声認証技術がもたらす安全・安心で便利な社会

越仲 孝文　リー コン エイク

要 旨

私たちが日常のコミュニケーションに使う音声は、最も簡便で手軽な情報伝達の手段です。音声の個人性に基づく生体認証技術・声認証は、生体認証が本来有する安全・安心に加えて、便利さを提供します。本稿では、昨今の人工知能（AI）技術の中心である深層学習（ディープラーニング）との関わりも含め、声認証技術を概説するとともに、世界最高水準の声認証技術を有するNECの取り組みについて紹介します。また、パブリックセーフティをはじめとして、今後の普及が期待される声認証技術の産業応用の可能性についても述べます。

KeyWords　話者照合／話者識別／話者認識／深層学習／音声認識

1. はじめに

私たちは日常生活において、さまざまな手段で人とコミュニケーションを取っていますが、話す・聞くコミュニケーションはそのなかでも最も基本的なものといえます。話す・聞くという行為は、電子機器はおろか、紙と鉛筆すら必要としません。人間にとって、これ以上に簡便で気軽なコミュニケーション手段はないでしょう。

話す・聞くコミュニケーションの媒体である音声は、人間の生体情報の一種であり個人性を有するため、生体認証に用いることができます。音声を用いた生体認証が、声認証です。声認証は、ユーザーにとって簡便で気軽な個人認証の手段を提供します。加えて、既存のマイクや電話などの機器が利用でき、高価で特別なセンサー機器を必要としないことから、システム構築の面からも安価で手軽な生体認証といえます。

本稿では、声認証技術を、昨今の人工知能（AI）技術の中心である深層学習（ディープラーニング）との関わりも含めて解説するとともに、世界最高水準の声認証技術を有するNECの取り組みについて紹介します。また、パブリックセーフティを含む産業応用の可能性についても述べます。

2. 声認証技術

どこからか人の声が聞こえたとき、たとえその人の姿が見えなかったとしても、「あの声は○○さんね」と察しがついたという経験は誰しもあるでしょう。人の声、すなわち音声には、声帯や口腔などの発声器官の形状（身体的特徴）に起因する個人性を含みます。また、話し方には癖のようなもの（行動的特徴）があり、これも人それぞれ違います。このような個人性に関わる特徴を音声から抽出して話し手（話者）を認識する技術が、声認証です。

2.1 声認証の技術要素

声認証は、技術的には話者認識（Speaker Recognition）、あるいは話者照合（Speaker verification）と呼ばれ、多くの場合、1対の音声を比較してそれらが同一人物のものか否かを推論する技

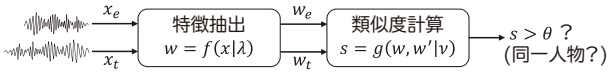


図1　声認証システムの基本構成（対比較）：特徴抽出、類似度計算のモデル（λ, ν）は学習によりデータから決定される

術を指します。1対多の比較（話者識別：Speaker identification）も、つまるところ対比較の反復に帰着するので、対比較が処理の基本単位となります。その構成を**図1**に示します。

特徴抽出では、i-vector（アイベクター）と呼ばれるフレームワーク[1)]が、現在もっともポピュラーです。i-vectorは、多数の話者の音声から作られた音韻（各種の母音、子音）の標準モデルを用いて、この標準モデルと入力音声の差分を特徴量として抽出します。単純にすべての差分を取ると10万次元に及ぶ巨大な特徴量となるので、因子分析の手法を用いて400次元程度のコンパクトな特徴量に圧縮するのがi-vectorのポイントです。類似度計算には確率的線形判別分析（Probabilistic Linear Discriminant Analysis：PLDA）というモデル[2)]が、よく用いられます。PLDAは古典的な機械学習の手法、LDAを確率論的に再定式化したモデルで、400次元のi-vector特徴量から話者の識別に有効な特徴量を自動的に選択し、尤度比（ゆうどひ）の形式で類似度を計算します。

i-vectorとPLDAは、いずれもガウス分布（正規分布）仮定に基づく確率モデルによって定式化され、大量のデータから最適なモデルパラメータを自動学習できます。機械学習のさまざまなテクニックが導入されて発展したi-vectorとPLDAは、声認証におけるデファクト標準となっています。

