論文 / 著書情報
Article / Book Information

論題(和文)	声の個人性の話
Title	
著者(和文)	古井 貞熙
Author	SADAOKI FURUI
出典(和文)	日本音響学会誌51巻11号, Vol. 51, No. 11, pp. 876-881
Journal/Book name	, Vol. 51, No. 11, pp. 876-881
発行日 / Issue date	1995, 11

日本音響学会誌51巻11号（1995），pp. 876-881

小特集—声質：音声言語の多様性に迫る—

声の個人性の話*

古井貞熙（NTTヒューマンインタフェース研究所/東京工業大学）**

1. まえがき

人の声に個人差がなかったら，つまり，みんなが活字のように同じ声だったら，人生はどんなにつまらないものになってしまうことでしょう。声に個人差があるおかげで，好きな人の声に胸をときめかせたり，声からその人の性格を推し量ったりすることもできるのです。人の声には，伝えようとする言葉の内容（言語的情報）だけでなく，話している人の感情（情緒性情報）など，様々な情報が含まれています。その中で，誰が話しているかという情報（個人性情報）は，実用的にも極めて大切で，私達の日常生活の中で重要な役割を果たしています[1]。テレビやラジオの討論会で，画面を見ていなくても，誰が発言しているかが分かったり，電話で相手が名乗らなくても，親しい相手なら，誰からかかってきたのかがすぐ分かるのも，声が人によって異なるからです。同時に複数の人が発言している中から，ある特定の人の発言に注目できるのも，声の個人差のおかげです。

それでは，声にはどうして個人差があるのでしょう。人は，どうやって声の個人差を聞き取っているのでしょう。これらは音声科学のテーマですが，意外によく分かっていません。声の個人差に関連する音声工学には，話者認識，話者適応などがあります。話者認識は，個人性情報を自動的に抽出して，誰の声であるかをコンピュータなどで自動的に判定することです。話者適応とは，音声認識に用いられる種々のモデルやアルゴリズムを，新しい発声者に合わせて，自動的に変えることです。これは，音声認識の精度を高めるために，欠くことのできない技術です。

これ以外の，個人性に関連した音声工学には，合成音声の声質を，ある特定の人の声にする技術があります。自然な声には，必ず個人差があり，それによってある種の人格を伴うものですから，自然な合成音声を実現するためには，何等かの個人性を持たせることが不可欠です。合成音声に個人性を付与する話については，本特集の中で，別に解説がされることになっていますので，ここではこれ以上は触れないことにします。

2. 声の個人差を生ずるしくみ

声の中には，声帯が振動して音源となる有声音と，口の中や唇などの声帯以外の狭いところを空気が通り過ぎるときの空気の流れの乱れが音源となる無声音があります[2]。これらの音には，声道と呼ばれる声帯から唇までの部分の形が，舌や顎の動きによって変化することによって，アやクのような特有の音色がつけられて，音声として外へ出てきます。声の高さは，声帯の振動周期によって決定され，声の音色には，声帯波，空気の乱流を生ずる狭めの形状，声道全体の形状など，いろいろな要因が寄与しています。

個人差を生ずるのは，音声の生成に関与するこれらの発声器官の形や動きが，人によって異なることによりますが，それにはそもそも先天的なものと，習慣として身に付いたものとがあります。この両者を明確に区別することはできませんが，親子や兄弟姉妹の声がよく似ていることから，先天的なものがかなりあると推測されます。方言や話し方のくせは主として後天的なもので，家族の声が似ている要因には，一緒に生活しているために話し方のくせが似てくるということもあると考えられます。声帯模写の人は，各個人の先天的な声の質を真似るのは極めて難しいため，（「声帯を

* Key issues in voice individuality.

** Sadaoki Furui (NTT Human Interface Laboratories Musashino, 180/Tokyo Institute of Technology, Tokyo, 152)

模写」するのではなく）主として後天的な話し方のくせを真似ていると考えられます。

声帯は柔らかいものですから，地声で乱暴に歌ったりしていると，縁がガサガサに傷んだりすることがあります。ひどい声を聞くと，この人の声帯はそうとう荒れているなと感じます。自然に直る場合もありますが，手術が必要になる場合もあります。

