論文 / 著書情報
Article / Book Information

論題(和文)	話者認識の新技術 テキスト指定型話者認識
Title(English)	
著者(和文)	古井貞熙, 松井知子
Authors(English)	SADAOKI FURUI
出典(和文)	NTT技術ジャーナル, , , pp. 66-69,
Citation(English)	, , , pp. 66-69,
発行日 / Pub. date	1994, 4

話者認識の新技術

テキスト指定型話者認識

古井 貞熙 †1／松井 知子 †2

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への期待が高まっています．従来の話者認識法には，録音した音声によって装置がだまされてしまうという問題がありました．これに対処するため，テキスト指定型話者認識法を提案し，高い精度で話者認識ができることが確認されました．

話者認識とは

■話者認識の応用と特徴

音声に含まれる発声者の特徴（個人性情報）を用いて，コンピュータなどにより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います[1),2),3),4)]．これが実現できれば，①車や建物の入口の鍵，②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，③音声メール・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ています．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．

■話者認識の方法

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周波数スペクトルを計算します．図2に，2人の話者が「大声を」と発声した音声のスペクトル時系列（周波数スペクトルの時間的な変化）の例を，サウンドスペクトログラムで示します．この図の横軸は時間，縦軸は周波数で，色の濃い部分はその周波数成分のエネルギーが大きいことを示します．この図から分かりますように，同じ人の声でも変化しますので，各話者に対して，その代表的なスペクトル時系列を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を蓄えます．

次に，新たに発声された音声の話者を認識する場合には，その音声について学習のときと同じ方法で話者の特徴を抽出します．その特徴と，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，その値によって同一の話者によるものかどうか話者の判定を行います．

■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．図3に示すように，音声が，あらかじめ登録されている多数の人（A_1，A_2，…，A_M）のうちの誰の声であるかを判定するのが話者識別です．犯罪捜査などで複数の容疑者の中から最も声の似ている人を選ぶのは話者識別に該当します．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（A_C）の声であるか否かを判定するのが話者照合です．音声を本人確認の鍵として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，