2.2 深層学習の導入

近年、画像認識や音声認識の分野では、深層学習（ディープラーニング）を適用して精度向上を図る試みが数多くなされています。声認証の分野でも、2014年頃から深層学習に関する研究が活発化し、最近、一つのパラダイムシフトが起こりつつあります。

深層話者埋め込み（Deep speaker embedding）あるいはx-vectorと呼ばれるその方式は、声認証の精度を大幅に向上させ、従来のi-vectorに代わる新たな特徴抽出器として、この分野の研究者の間に急速に広まっています[3)]。その概念を**図2**に示します。まず、特徴抽出部と識別部からなる深いニューラルネットワークを、音声から話者を正しく推論できるように訓練します。こうしてできたニューラルネットワークの特徴抽出部は、音声から話者の識別に有用な情報のみを取り出す優れた特徴抽出器となります。

音声はその長さが一定ではない可変長の時系列データであり、したがって、ニューラルネットワークに入力するデータ量も可変です。この点が画像と違う音声の扱いの難しいところですが、x-vectorでは、時間方向にデータを集約するプーリング層を特徴抽出部の途中に設けることにより、一定次元数の特徴量を出力します。

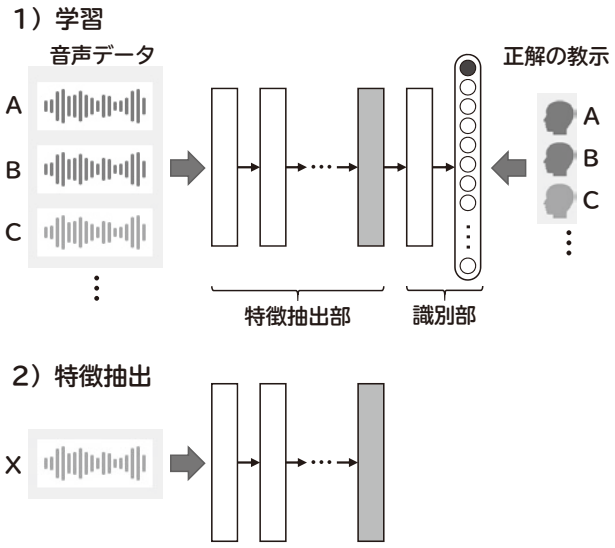


図2　深層学習に基づく新たな特徴抽出のパラダイム

声認証への深層学習導入の試みは、特徴抽出にとどまらず、フロントエンド（雑音下での音声/非音声識別、音声強調など）からバックエンド（類似度計算）まで多岐にわたります。また、すべての技術要素をニューラルネットワークに置き換えてシステム全体を一括学習するEnd-to-endシステムも登場しています[4)]。このような流れは、今後も続くとみられています。

2.3 NECの取り組み

NECは近年、指紋や顔に続く次代の生体認証の一つとして声認証の技術開発に力を入れ、世界をリードする水準にまで技術を高めています。

特に深層学習の導入については、早い時期から研究を始めて、多くの技術成果を上げています。例えば、雑音環境下で非音声と音声を識別して音声区間のみを正確に検出する技術[5)]、雑音が重畳した音声の特徴量から雑音成分を除去する技術[6)]、個人性に関する情報が得難い短時間の音声から長時間相当の特徴量を復元する技術[7)]などの独自技術があります。

またNECは、米国国立標準技術研究所（NIST）が

開催するベンチマークテスト、Speaker Recognition Evaluation（SRE）[8]に積極的に参加しています。SREは、世界中の産官学研究機関60チーム以上が参加し、おのおのの開発したシステムの認証精度を同一のデータセットで競うコンペティションです。このコンペティションで、NECは高い技術力を実証しています。

