4. 音声による本人認証

第1部 音声による本人認証のしくみと技術動向

東京工業大学大学院情報理工学研究科計算工学専攻
古井 貞熙

音声を用いて，本人かどうかを認証できるようにしたいという期待が高まっている．その背景には，最近の音声ワープロにみられるような音声処理技術の発展と，音声なら通常の電話やパソコンに付属のマイクを入力手段として用いることができるので，容易に通信関連のシステムが実現できるというメリットがある．しかし音声には，同じ人でも変化する，雑音や伝送系の影響を受けて変化するなどの難しさがあり，声の変動への対処法などが活発に研究されている．

□ 話者認識の歴史 □

音声の個人差を用いて，誰の声であるかを自動的に判定することを話者認識という．これによって，個人を認証しセキュリティを守るのに音声を用いることができる．

音声を個人の特定に用いようという試みは，犯罪捜査から始まった．音声が初めて犯人の割り出しの手掛かりとして用いられたのは，1660年の英国チャールズI世の死に関する裁判のときであるといわれている[1]．その後，時代を経て，リンドバーグ氏の子息の誘拐事件をきっかけとして，1937年に初めて音声の個人性に科学的なメスが加えられた．そして，1945年に米国のベル研究所のポッターによって，いわゆる「声紋」(科学的には「サウンドスペクトログラム」) を自動的に描く技術が発明され，1962年に同研究所のカースタが初めてこれによる話者認識の可能性を発表した．これは，人が声紋を目で見て判断するもので，客観性に乏しい問題があった．その後，コンピュータなどの発展により，自動的に話者を認識する研究が活発に行われるようになり，今日に至っている．

最近では，情報化社会の中で，インターネットや電話網を用いたバンキングサービス，買い物サービス，情報提供サービス，コンピュータのリモートアクセスなどで，音声を用いて本人か否かを確認したいというニーズに答えるための種々の研究が進められている．テレフォンカードやクレジットカード（現金引き出し機）に，話者認識機能を付け加える実験もすでに行われている[2]．

また，本人認証からは少し離れるが，マルチメディア情報検索や，会議の議事録などの作成を目的として，複数の人が発声した一連の音から，各話者とその発声区間を自動的に検出する研究も行われている．

□ 話者認識のしくみと特徴 □

話者認識の長所と短所

カードや暗証番号は，他人に盗まれたり落としたりするおそれがあるが，音声のようなバイオメトリクスにはそのような心配がない．また音声は電話によって伝えることができるので，音声を用いて本人確認ができれば，他の情報を用いる場合に比べて，電話やマイクロホン以外に特殊な装置を，ユーザ側に用意する必要がないという大きなメリットがある．ネットワークに関連したセキュリティを守る手段として，最も便利であるといえる．一方欠点としては，指紋などと違って本人の音声でも変化するという点がある[2]．音声の変化要因には，付加雑音，電話機，伝送系の歪みなどもある．このために，よく似た他人の声を本人の声と認識したり，本人の声なのに他人の声と認識する誤りを完全に回避するのは難しい．