声は訓練によってある程度変えることができます。一般にアナウンサの音声は3～4 kHz付近のスペクトルのレベルが普通の人に比べて高く，これがアナウンサ音声の響きを良くしていると言われています[3]。「ボイストレーナ」という，歌手の声の出し方を教える人がいますが，その役目は，歌手が本来持っている声の素材をどのように磨くかということなのだそうです[4]。その人が持っている一番いいところを引き出し磨くことで，魅力ある声を生み出すことができるのだそうです。声の質がいい，悪い，ということよりも，人間性のある，温かみのある，その人の心を表現する声，そういう意味で持ち声のいいところを引き出すという点が，大事なポイントなのです。

3. 声の個人差を聴き取るしくみ

人は声の音色，高さ，大きさなど，様々な情報から個人性を知覚しています。分析合成音を用いた実験によれば，種々の物理的特徴要素の聴覚的効果は，スペクトル包絡特性（音色），基本周波数（高さ），発話時の時間特性（テンポ）の順に大きいことが分かっています[5]。スペクトル包絡情報として最も基本的で大切なのは，ホルマントと呼ばれるスペクトルのピークです。中でも特に重要な，第1，第2ホルマントを多数の人の母音について調べてプロットすると，基本的には5母音の相対的位置関係を維持したまま対数周波数で平行移動する傾向がありますが，一人一人について見ると現象はかなり複雑です[1]。個人を特徴づけるスペクトル情報に関しては，かなり広い周波数範囲に分散していること，その中では，2.5～3.5 kHzくらいの周波数範囲のスペクトルが，特に重要な役割を果たしていることなどが分かっています[6]。

人は，相手がどんな言葉を話していても，言葉によらない共通した声の個人性があると感じていると思われます。多数の人の音節を用いた受聴実験の結果，声の個人差のために，急に人が代わると音節の正聴率が下がりますが，音節の種類に関わらず，同じ人の5音節程度を聴けば，その声に慣れて聴きやすくなることが分かっています[7]。音節の種類に普遍的な個人性をつかんでいるというわけです。しかしこれが，音声波形やスペクトルのどのような特徴によっているのかは，ほとんど分かっていません。これが明らかになれば，後述するコンピュータによる話者認識に大いに寄与すると期待されます。

4. コンピュータによる話者認識

4.1 話者認識への期待

今後の情報化社会の発展の中で，音声によって誰であるかを自動的に判定する話者認識[8),9)]への期待が高まっています。これが実現できれば，車や建物の入口の音声キーが可能となるほか，特定の会員や利用者を対象とした電話によるバンキング，買い物，情報案内サービス，音声メールやコンピュータのリモートアクセスなどにおいて，音声による本人確認ができるようになり，極めて便利になると期待されます。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りますが，音声にはそのような心配がありません。ただし，本人の声が変化しても正しく動作し，他人が真似た声では誤動作しないようにすることが必要です。

話者認識の一般的方法としては，図-1に示すように，まず登録すべき各話者の声の特徴を学習します。具体的には，各話者に幾つかの単語や文章を発声してもらい，その音声波形から話者の特徴を抽出します。通常は10ミリ秒程度の細かい時間ごとに，ケプストラムとその時系列の多項式展開係数を計算します。ケプストラムは，対数スペ

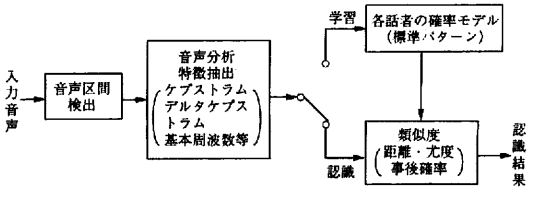

図-1 話者認識系の基本的構成

クトルをフーリエ逆変換したもので，比較的少数のパラメータで，スペクトル包絡を安定に表現できる特徴があります。ケプストラム空間における話者間距離は，音声を受聴したときの話者間の心理的距離とも比較的よく対応することが分かっています[6]。ケプストラムの時系列の多項式展開係数は，デルタケプストラムと呼ばれており，これによって，聴覚的に重要な音声スペクトルの時間的変化情報を表現することができます[10]。

抽出したパラメータ時系列を，各話者ごとに，標準パターンとして蓄えるか，スペクトルの変動を考慮した確率的なモデルとして蓄えます。確率的なモデルとしては，音声認識の場合と同様に，HMM (Hidden Markov Model；隠れマルコフモデル)[2] が広く用いられています。HMM は，音声のスペクトルの時系列が生成されるモデルを確率状態遷移マルコフ過程（マルコフモデル）で表したものです。話者認識をする場合には，未知音声について，学習のときと同じ方法で話者の特徴パラメータを抽出します。そのパラメータ時系列と，登録されている各話者の標準パターンあるいはモデルとの類似度を調べ，その度合によって話者の判定を行います。

