T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	音声による個人認識
Title(English)	
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	SUT BULLETIN, Vol. , No. 7, pp. 26-31
Citation(English)	, Vol. , No. 7, pp. 26-31
発行日 / Pub. date	2000,

Powered by T2R2 (Science Tokyo Research Repository)

音声による個人認識

古井　貞熙

音声を用いて本人かどうかを認証する話者認識技術への期待が高まっている。音声には，バイオメトリックス共通のメリットのほかに，電話やマイクロフォンを入力手段に使えるという便利さもある。最近では，HMMによる方法，テキスト指定型，発声内容照合法などが研究されている。同じ人の音声でも発声のたびに変化するので，高い認識精度を得るためには，変動に対する種々の正規化法を適用することが必要である。

話者認識とは

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[1,2]。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。バイオメトリクスによる他の個人認証技術と同様に，カードや暗証番号と違って他人に盗まれたり落としたりするおそれがない長所がある。さらに音声には，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，他のバイオメトリクスによる方法と比べて，新たな機器やインフラを整備しなくても，容易に通信ネットワーク関連のシステムが実現できるというメリットもある。

しかし音声には，指紋などと違って，同じ人でも変化するという問題がある。音声の変動要因には，付加雑音，電話機，伝送系の歪みなどもある。このために，人も経験するように，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。

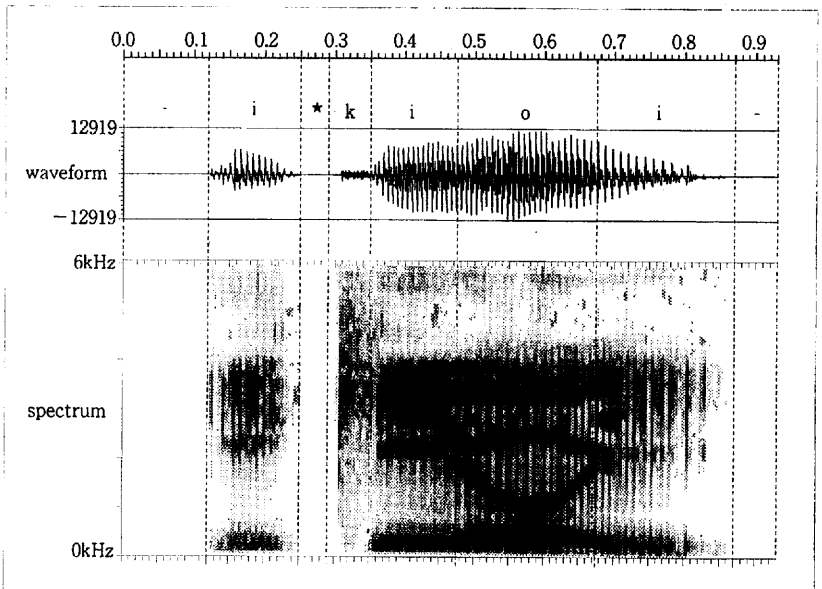

図1　男性が「いきおい」と発声した音声の波形とスペクトログラム

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。音声の個人性に初めて科学的なメスが加えられたのは，リンドバーグ氏の子息の誘拐事件がきっかけで，1937年のことであった。そして1945年に米国のベル研究所のボッターによって，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。図1に，男性が「いきおい」と発声した音声の波形とスペクトログラムの例を示す。

