4. 音声による本人認証

第2部 話者認証システム

（株）アニモ　開発推進部
鈴木　晃

音声による個人認証の利点は，新たな機器・インフラを新規に整備する必要がない点である．すなわち，音声を通す媒体として，電話網がすでにほぼ100％近い普及率で各世帯・各事業所に整備されている．したがって，音声による話者認識（本人確認）技術が実用化されれば，電話を通して，「いつでも，どこでも，だれでも」本人であることが証明可能となる．ここでは，この話者認識技術の具体的実現について，事例を絡めて述べる．

□ 話者認識エンジン □

テキスト依存型の商用話者認識エンジンとしては，米国においてはTI社の技術を応用したSprint社の話者認識サービス，T-NETIX社の開発したSpeakEZがあり，これらは日本においても導入されている例がある．

また日本ではアニモと富士通によりVoiceGATEの開発・製品化が行われている．VoiceGATEはテキスト依存型話者認識を実行する下位レベル機能を提供したものであり，現在はWindowsおよびWindowsNT上で動作可能であり，UNIX系への移植も計画中である．その性能を**表-1**に示す．

表-1中の本人拒否率の算定は50名が異なった日に発声した10回の音声データによるものであり，他人受入率は20名がそれぞれ他の19名分のキーワードを発声した3回の音声データによるものである．

識別アルゴリズムは言語構造に依存しないため，使われる言語には制約がない．話者認識技術には，「与えられた音声が，該当する話者であるかどうかを判定する」話者照合（Speaker Verification）と「あらかじめ登録された話者集合の中から，入力音声に最も近い話者を選択する」話者識別（Speaker Identification）があるが，このエンジンでは話者照合の機能が実現されている．

次にテキスト独立型の話者認識エンジンの例としては，アニモが開発・製品化したVoiceSyncがある．これはテキスト独立型話者認識を実行する下位レベル機能を提供するものであり，現在はWindowsおよびWindowsNT上で動作する．その性能を**表-2**に示す．

表-2中の本人拒否率の算定は，24名が異なった日に発声した10回の音声データによるものであり，他人受入率は20名が異なった日に発声した5回の音声データによるものである．識別アルゴリズムは言語構造に依存しない

認証方式	テキスト依存型
認証時間	0.05秒
音声品質	電話音声
本人拒否率（FRR）	1％以下
他人受入率（FAR）	5％以下
対応言語	マルチリンガル対応
認証パターン	話者照合
特殊ハード	不要

表-1　VoiceGATEの仕様概要

認証方式	テキスト独立型
認証時間	10秒～
音声品質	携帯電話音声
本人拒否率（FRR）	1％以下
他人受入率（FAR）	5％以下
対応言語	マルチリンガル対応
認証パターン	話者照合・話者識別
特殊ハード	不要

表-2　VoiceSyncの仕様概要

ため，使われる言語には制約がない．このエンジンでは話者照合および話者識別の機能が実現されている．

□ 運用事例 □

ここでは話者認識システムの現状を示す実運用システムの例として，テキスト独立型の話者認識エンジンを利用したテレホンバンキングシステムを紹介する．このシステムは1997年より実稼働している．このシステムは，電話による依頼のみで各種の銀行サービスを提供するもので，オペレータによる顧客との対応を前提とした有人テレホンバンキングサービスである．

図-1 は本システムの音声対話および話者認識機能の構成を示すものであり，オペレータ端末とテレホンバンキング用サーバによるサーバ・クライアント構成を採っている．基幹系のホストシステムとの連携はテレホンバンキング用サーバを介して行う．各オペレータ端末には，それぞれTelephony board（電話回線ボード）が搭載され，音声対話はそのボードを介してディジタル化され，さらにテレホンバンキングサーバを通して，録音・保存される．

その性格上，厳密さが要求される銀行サービスをこのようなテレホンサービスで実現するには，以下の2つの問題を従来の対面取引とは異なる技術によって解決することが必要であり，話者認識は2番目の問題をクリアするための要素技術である．