話者認識の難しさは，同じ人の同じ言葉でも，発声するときによってその物理的特徴が変化し，認識精度が低下してしまうところにあります。これが指紋と大きく異なる点です。同じ日に繰り返し発声した音声の変動は極めてわずかですが，数か月たつとかなり大きな変化を生じます[8),9)]。風邪をひいたり，電話を通したりすれば，変化は一層大きくなります。私共の実験結果によれば，プロの声帯模写の人が声を真似たときの問題よりも，本人の声が時期と共に無意識の内に変化する問題の方が，はるかに大きいことが分かっています。このため，変化を自動的にキャンセルする方法，各個人の標準パターンやモデルを逐次更新する方法などが研究されていますが，まだ解決されたとはいえません。

4.2 話者識別と話者照合

話者認識は，話者識別と話者照合に分けることができます。音声が，あらかじめ登録されている多数の人の内の，誰の声であるかを判定するのが話者識別です。自分が誰であるかを音声，カード，登録番号などで名乗り，音声が，名乗った本人の声であるか否かを判定するのが話者照合です。音声を本人確認に用いる応用のほとんどは，話者照合に該当します。話者識別の場合は，多数の登録話者の中から，未知音声と最も類似度の高い話者を選びます。話者照合の場合は，類似度が一定のしきい値よりも大きければ名乗った本人の音声であるとして受理し，そうでない場合は他人の音声と判定して棄却します。

話者識別の性能は，登録話者の数が大きくなると，それだけ選択肢が増えるので低下しますが，話者照合の性能は，常に本人か他人かの二者択一の判定を行いますので，その数には依存しません。話者識別の計算量は，登録話者の数に比例して増えますが，話者照合の計算は基本的に，未知音声と名乗った話者の標準パターン（モデル）との類似度の計算のみですので，登録話者の数が増えても変わりません。

4.3 テキスト依存型とテキスト独立型話者認識

話者認識の従来の方法は更に，あらかじめ決まっている言葉（キーワード）を発声するテキスト（発声内容）依存（限定）型と，任意の言葉を発声してよいテキスト独立型に分類できます。前者の場合は，ア，イなどの個々の音素に依存した個人性情報を直接的に用いることができますので，一般にテキスト独立型の方法よりも高い認識性能を得ることができます。言い換えると，同じ認識性能を得るためには，テキスト依存形の方が，短い音声ですみます。発声する言葉をあらかじめ決めておいてもよいから，なるべく簡単な方法で，認識精度を高くしてほしいというニーズには，この方法によって対処するのが適当です。

しかし応用によっては，決まった言葉を用いるのが難しい場合もあります。また，人は発声内容にかかわりなく，話者を認識することができます。このため，テキスト独立型やそれを基本とする方法も，最近活発に研究されています。テキスト独立型の代表的な方法には，長時間平均スペクトルによる方法[8]，ベクトル量子化による方法[11),12)]，エルゴディックな HMM を用いる方法[13] などがあります。しかし，いずれの方法も，発声内容にかかわりなく共通に含まれている個人性特徴に着目

していると思われる人の個人性知覚の仕組みと結び付けるのには無理があるように思われます。人と同じような機能を実現するには，音声知覚と工学的な話者認識アルゴリズムの両方からの一層の研究が必要です。

4.4 テキスト指定型話者認識

テキスト依存型，独立型に共通して，本人のキーワードの音声を録音して装置の前で再生すれば，本人になりすまして装置を容易にだますことができるという問題があります。この欠陥は，アメリカ映画「スニーカーズ」の中でも，装置をだます方法として利用されています。このため，テキスト依存型において，数字や決められたキーワードの発声順序を，認識のたびに装置の側から指定するという方法が試みられています。しかしこの方法でも，最近の電子化された録音装置を用いれば，比較的容易に任意の単語系列を再生することができますので，万全ではありません。このため我々は，テキスト指定型話者認識法を提案しました[14]。

この新しい話者認識法では，認識装置を用いるたびに，装置の側から新しいテキストを，文字あるいは合成音声でユーザに提示し，ユーザが本人の声でそのテキストを正しく発声したと判定できるときのみ，話者照合の受理判定をします。その言葉としては，初めてでも比較的発声しやすい，短い文章を用います。決められたキーワードの発声順序を変えるのではなく，発声する言葉をその都度根本的に変えてしまうという方法です。予想のできない言葉が文字表示あるいは音声合成によって指定されますので，テープレコーダによってだますことは極めて難しくなります。