2018年のSRE では、背景雑音、劣悪な通信回線のもとでの電話会話から特定の人物を見つける課題、及びYouTubeに代表されるネット動画に登場する複数の人物から特定の人物を見つける課題という、2種類の課題でテストが行われました。いずれの課題も技術的に厳しい条件設定で難易度が高く、例えば電話会話の課題では、NISTより提示されたベースラインシステムの精度はわずか88.8%（等誤り率11.2%）でした。これは決して、NISTのベースラインシステムの技術水準が低いということではありません。事実、このベースラインシステムは前述のx-vector特徴抽出を搭載した最新鋭システムです。これに対してNECのシステムの精度は95.0%（同5.0%）、誤り率で最新鋭システムの半分未満という卓越した成績を上げました。

この種のシステム開発では、あらゆる技術要素のレベルを極限まで高める必要があります。とりわけNECは、x-vectorに改良を加えた独自の特徴抽出方式を新規開発し、顕著な精度向上に成功しました。この方式では、x-vectorに注意機構（Attention mechanism）と呼ばれる補助ネットワークを加えて、入力音声のなかで個人の特徴がよりよく表れている箇所を自動的に選択する機能を実現しました[9]。また、一般に膨大な訓練データを必要とする深層学習において、限られた音声データに変換を施し、見かけの話者数を数倍に増強するデータ拡張の新手法を開発、深層学習の実力を最大限発揮できるよう工夫を施しました。

3. 産業応用

最後に、声認証技術が社会でどのように役立つのかを見ていきましょう（**図3**）。

Eコマース：近年、少額のクレジットカード払いの多くはサインレスになっています。煩雑な手続きを省略して購買行動の障壁を減らすことが、売り手と買い手の双方にメリットをもたらすからです。昨今の流通業界では、安全・安心に加えて利便性が求められているといえます。本稿の冒頭でも述べたように、音声は人々が日々のコミュニケーションで使う手軽なメディアであり、音声を用いた生体認証はユーザーにとって簡便で気軽な本人確認の手段を提供します。声認証は、Eコマースやテレホン/ネットバンキングなどの商取引における本人確認に適した認証手段です。

コールセンター業務：企業の顧客志向の高まりに伴い、コールセンターなどの顧客接点でのサービス品質向上がさまざまな業種で推進されています。そのなかで、頻繁に電話をする重要顧客の本人確認手続きの簡略化（**図4**）、苦情が多いなど問題視される顧客の早期特定などが、現場の課題として挙げられます。声認証は、相手の姿が見えない電話で使える唯一の生体認証であり、自然な会話から顧客を特定できることから、コールセンター業務支援に有効です。

犯罪捜査：近年、振り込め詐欺をはじめとした電話を使った犯罪を撲滅するために、さまざまな対策が講じられています。しかし、この種の犯罪は巧妙化、組織化し、衰えることがありません。声認証は、犯罪者の足取りを追い捜査を支援する分析ツールとしての活用が期待されています。また、電話やソーシャル・ネットワーキング・サービス（SNS）などでの犯罪組織の動きを監視する手段としても、有用です。音声を手掛かりとした分析は、近年街頭で急速に普及する監視カメラ以上に、目に見えない水面下の予兆を発見

図3　声認証の応用場面は多岐にわたる

図4　コールセンター業務支援：迅速に顧客IDを確認

し、犯罪の抑止に寄与すると期待されます。

その他：近年普及が著しいスマートスピーカ、スマートイヤホンなどのヒアラブルデバイス[10]、ロボットなどのユーザーフレンドリーなインタフェースとして、声による個人認証の実用化が期待されています。

4. おわりに

本稿では、音声を用いた生体認証である声認証について、その技術概要、近年の深層学習との関わり、世界をリードするNECの取り組みについて紹介しました。また、安全・安心に加えて、高い利便性を有する声認証の産業応用の可能性について述べました。声認証がこれからの生体認証技術として、各方面で社会課題の解決に貢献することを期待します。