話者認識の分類

話者認識は，次の3種類に分類できる[3), 4)]．

(a) テキスト依存（限定）型：「ひらけごま」のように，用いる音声の発声内容（キーワード）をあらかじめ決めておく

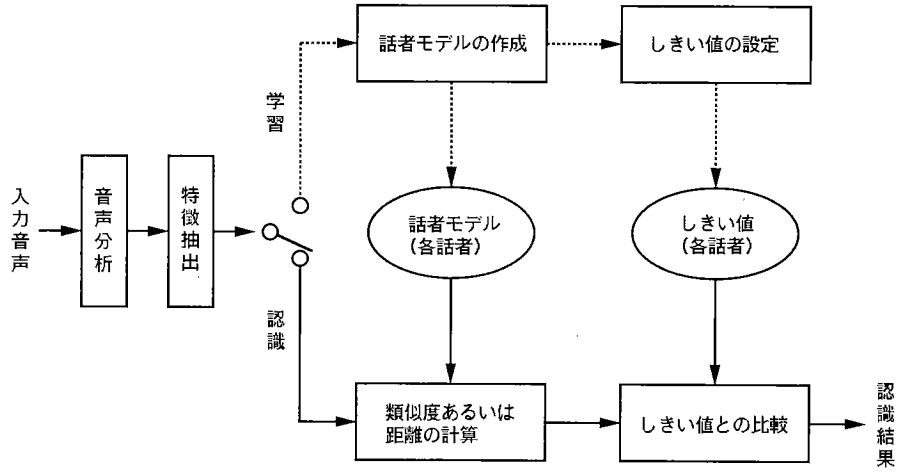

図-1　話者認識システムの基本的構成

(b) テキスト独立型：どんな言葉を発声してもよい
(c) テキスト指定型：装置を使うたびに新しいキーワードを装置側から指定する

テキスト依存型と独立型を比較すると，前者の方が，キーワードの違い（キーワードが盗まれないことが前提）に加えて，その言葉特有の音声の個人差も利用することができるので，一般に後者よりも短い音声で比較的高い性能を得ることができる．たとえば，テキスト依存型で数秒程度の音声と同じ精度を得るために，独立型では10秒ないし数10秒の音声が必要である．実用的には，話者認識の応用の多くにおいては，キーワードをそれぞれの人について固定することが可能であるが，応用によっては，必ずしも同じ言葉どうしを比較することができない場合もある．この場合には，テキスト独立型の話者認識技術が必要になる．

テキスト依存型と独立型にはどちらにも，本人が発声した音声を他人がひそかにテープレコーダなどで録音して持ってきて，装置の前で再生すれば装置が簡単にだませてしまうという致命的な問題がある．テキスト指定型は，この問題を解決するために開発された方法である．この方法では，装置がディスプレイや音声合成によって，その場で新たなキーワードを提示し，本人が正しくそのキーワードを発声したか否かを判定する．いわば，音声認識と話者認識の両方を同時に行う．この方法では，どんな言葉を発声しなければいけないかが事前に分からないので，録音テープを用意することができない．このようにして，従来の方法が持っていた問題を回避することができる．テキスト指定型の簡易的な方法として，発声内容は数字の組合せに限定するが，毎回新しい数字の並びを装置側から指定して，発声させるという方法もある．

話者認識システムの基本的構成

話者認識システムの基本的構成は，**図-1** に示す通りである[2]．音声波は，まず10ミリ秒程度の細かい時間ごとにスペクトルに変換される（音声分析）．実際にスペクトルを表現する方法としては，ケプストラム（cepstrum）パラメータ[1), 3)]が広く用いられている．ケプストラムには，音声のスペクトルを滑らかな形で表現する優れた性質があるため，現在の音声認識システムで広く用いられている．そのパラメータの時系列を各話者ごとにそのまま登録するか，話者の特徴を表現するパラメータに変換し（特徴抽出），それに基づいて各話者のモデルを作成して登録する（学習）．モデルを作成する方法としては，ベクトル量子化，HMM（隠れマルコフモデル；hidden Markov model）などの技術を用いる[3]．HMMも，最近の音声認識で広く用いられている方法で，統計理論に基づいている[5]．音声は上述のように，同じ人でも変化するので，あらかじめ各話者の音声の変動の幅を調べて，その人の典型的なモデルを作成すると同時に，本人の音声と判定する許容限界のしきい値を決めておく．

認識する際には，学習時と同じ方法で音声分析と特徴抽出を行い，蓄積されている各話者のモデルと比較を行う．その類似の度合いが，あらかじめ設定されているしきい値よりも大きければ本人の音声と判定（受理）し，そうでなければ他人の音声として判定（拒否，棄却）する．

話者認識の誤り（誤認識）には，本人の音声を棄却す

る誤り（タイプ１の誤り，本人拒否率，本人棄却率ともいう）と，他人（詐称者という）の音声を受理する誤り（タイプ２の誤り，他人受入率，詐称者受理率ともいう）の２種類がある．しきい値をどこに設定するかによって，両者の誤りにはトレードオフの関係がある．このため実際のしきい値は，両者の誤りの影響の相対的重要性に応じて設定する必要がある．研究論文では，通常，２種類の誤り率が等しくなるしきい値に事後的に設定して，そのときの誤り率（等誤り率）で性能を評価している．実際の認識システムではこのようなことはできないので，あらかじめいかにして最適なしきい値を推定しておくかが，重要な課題である．