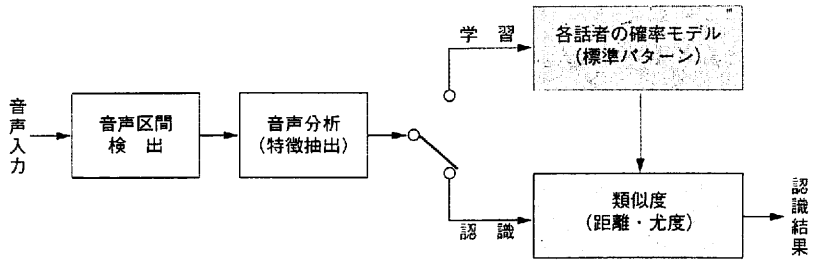

図1 話者認識系の基本的構成

†1 ヒューマンインタフェース研究所
古井特別研究室長
†2 同 上 研究主任

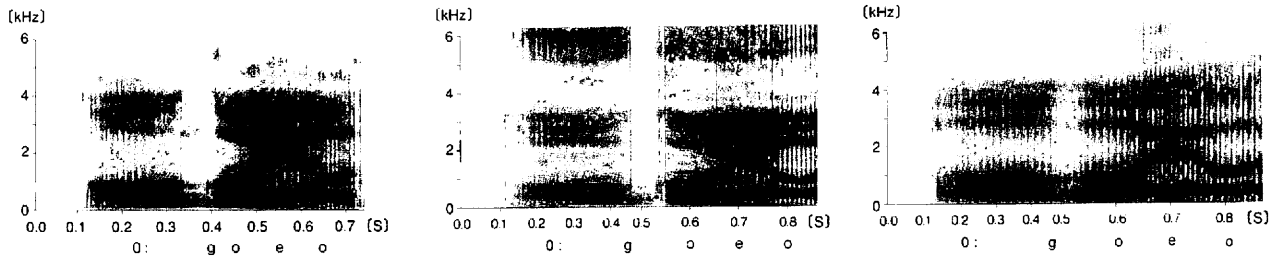

図2　「大声を」と発生した音声のサウンドスペクト グラムの例

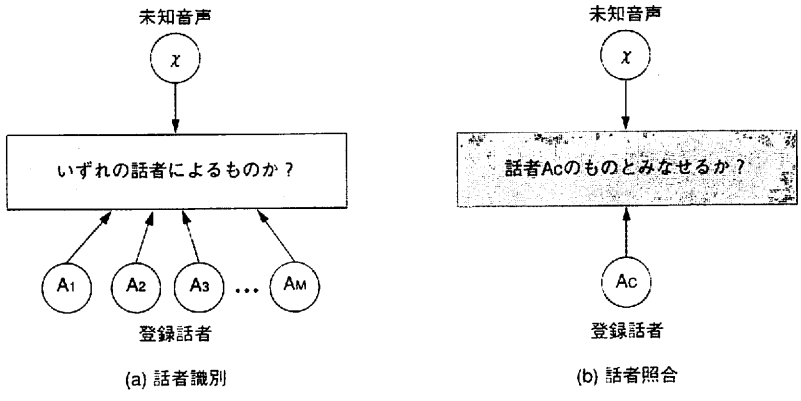

図3　話者識別と話者照合

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．したがって，例えば100人の音声について実験を行って99%の話者照合の精度が得られれば，登録話者の数が仮に1億人になっても，ほぼ99%の精度が実現できるといえます．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．

■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，各話者のキーワードの標準パターン（あるいはモデル）と入力音声とを比較して，どれだけ似ているかを調べるので，音声認識において確立されている技術を用いることができます．しかも，「ア」，「イ」などの個々の音韻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．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．

ところが，テキスト依存型・独立型に共通して，従来の方法には，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致命的な問題がありました．この欠陥は，昨年公開されたアメリカ映画「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

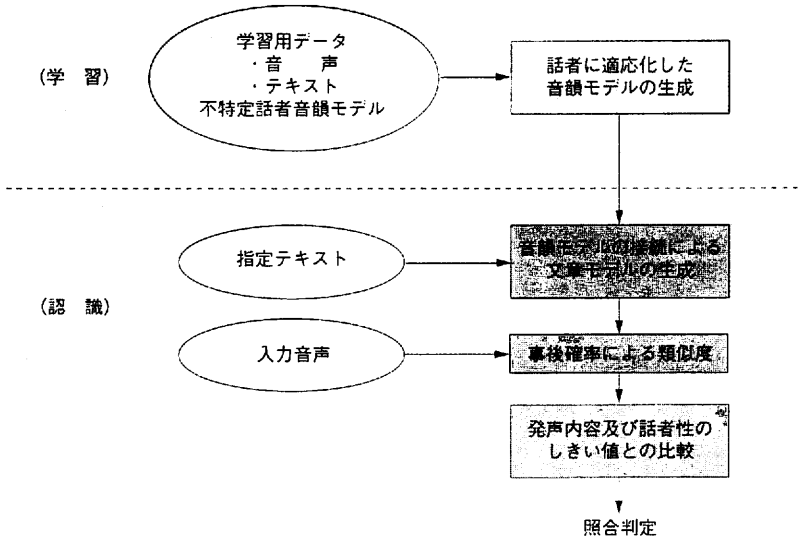

図4　テキスト指定型話者認識法

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，私たちは最近，テキスト指定型話者認識法を提案しました．

テキスト指定型話者認識

■テキスト指定型話者認識法の原理と方法

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[5),6)]．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．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．このようにすればテープレコーダで再生できる言葉とは全く異なった，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4に示します．あらかじめ各登録話者ごとに，学習用音声を用いて各音韻のスペクトルを表すモデル（音韻モデル）を作成しておきます．未知音声を，各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です．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韻モデルを接続し，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て類似性が十分に大きいときのみ，その音声を受理します．

■技術的ポイント

この方法の重要な技術的ポイントは2つあります．1つは，各登録話者の限られた量の学習用音声から，いかにして話者の特徴と音韻の特徴を十分に表現した音韻モデルを作成するかということです．始めの学習時のみとはいっても，各話者に例えば数十あるいは百種類にも及ぶ文章を発声してもらうのは，負担が大きすぎるので現実的ではありません．このため，せいぜい10種類程度の文章を発声した音声から音韻モデルをつくろうとすると，この中に含まれない音韻が出てきます．含まれている音韻でも，前後の音韻の影響を受けてスペクトルが変形しています．このため，すべての音韻のモデルを適切につくるのが極めて難しいという問題があります．2つ目は，正しいテキストが発声されているか，本人の声であるかの尺度をいかに適切に計算するかということです．このため筆者らは，次のような方法を用いることにしました．