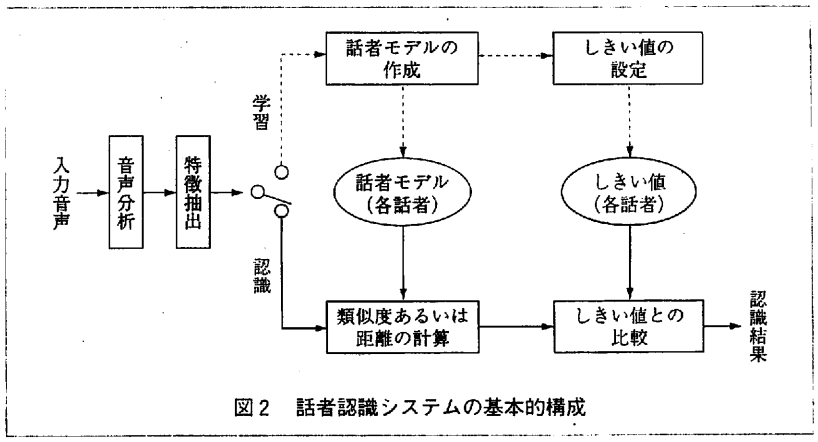

図2　話者認識システムの基本的構成

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応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。

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

話者認識の種類

話者認識は，次の3種類に分類できる。

(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

(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発声内容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，人によるキーワードの違いに加えて，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。例えば，テキスト依存型で数秒程度の音声で得られる精度を，テキスト独立型で得るためには，10秒ないし数十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，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(なりすまし)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(後述)。この方法では，装置がディスプレイや音声合成によって，その場で新たな文あるいは単語を提示し，本人が正しくそれ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なら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。こうして，従来の方法がもっていた問題を回避することができる。

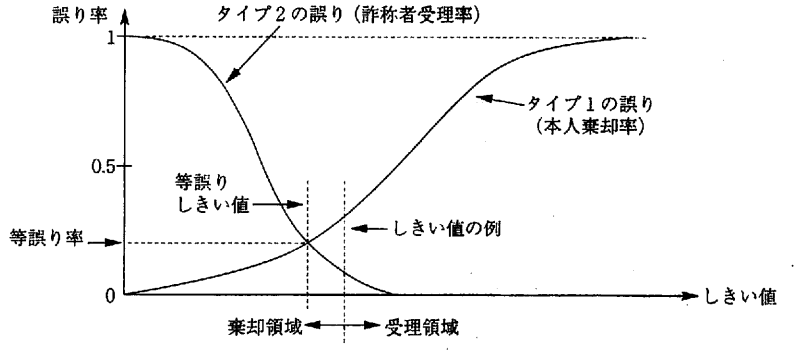

図3　話者認識のしきい値（許容範囲）と誤り率の関係

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。図3に示すように，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（等誤りしきい値）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，あらかじめ如何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2に示すとおりである。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）パラメータが広く用いられている[1]。ケプストラムには，音声のスペクトルを安定な形で表現する優れた性質があるため，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（ベクトル）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），それを用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[1]。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（次章参照）。

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

HMM（隠れマルコフモデル）

音声認識および話者認識で使われる典型的なHMMの構成を図4に示す。HMMは，確率的な状態遷移と確率的な出力信号（音声の場合はケプストラムなどからなる特徴ベクトル）を備えたオートマトンである。観測された出力信号系列から，その出力信号系列を生成した状態遷移系列を一意的に復元することができないため（状態遷移系列がモデル内部に隠れていて見えないため），「隠れ」マルコフモデルと呼ばれる。

図4-(a)はエルゴード（Ergodic）型と呼ばれ，どの状態も開始と終了状態になることができ，すべての状態間の遷移が可能な構成になっている。テキスト独立型話者認識のように，任意の発声内容の音声を表現するのに用いることができる。このとき各状態は，母音，子音など特徴的な音素のグ

ループに自動的かつ確率的に対応付けられる。もう一方の(b)は，Left-to-right型と呼ばれ，音声認識で各音素や単語を表すのに広く用いられている。これらの音声ではスペクトルの時間的経過が重要で，状態の遷移が時間的に逆転することはないので，(b)の構成が適している。

HMMを特徴付ける状態遷移確率と特徴ベクトルの出現確率は，大量の音声を用いて自動的に学習することができ，音声の特徴ベクトル系列がある特定のHMMから出現する確率(尤度)も，容易に計算することができる。