この基本的方法を図-2に示します。未知音声を，登録話者が多種多様なテキストを発声したときの文章（音声）モデルと比較できるようにするため，あらかじめ各登録話者ごとに，学習用音声を用いて，各音素のスペクトルを表すモデル（音素モデル）を作成しておきます。各登録話者の限られた量の学習用音声から，話者の特徴と音素の特徴を十分に表現した音素モデルを作成するために，多数話者の音声を用いて作った不特定話者音素モデルを，たねとして用います。この音素モデルは，HMM で表現されており，音素性の情報を

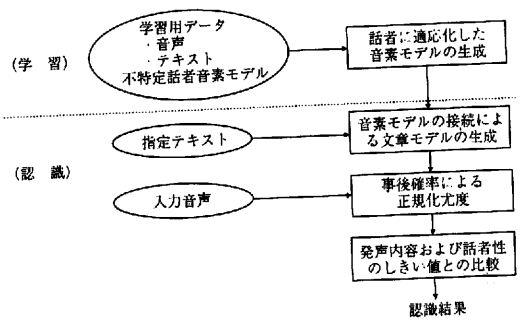

図-2 テキスト指定型話者認識法

十分に持っていますが，声の個人性に関する情報は持っていません。これを各話者の学習用音声を用いて，その話者に合うように自動的に適応化します[15]。

認識時には，指定したテキストに対応して，名乗った話者の音素モデルを接続して，そのテキストの文章（音声）モデルを作成します。入力音声をその文章モデルと比較し，類似性が十分に大きいときのみ，その音声を受理します。この類似性の尺度として，従来は尤度を用いていましたので，テキストや発声時期が変わると，その値が大きく変動する問題がありました。私共は，類似性を事後確率（入力音声がその文章モデルから生成されたものである確率）で表すことにより，信頼性を高める方法を提案しました[16]。

以上で述べた方法を用いて認識実験を行いました。各話者の音素モデルを作成するための適応化には，ある1時期に発声した10文章（1文章は平均約4秒）を用いました。話者認識のテストには，それと異なる2時期に本人を含む多数話者が発声した，異なる内容の5文章を1文章ずつ用いました。その結果，高い話者照合率が達成でき，本人の音声でも異なる発声内容である場合，すなわち他人が録音した音声を再生したような場合には，ほぼ正しく棄却できることが確かめられました[16]。

4.5 話者認識の実用システム

米国では，クレジットカード式のテレホンカードが用いられており，その盗難などによる悪用が社会的な問題となっています。このため，テレホンカード（"Voice Phone Card"）に話者照合機能を付加し，セキュリティ性の向上をはかるサービ

スが 1994 年から開始されています。話者照合機能自体は極めて高い認識性能を有しているわけではありませんが，本人がカードを紛失する確率，それを悪用される確率，話者照合の確率をかけ合わせれば，非常に高いセキュリティ性が得られるというのがセールスポイントです。具体的には，10 桁の社会保障番号 (Social Security Number) を発声します。うまく本人と確認されれば，例えば“Call Suzuki” (Suzuki さんの名前はあらかじめ登録しておきます) と発声すると，Suzuki さんに電話がかかります。すでに数十万人のユーザによって使われていると言われています。

米国では 1993 年より，テレホンカードや通常のクレジットカードに話者照合機能を付加するためのフィールドテストも行われています。ユーザはあらかじめ，10 数字あるいは 7 種類程度の短い言葉を発声し，カードにそれぞれの言葉に対応したユーザの声の特徴がパターンとして蓄えられます。装置には，入力された音声とカードに蓄えられたパターンとの照合をとる機能が取り付けられています。照合時には，装置側でその言葉の中から複数をランダムに選び，ユーザに提示します。ユーザがその言葉を順に正しく発声したときのみ，本人であると確認されます。

5. 音声認識における話者適応

音声認識において誰の声でも認識できるようにする，不特定話者の音声認識技術は，重要な課題の一つです[2]。音声から個人性情報を取り出すことができれば，それをキャンセルすることによって，個人差に起因する認識誤りを減らすことができるのではないか，と考えるのは自然ですが，それほど単純ではありません。話者が認識できることと，その情報を音声から取り除くことの間にはまだかなりギャップがあります。不特定話者音声認識の実現のためには，できるだけ多くの人の音声データを集めて，それらを包含できるような確率的音声認識モデルを構築し，それでも対応できない個人差に話者適応機能で対処するというのが，現実的な方法です。音声認識系は，発声者の個人差だけでなく，周囲の雑音やマイクロホン，伝送系などの特性の変化にも対応できなければならないので，適応機能を持つことは極めて大切です。

