーションデータセットと呼ばれる小規模な代表的データセットを用いて，モデルの重みと活性化の両方を量子化する手法である。精度向上の課題の対策として，対象ドメインに特化したデータを使用して元のモデルをファインチューニングし，量子化によるEERの増加を補償することを提案している。

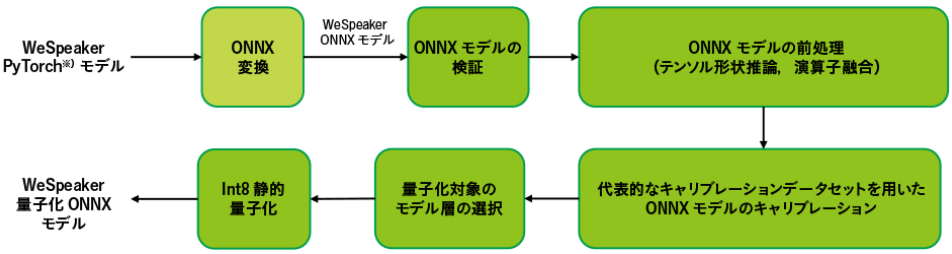
さらに，ONNX（Open Neural Network Exchange）フォーマットも使用している。ONNXはさまざまなフレームワーク間の互換性をサポートしており，デバイスの種類が多様でハードウェアに制約があるエッジ環境への適用に最適である[8)]。ONNXフォーマットを活用することで，日立の量子化VAソリューションのAndroid※2）スマートフォンへの展開を検討することが可能となる。エッジ処理，効率的なモデルフォーマット，デバイスの互換性を組み合わせることで，音声AIは誰もが簡単にアクセスできるインクルーシブなFinTechのための強力なツールとなる。

※2） Google AssistantおよびAndroidは，Google LLCの商標または登録商標である。

※3） Siriは，Apple Inc.の商標または登録商標である。

※4） Raspberry Piは，Raspberry Pi Foundationの商標である。

図3｜オープンソースRawNet3 PyTorchモデルの量子化ONNXモデル構築のパイプライン



注：略語説明　ONNX（Open Neural Network Exchange）

※）PyTorchは，Linux Foundationの商標である。

オープンソースのPyTorchモデルをONNXフォーマットに変換し，機能を検証した。このONNXモデルには，前処理としてモデル演算子の融合とモデルテンソル形状の推論を実施した。前処理したモデルに対して代表的なデータセットでキャリブレーションを行い，量子化の対象とするモデル層を選択した。最後にモデル量子化を行い，元のサイズの1/4に縮小した。また，量子化したモデルをRaspberry Pi 4やスマートフォンなどのエッジデバイスに実装すると，推論レイテンシも短縮される。

5. おわりに

音声ベースのAIを活用したファイナンシャル・インクルージョンは，VA，CNR，エッジベースの音声AIソリューションを組み合わせることで，金融サービスへのアクセスを向上させるとともに，VAによる安全でシームレスなアクセスを提供する。また，CNRにより，読み書きに課題を抱えるユーザーやデジタルデバイスの操作に不慣れなユーザーでも簡単にやりとりすることが可能となる。さらに，接続環境が整っていない地域でもエッジ処理によって信頼性の高いサービスを提供することができる。これらの技術を組み合わせることで，多様なユーザーのニーズに応える，アクセスしやすい金融エコシステムが実現できると考えている。一方で重要な課題も残されており，特に量子化精度の向上や，計算リソースの制約によって性能に影響の出るデバイスへのモデル展開では，最適化が求められる。

ファイナンシャル・インクルージョンの取り組みは，十分な金融サービスが行き渡っていない地域のエンドユーザーに利益をもたらすとともに，FinTech企業にとってはより広範な市場に参入し，金融イノベーションにおけるリーダーシップを示す機会となる。今後，技術のさらなる進歩に伴って，より幅広いユー

ザー層にリーチし，よりインクルーシブなデジタル経済を確立するためには，こうしたソリューションを追求することが求められる。

参考文献など

1) H. Wang et al., "WeSpeaker: A Research and Production Oriented Speaker Embedding Learning Toolkit," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1-5 (2023).
2) He, K. et al., “Deep residual learning for image recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016).
3) D. Povey et al., “The Kaldi speech recognition toolkit,” IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (2011).
4) A Baevski, et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in neural information processing systems (2020).
5) A. Radford, et al, "Robust speech recognition via large-scale weak supervision." In International conference on machine learning, pp. 28492-28518. PMLR (2023).
6) R. Mishra, et. al., “Revisiting Automatic Speech Recognition for Tamil and Hindi Connected Number Recognition,” In Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages, pp. 116-123 (2023).
7) Quantization
8) ONNX

© Hitachi, Ltd. 1994, 2025. All rights reserved.

日立評論

日立評論は、イノベーションを通じて社会課題に応える
日立グループの取り組みを紹介する技術情報メディアです。

日立評論Webサイトでは、日立の技術者・研究者自身の執筆による論文や、
対談やインタビューなどの企画記事、バックナンバーを掲載しています。ぜひご覧ください。

日立評論(日本語) Webサイト

https://www.hitachihyoron.com/jp/

Hitachi Review(英語) Webサイト

https://www.hitachihyoron.com/rev/

日立評論メールマガジン

Webサイトにてメールマガジンに登録いただきますと、
記事の公開をはじめ日立評論に関する最新情報をお届けします。