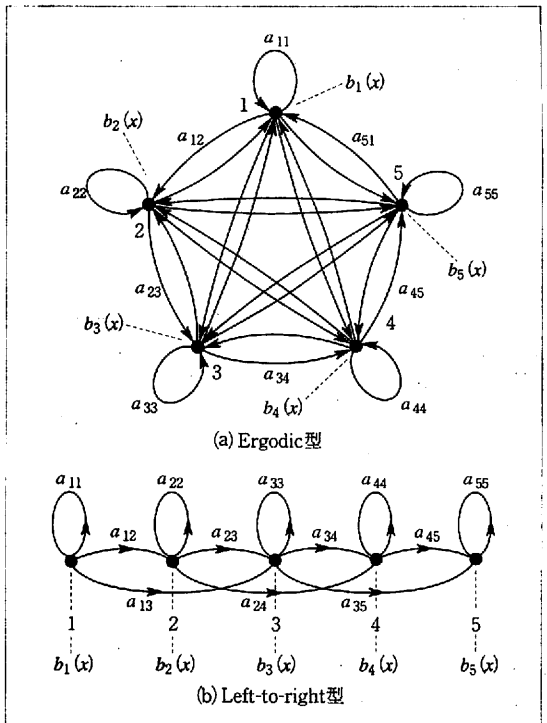

図4　音声認識および話者認識で使われる典型的なHMMの構成
(a_{ij}：遷移確率，$b_{ij}(x)$：特徴ベクトルxの出現確率)

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，音声の変化に対処しながら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。

テキスト指定型話者認識

テキスト指定型話者認識システムの構成を図5に示す。登録されている各話者が，装置が指定した任意の新しい文(あるいは単語)を発声したときの，音声のスペクトル系列をモデル化するために，あらかじめ各話者の各音素のHMMを作っておく。そのために各話者にあらゆる音素を含む大量の文を発声してもらうことは非現実的なので，不特定話者の音素モデルを種として，限られた量の学習用音声とそのテキストを用いて，その話者の声に適応した音素HMMを作成する。

認識の際には，装置から指定するテキストに応じて，認識すべき話者の音素モデルを接続して文モデルを作り，音声から抽出した特徴ベクトル系列の出現確率(尤度)を計算する。この尤度をあらかじめ決めてあるしきい値と比較して，認識の判定を行う。

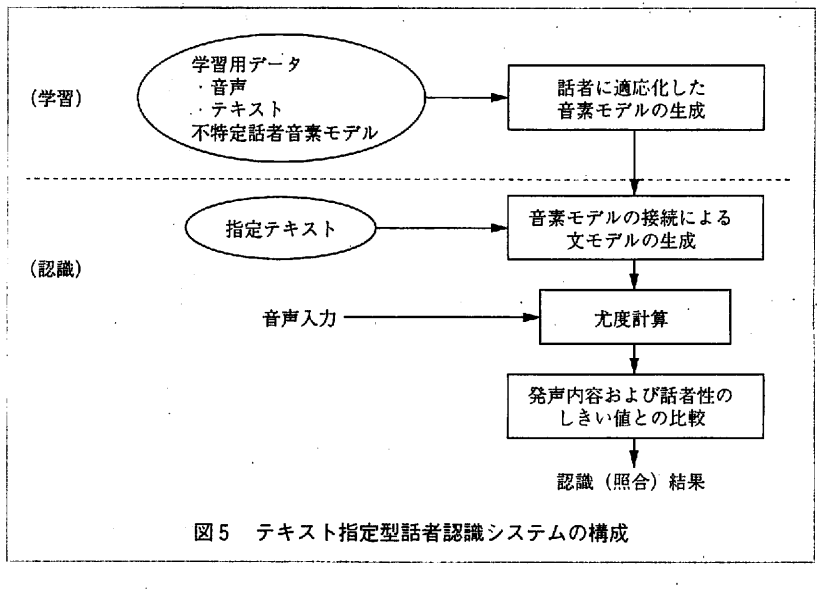

図5　テキスト指定型話者認識システムの構成

テキスト指定型でも，提示された文を，本人そっくりの声で直ちに合成できる音声合成装置ができれば，これを用いて装置をだますことは可能であるから万全とは言えないが，それにはまだしばらく時間がかかるであろう。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み合わせに限定するが，毎回違う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。