❑ 話者認識の技術動向 ❑

ここでは，比較的最近の研究例を紹介する．主として，音声の変化に対応して高い認識性能を実現するための研究が，活発に行われている．

HMMかDTWか

テキスト依存型話者認識の古典的方法は，各登録話者が発声したキーワードあるいはキーフレーズから抽出したパラメータ系列をそのまま蓄えておいて，入力音声のパラメータ系列との類似度を調べることである．この際に音声の時間的伸縮を正規化するために，動的計画法を用いた類似度計算（DTW; dynamic time warping）[1), 3)]を行う．最近では，音声の変動を統計的モデルで表すHMMが多く用いられるようになったが，このためには，各話者についてある程度の数の発声が必要であり，しかも，声の変動をモデル化するためには，時期を変えた発声が必要になる．これは，発声者にとって大きな負担であり，実用システムを立ち上げる上での障害となる．このため，むしろ１回あるいは少数の発声の音声をそのまま蓄えるDTW法の方が有利なこともある．

羊，山羊，小羊，狼

話者認識の精度には，話者による大きな偏りがあり，誤認識のきわめて大きい（認識の難しい）一部の話者によって全体の認識性能が決まってしまうことが，よく知られている．このような現象は，一般に“Sheep and goats現象”と呼ばれる[3)]．話者による誤認識率の違いが10倍にもなることがある．実際のシステムでは，全話者の平均値だけでなく，このような特に難しい話者の誤認識がどの程度に抑えられるかが重要である．文献6）では，特殊な話者を次のように呼び，これらの話者を統計的に検出する方法について検討している．

- Sheep（羊）：誤認識の少ない大多数の話者（やさしい話者）
- Goats（山羊）：誤認識のきわめて大きい一部の話者（難しい話者）
- Lambs（小羊）：他人が真似しやすい声の（似た声が多い）話者
- Wolves（狼）：他人の声の真似が得意な（他人の声として受理されやすい）話者

このうちSheep（羊）は最も問題の少ない話者で，通常，話者集団の大部分を占める．実験に用いた話者数が少なくて，たまたまこのような話者だけから構成されていると，思いのほか良い認識率が得られることになる．ところが実験の規模を拡大していくと，Goats（山羊）が含まれるようになり，話者数としてはわずかな割合であっても，平均認識率を大きく下げることになる．このような話者の声でも，同じ日の声の比較では大きく変動しなかったり，静かな理想的な環境では変動が顕著に現れなかったりするので，実用を目指した実験では注意をする必要がある．さらに，このような話者の本人の音声を拒否しないように，しきい値を緩く設定すると，他人の音声を受け入れやすくなってしまうので，何らかの特別の防御策が必要になることもある．

Lambs（小羊）は，上で述べた理由でGoats（山羊）に対するしきい値を緩めることによって生ずるのが普通なので，Goats（山羊）と同じ話者になることが多い．Wolves（狼）に関しては，よく分かっていない．少なくとも，プロの声帯模写者がコンピュータによる話者認識を容易に破れるということはないことが，我々の実験によって確認されている．声帯模写者でも声の質をそっくりに真似ることは難しいので，主として話し方のくせを真似しており，声の質に関する特徴を用いている話者認識システムには影響が少ないためである．

特徴パラメータと尤度の正規化

日常的な声の変化に加えて，風邪を引いて声が変わってしまったらどうするか，騒音が大きいときにどうするかなど，声の変動の正規化は重要な課題である[4)]．適切な正規化なくしては，話者を判定するしきい値を事前に設