取引内容の証拠性

現行窓口取引においては顧客が記載した取引申請書等が書類として保存され，またATM端末取引では入力された取引履歴が電子的に保存される．しかし電話によるサービスでは，このような形での取引内容の保存ができないため，ディジタル化された音声データを保存することにより，取引の証明としている．さらに各オペレータによる決済は，検証端末による各取引ごとの検証を受けることによって事故の発生を防いでいる．

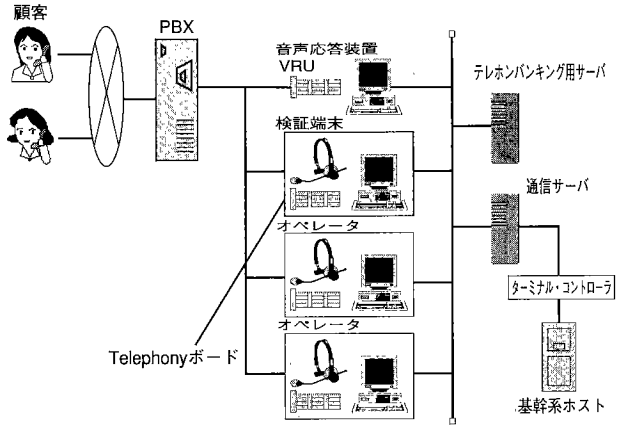

図-1

本人認証

現行の窓口またはATM端末では，通帳またはカードを用いて銀行口座所有者の本人性確認が行われているが，電話によるサービスを実現する場合には，この確認手段をとることができない．事前の本人確認手段は暗証番号のダイヤル入力であるが，非対面であるため成りすましの危険性が高いことが懸念される．このため，本サービスにおいては音声による本人認証（話者認識）技術を取り入れている．暗証番号のダイヤル入力と話者認識との併用により，従来以上のセキュリティを確保することが可能となった．

□ 実運用上の問題点とその対処法 □

ここでは上述の例を含めて話者認識システムの実用化において，筆者が遭遇した問題点およびその対応を述べる．対応策に関しては，理論面な手法よりもむしろ運用から得られた実践的アプローチを中心に述べる．

雑音した音声の取り込み

バイオメトリクス技術における認証精度を決定する基本的な要素には，当然のことながら認証アルゴリズムが挙げられる．しかしながら開発時の理論的認証精度と実運用での認証精度には，明らかに差異が生じる．このような差異が生じる原因としては，電話機の特性，通信路における歪み等に起因する入力音声のばらつきが考えられる．これに対する技術的な開発要件としては，雑音に対するロバストネスを上げることが重要である．

このようなセンサからの入力のばらつきはすべてのバイオメトリクスに共通な悩みであるが，音声等の行動特性に基づくバイオメトリクスでは行動自体の不安定さも大きな問題である．したがって品質的に安定した音声を取り込むためには，音声の入力を要求する際に自然な発声を促すようにソフト面において工夫された人間系インタフェースが最も重要なファクタとなる．

画面と人とのインタフェースであるGUI（Graphical User Interface）に対応する概念として，音声対話におけるマシンと人，オペレータと人とのインタフェースであるVUI（Voice User Interface）の概念を確立することが重要である．よく考えられたVUIによって自然な音声の発話を促すことによって，話者認識の入り口である音声入力の安定性が確保できると考える．

録音された音声による成りすまし問題

テキスト依存型の話者認識を使用する方式は，実は2つの本人認証方式が組み合わされたものとみることができる．すなわち，1つ目は単語・フレーズを登録者が設定できる点であり，これによって通常の暗証暗号・パスワードと同程度のセキュリティ強度が実現される．いわゆ

る「ボイスパスワード」という見方である．それに加えて，発声された音声データそのものに個人性がある点を利用することによって，セキュリティ強度がさらに高められているとみてよい．

しかし，電話回線のタッピング等の手段によって，本人が発声した単語・フレーズを盗聴・録音し，その録音音声を再使用する成りすまし手法が想定される．最近のMD等によるディジタル録音機の性能は，以前のテープによる録音と比較して機器特有のノイズというものが検出されにくくなってきたため，テープ特有のノイズの検出では対抗できないケースが想定される．

この問題点に関しては，先に述べられた通りテキスト指定型を用いて，限定されたキーワードの繰り返し使用を避けるという技術的な問題解決の方法もあるが，運用面においては，次のような方策で対抗することも可能である．