羊，山羊，子羊，狼

話者認識の精度には，話者による大きな片寄りがあり，誤認識の極めて大きい(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。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。話者による片寄りは，次のように分類できる[2]。

- Sheep(羊)：誤認識の少ない大多数の話者(やさしい話者)
- Goats(山羊)：誤認識の極めて大きい一部の話者(難しい話者)
- Lambs(子羊)：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，話者数としては僅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(子羊)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。

声の変動の正規化とモデルおよびしきい値の更新

日常的な声の変化に加えて，風邪をひ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。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。

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。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研究されている。

発声内容照合法

声が時期的に変動する性質があるため，各話者の典型的なモデルを作るには，複数の時期にわたった音声を収録して用いる必要がある。しかしこ

れでは，システム立ち上げ時の各ユーザの負担が大きくなってしまう。この問題を解消し，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。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 (例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6に示す。

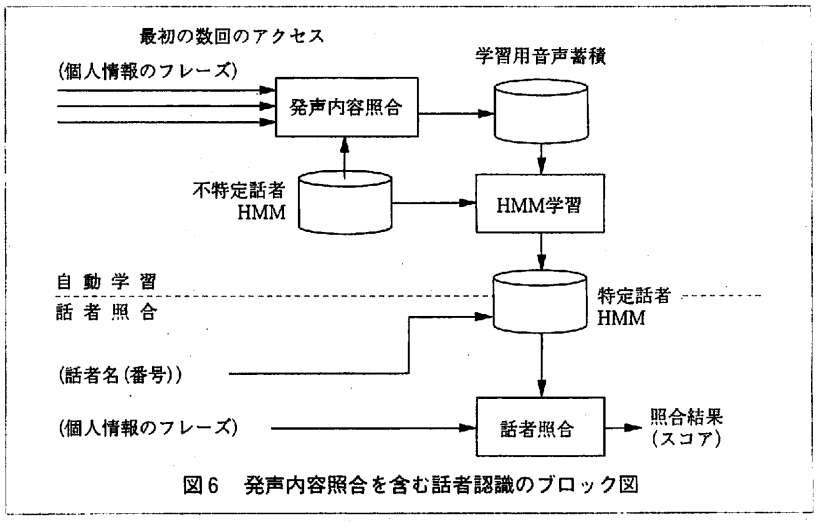

図6　発声内容照合を含む話者認識のブロック図

むすび

音声を用いた本人確認への期待の高まりを受けて，種々の研究が進められており，近年の音声認識技術の進歩と歩調を合わせて，新しい有望な技術が提案されている。しかし本人の音声でも発声のたびに変化する問題への対処法は，まだ十分とは言えない。日常的な変化に加えて，風邪をひ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という問題もある。近年の音声認識の進歩と同様に，話者認識技術のこれからの飛躍的進歩のためには，大量の音声データベースを構築し，コンピュータの統計的処理による認識アルゴリズムを確立することが必要であろう。

参考文献

1) 古井貞熙：『音声情報処理』森北出版，1998年
2) 古井貞熙：「音声による本人認証」『情報処理』，1088～1091，1999年

古井　貞熙(ふるい・さだおき)
東京工業大学　大学院情報理工学研究科計算工学専攻　教授
[専門など] 話者認識，音声認識，音声知覚，ヒューマンインタフェースなど
[主な著作論文]『ディジタル音声処理』(東海大学出版会)，『音響・音声工学』(近代科学社)，『音声情報処理』(森北出版)，「Digital Speech Processing, Synthesis and Recognition」(Marcel Dekker)，「テキスト指定型話者認識」『電子情報通信学会論文誌』1996年