■音韻モデルの作成法

まず準備として，多数話者の音声を用いて共通の音韻モデルを作成しておきます．この音韻モデルは，音韻性の情報を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．これを不特定話者音韻モデルと呼び，各登録話者の音韻モデルを作成するための種として用います．各話者に対しては，その学習用音声を用いて，不特定話者音韻モデルをその話者に合うように変更します．この過程を適応化と呼びます．その処理の流れを図5に示します．学習用音声のテキスト（発声内容）は既知ですので，それに応じて不特定話者音韻モデルを接続して文章モデルをつくり，それと学習用音声を対応付けます．両者ができるだけ合うように，モデルのパラメータを修正します．これを図に示すように何度か繰り返すことによって，モデルのパラメータが話者に適応化されます．

具体的な音韻モデルとしては，音声認識において現在広く用いられている

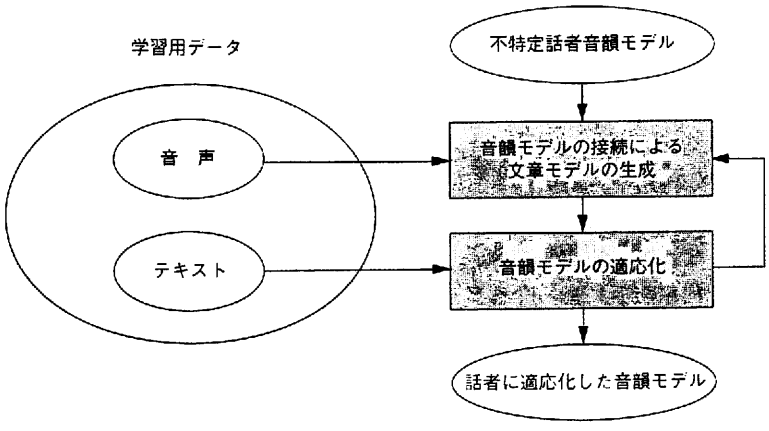

図5　話者に適応化した音韻モデルの作成法

HMM (Hidden Markov Model:隠れマルコフモデル)を用います[7]. 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，複数の状態間の遷移確率と，各状態遷移におけるスペクトルの出現確率分布で，モデルが表現されます．上記の適応化では，出現確率分布関数の平均値と重み係数を修正します．出現確率分布の分散の値は，少数の学習音声を用いて適応化することが困難なので，不特定話者音韻モデルの値をそのまま用います．

■事後確率による類似性の表現

話者照合と発声内容に関する判定の尺度に関しては，文章モデルと入力音声との類似性を事後確率（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．従来は，仮定した文章モデルからその入力音声が生ずる確率値である尤度という量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．この事後確率を用いる方法によって，認識性能が大きく向上することが確認されています[8].

■評価実験

以上で述べた方法を用いて認識実験を行いました．使用した音声サンプルは，男10名，女5名が約5カ月にわたる3つの時期に発声した種々の文章音声です．音声のスペクトルを表す特徴パラメータとしてはケプストラムを用いました．ケプストラムは，対数スペクトルを逆フーリエ変換したもので，1～16次程度の比較的低次のパラメータによって，細かい変動を除去したなめらかなスペクトルが表現できます．各話者の音韻モデルを作成するための適応化には，話者ごとに，ある一時期に発声した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つの時期に，本人を含む全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，100%の話者照合率，すなわち，本人の音声と他人の音声の完全な分離が達成できました．また，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，99.7%の精度で正しく棄却できることが確かめられました[6),8)].

おわりに

これまでの実験では，テキスト指定型話者認識により，極めて高い話者認識性能が得られましたが，ここで用いた音声データは，静かな環境で，高品質のマイクロホンを用いて録音した音声です．マイクロホンが変わったり電話を通したりすると，音声のスペクトルは変化します．実際の場でどの程度の性能が得られるかを確かめるため，現在，多数話者による実際の電話音声を収録しています．今後これらの音声を用いた評価実験を行う予定です．

■参考文献

1) 古井貞熙："音声による個人識別の技術"，システム/制御/情報，35，7，pp.408-414，1991
2) 古井貞熙："音声の個人性情報と話者認識"，電子情報通信学会誌，75，6，pp.631-632，1992
3) 古井貞熙："音響・音声工学"，近代科学社，pp.211-219，1992
4) 古井貞熙："ディジタル音声処理"，東海大学出版会，pp.193-206，1985
5) 松井，古井："連結音韻モデルによる話者認識"，電子情報通信学会音声研究会資料，SP92-128，1993
6) T. Matsui and S. Furui: "Concatenated phoneme models for text-variable speaker recognition", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Minneapolis, II-391-394, 1993
7) 菅村，北脇："音声認識"，本誌，Vol.6 No.2，pp.58-62，1994
8) 松井，古井："テキスト指定型話者認識のための事後確率に基づく尤度正規化法"，日本音響学会秋季研究発表会講演論文集，1-7-20，1993