すなわち，行動特性に基づくバイオメトリクス，特に音声の場合には本人の繰り返し発話においても完全に同じデータではあり得ないという特性がある．このため入力された音声データと前に入力された音声データとの比較を行ない，ほぼ完全に一致した場合には録音物の疑いありとして「グレー（灰色：本人／他人の区別が明確にできていない状態を呼ぶ）」判定を行う．グレー判定に対しては別の本人認証手段を用いる特別処理に回すようにシステムを構築している．

さらにテキスト独立型との併用による二重の話者認識システムの構築が可能である．テレホンバンキングを含むテレホントレーディング分野においては，本人認証処理が終了した後に本来の取引にかかわる発話内容（取引指示）が続く．その自由な発話内容（テキスト独立な発話内容）をテキスト独立型の話者認識システムに処理させることにより，話者照合の実運用認証精度を確保するという考え方である．本アプローチでは，テキスト依存型×テキスト独立型という組合せによって，認証精度の確保と録音物への対応を可能としている．

経年変化

音声の経年変化，特にその個人性に関しては系統だった調査はきわめて困難である．発声には肺，腹筋，喉頭，咽頭，口腔，鼻腔の諸器官が関係しており，特に声帯の振動が声の基本周波数を決定している．このような身体的な器官の経年変化が，個人の発話の変化に影響を与えることは当然考えられるが，住環境の変化による言語的・発音的変化の影響も考えられる．また少年期から青年期に起こる「声変わり」も非連続ではあるが，経年変化と捉えられる．

音声による話者照合において，経年変化，特に長期間に渡るゆっくりした変化に関しては，筆者はまだ十分なデータを蓄えるには至っていない．このため，特定個人の経年変化への対応という点に関しては現在の技術はいまだ研究途上という段階であるが，実運用面においては次のような学習機能を取り入れた対応策をとっている．

個人を特定する基準となる特徴パラメータは，顧客のサービス加入時に採取・登録されるが，以下のような方法でその経年変化を学習して修正することができる．すなわち，認証時に正しく本人であると判定された場合に限って，入力された音声が安定で，かつ既登録の基準特徴パラメータとの差がある程度の分散の範囲であることを条件として，認証用に入力された音声データを加味して新たな特徴パラメータを作成する．

この方法によって，近似的に経年変化への対応を可能としている．しかし，テレホンバンキングのような短いサイクルで取引が行われる業態においてはこの対応策が有効であるが，長期間アクセスのない取引業態や取引者に対する対応についてはなお今後の課題である．

□ 話者認識の応用と今後の展開 □

21世紀には電子商取引による情報通信社会革命が本格的に展開するといわれている．この電子情報社会は自己責任能力のある個人を主たるプレーヤーとするものであり，健全な電子商取引システムの発展には個人認証技術が不可欠である．現在この個人認証は多くのアプローチで研究開発されているが，音声には取引の意思を伝える傍らで本人性を確認できる特長に加えて，既存のネットワークに乗せやすいという利点があり，個人認証に用いるバイオメトリクスとしては社会的受容性が高いと考えている．

通産省の先進的情報システム開発実証事業において，トータル的な実運用に十分耐え得る性能の話者認証システムの開発と1000人超規模の音声によるロバストネスの実証実験が予定されており，音声による本人認証の定着への起爆剤として期待されている．

他方，現在，WWWとディジタル放送の融合が急ピッチに進み，ブロードバンドチャネルが全世界に張り巡らされ，大量の情報が流通する時代になってきた．データベースはすべてマルチメディア化される方向にあり，その中から有効な情報を検索する技術が早急に求められる．いわば，情報氾濫に対する治水技術である．この問題に対し，音声認識と話者認識を組み合わせたメタサーチエンジンのアイディアは1つの解決方向を示すものとなろう．

現在は，この20年間で最も本格的な音声技術ブームにみえる．しかしながら音声を用いたキラーアプリケーションの開発がなされなければ，今回も単なる一過性の現象に過ぎなくなるのではないかというのが率直な懸念である．

（平成11年8月4日受付）