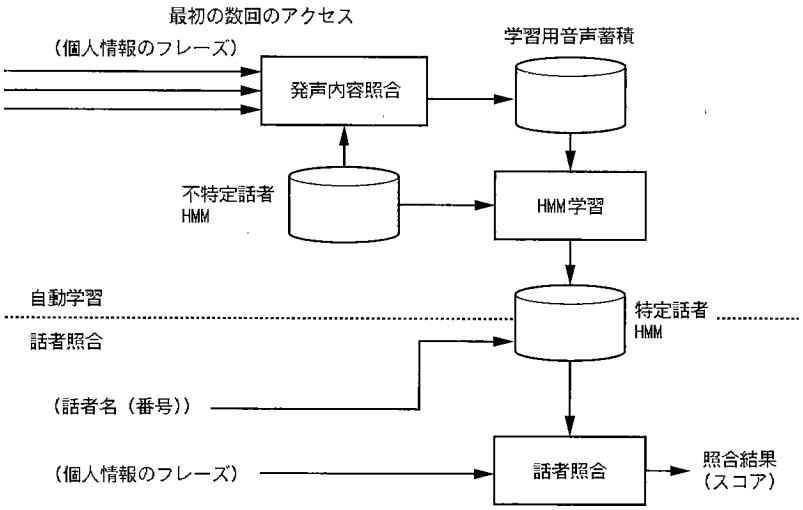

図-2　発声内容照合を含む話者認識のブロック図

定することができない．正規化の方法には，特徴パラメータあるいはモデルに対して適用する方法と，尤度（あるいは距離）に対して適用する方法がある[3]．

前者の方法として，比較的簡単であるが有効な方法に，文章中の各次数のケプストラムの平均値をゼロに正規化する，CMS（cepstral mean subtraction）あるいはCMN（cepstral mean normalization）法がある．HMMを用いた話者認識において，騒音が大きいときに高い認識性能を達成する方法として，HMM合成法が有効であることが実験的に確認されている．

後者の方法としては，尤度比や事後確率を用いて尤度を正規化する方法が，きわめて有効であることが確かめられている．

しきい値とモデルの更新法

各話者の少数のデータに基づいて，判定のしきい値をどのように設定するか，データの増加と声の変化に従って，モデルや判定のしきい値を定期的かつ自動的に更新するにはどうすればよいかという問題も重要である[3), 4)]．文献7）では，最適なしきい値を，過去のデータに基づいて推定する種々の方法が，実験的に比較されているが，まだ決定的な方法はない．

発声内容照合法

システム立ち上げ時に，各ユーザが時期を変えて複数回発声した音声が収録できるまで，システムが使えないという問題を解消し，かつ高い性能を得る方法として，発声内容照合（VIV; verbal information verification）法が提案されている[8]．この方法では，システムの使い始めでは，声の個人性による認識は行わず，本人しか知らない情報（たとえば本人の母親の旧姓など）を装置側から質問して，その答えを音声認識し，その答えが（十分に）正しければ本人と判定する．この音声認識では，不特定話者用の音素HMMを用いる．本人と判定されたら，その声を収録しておき，4～5回の音声が収録されたら，それを用いて，本人の音声の特徴を示すHMMを作成する．以後はVIVとHMMを用いた話者認識の両方，あるいは後者のみに移行する．VIVから話者認識（話者照合）へ移行する形態のシステムのブロック図を，**図-2**に示す．

参考文献

1）古井貞熙: ディジタル音声処理, 東海大学出版会（1985）.
2）古井貞熙:「音声」の識別でセキュリティを守る, エレクトロニクス, pp.38-40（1998）.
3）古井貞熙: 音声情報処理, 森北出版, 東京（1998）.
4）松井知子, 古井貞熙: 話者認識研究の現状と展望, テレビジョン学会技術報告, Vol.20, No.41, pp.19-24（1996）.
5）中川聖一: 確率モデルによる音声認識, 電子情報通信学会, 東京（1988）.
6）Doddington, G. et al.: Sheep, Goats, Lambs and Wolves－A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation, Proc. ICSLP 98, We4B4（1998）.
7）Pierrot, J.-B. et al.: A Comparison of A Priori Threshold Setting Procedures for Speaker Verification in the CAVE Project, Proc. ICASSP 98, pp.125-128（1998）.
8）Li, Q. and Juang, B. H.: Speaker Verification Using Verbal Information Verification for Automatic Enrollment, Proc. ICASSP 98, pp.133-136（1998）.

（平成11年6月17日受付）