話者適応の方法に関しては，これまで日本の研究も含めて，種々の有力な方法が提案されており，最近の国際会議でも新しい試みが発表されています。これからは，種々の性質の音声変動に総合的に適応できる枠組みの構築が重要になると思われます[17]。

6. むすび

声の個人性は，つきつめれば人の数だけあるわけで，その中から普遍的な原理を見つけるのは極めて難しい課題です。この研究課題は，音声のあらゆる分野と関連を持っており，この解明は極めて大きな波及効果を生む可能性を秘めています。それだけ奥が深い研究分野であると言えます。話者認識技術の発展として，二人以上の人が対話をしている音声から，それぞれの人の発声区間を自動的に区別して取り出そうという研究も，最近行われています。最近話題となっている聴覚の情景分析 (Auditory Scene Analysis)[18] においても，声の個人性情報を把握することが必須になると思われます。

声の個人性に関しては，まだ解明できていることは極めてわずかで，まだまだこれからやらなければいけないこと，やりたいことが山ほどあります。この分野に興味と関心を持って下さる方が増えることを期待しています。

文　献

1) 古井貞熙，“音声の個人性情報と話者認識，” 信学会誌 **75**, 631-632 (1992).
2) 古井貞熙，音響・音声工学 (近代科学社，東京，1992), pp. 211-219.
3) 桑原尚夫，大串健吾，“アナウンサー音声の音響的特徴，” 信学論 **J66-A**, 545-552 (1983).
4) 古井貞熙，米山文明，“声のはなし，” NHK・FM. 日曜喫茶室 (1988 年 5 月 22 日放送).
5) 伊藤憲三，齋藤收三，“音声の音響的特徴パラメータが個人性の知覚に及ぼす影響，” 信学論 **J65-A**, 101-108 (1982).
6) 古井貞熙，赤木正人，“音声における個人性の知覚と物理関連量，” 音響学会聴覚研資 H-85-18 (1985).
7) 加藤和美，古井貞熙，“話者の個人性に対する聴覚の適応特性，” 音響学会聴覚研資 H-85-5 (1985).
8) 古井貞熙，“音声による個人識別の技術，” システム／制御／情報 **35**, 408-414 (1991).
9) 古井貞熙，“話者認識，” テレビジョン学会誌 **47**, 1600-1603 (1993).
10) S. Furui, “Cepstral analysis technique for automatic speaker verification,” IEEE Trans. Acoust.

Speech Signal Process. **ASSP-29**, 254-272 (1981).
11) F.K. Soong and A.E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," Proc. ICASSP 86, 877-880 (1986).
12) 松井知子, 古井貞熙, "音源・声道特徴を用いたテキスト独立形話者認識," 信学論 **J75-A**, 703-709 (1992).
13) 松井知子, 古井貞熙, "VQ, 離散／連続 HMM によるテキスト独立形話者認識法の比較検討," 信学論 **J77-A**, 601-606 (1994).
14) 松井知子, 古井貞熙, "連結音韻モデルによる話者認識," 音響学会音声研資 SP92-128 (1993).
15) 松井知子, 古井貞熙, "テキスト指定形話者認識のための話者適応化法," 音講論集 3-7-5 (1994.3).
16) 松井知子, 古井貞熙, "音韻・話者独立モデルによる話者照合尤度の正規化," 音響学会音声研資 SP94-22 (1994).
17) Y. Minami and S. Furui, "A maximum likelihood procedure for a universal adaptation method based on HMM composition," Proc. ICASSP 95, Detroit, TP02. 3 (1995).
18) "小特集「聴覚の情景分析」," 音響学会誌 **50**, 1006-1028 (1994).

古井　貞熙

昭43東大・工・計数卒。昭45同大学院修士課程了。同年NTT電気通信研究所入社。以後，同研究所において，音声認識，話者認識，音声知覚などの研究に従事。昭53～54ベル研究所客員研究員。現在NTTヒューマンインタフェース研究所古井特別研究室長。東京工業大学客員教授。工博。科学技術庁，IEEE，日本音響学会，電子情報通信学会などから，論文賞，著述賞など受賞。IEEE Fellow。