参考文献

1) Najim Dehak et al. : Front-End Factor Analysis for Speaker Verification, IEEE Transactions on Audio, Speech, and Language Processing, Vol.19, pp.788-798, 2011.5
2) Simon J. D. Prince et al. : Probabilistic Models for Inference about Identity, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.34, 2012.1
3) David Snyder et al. : X-vectors: Robust DNN Embeddings for Speaker Recognition, 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018.4
4) Georg Heigold et al. : End-to-end Text-dependent Speaker Verification, 2016 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2016.3
5) Hitoshi Yamamoto et al. : Robust i-vector extraction tightly coupled with voice activity detection using deep neural networks, 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017.12
6) Shivangi Mahto et al. : I-vector Transformation Using a Novel Discriminative Denoising Autoencoder for Noise-Robust Speaker Recognition, INTERSPEECH, 2017.8
7) Hitoshi Yamamoto et al. : Denoising Autoencoder-Based Speaker Feature Restoration for Utterances of Short Duration, INTERSPEECH, 2015.9
8) 米国国立標準技術研究所：Speaker Recognition https://www.nist.gov/itl/iad/mig/speaker-recognition
9) Koji Okabe et al. : Attentive Statistics Pooling for Deep Speaker Embedding, INTERSPEECH, 2018.9
10) 荒川 隆行：人によって異なる耳穴の形状を音で識別する耳音響認証技術, NEC技報, Vol.71 No.2（本特集）, 2019.3

執筆者プロフィール

越仲 孝文
バイオメトリクス研究所
主幹研究員
IEEE、電子情報通信学会(IEICE)、人工知能学会(JSAI)、日本音響学会(ASJ)、各会員
工学博士

リー コン エイク
バイオメトリクス研究所
主幹研究員
IEEEシニア会員、International Speech Communication Association (ISCA)、Asia-Pacific Signal and Information Processing Association (APSIPA)、各会員
工学博士

NEC 技報のご案内

NEC技報の論文をご覧いただきありがとうございます。
ご興味がありましたら、関連する他の論文もご一読ください。

NEC技報WEBサイトはこちら

NEC技報（日本語）　NEC Technical Journal（英語）

Vol.71 No.2　バイオメトリクスを用いた社会価値創造特集

バイオメトリクスを用いた社会価値創造特集によせて
社会価値の創出に貢献するNECの生体認証への取り組み

◇ 特集論文

NECが推進するバイオメトリクスの取り組み
NECの生体認証ブランド「Bio-IDiom（バイオイディオム）」
バイオメトリクス研究の今後の進化発展
バイオメトリクス事業におけるプライバシーへの配慮

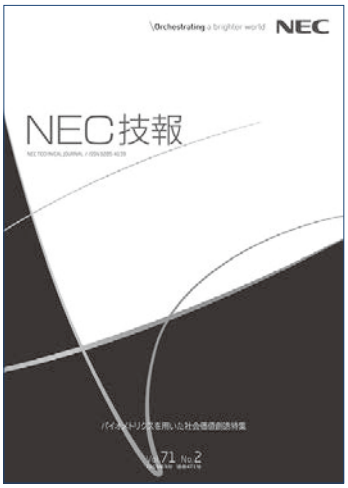
バイオメトリクスを用いたサービス・ソリューション
Western Identification Network：連携型アーキテクチャが提供するサービスとしての生体認証
マイナンバーカードに関わる顔認証システムの活用
顔認証クラウドサービス「NeoFace Cloud」
高度映像分析ソリューションを提供するNEC 映像分析基盤
将来のリテールサービスを支える生体認証技術による新しい店舗ソリューション
ユーザーが使いたい金融サービスを即時利用可能にする「本人確認サービス」の提供
バイオメトリクスを活用した非日常空間体験向上の取り組み
顔認証と位置情報を活用した建設現場における現場作業員の入退場管理サービス
次世代ものづくりの現場における個人特定の重要性

バイオメトリクスを支えるコア技術・先進技術
安全・安心な社会を実現する顔認証・人物照合技術
フュージョン照合を活用した虹彩認証高度化技術
新特徴量を利用した遺留指紋照合高度化技術
声認証技術がもたらす安全・安心で便利な社会
人によって異なる耳穴の形状を音で識別する耳音響認証技術
映像から不審者を高精度で絞り込む行動パターンの自動分類
安価なIoT端末上で動作する顔映像からの眠気推定技術

◇ NEC Information

C&Cユーザーフォーラム＆iEXPO2018　Digital Inclusion
基調講演
展示会報告

NEWS
2018年度C&C賞表彰式典開催



Vol.71 No.2
（2019年3月）

特集TOP