論文 / 著書情報
Article / Book Information

論題(和文)	話者認識
Title(English)	
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	テレビジョン学会誌, Vol. 47, No. 12, pp. 1600-1603
Citation(English)	, Vol. 47, No. 12, pp. 1600-1603
発行日 / Pub. date	1993, 9
権利情報 / Copyright	本著作物の著作権は映像情報メディア学会に帰属します。 Copyright (c) 1993 Institute of Image Information and Television Engineers.

4-4 話　者　認　識

古　井　貞　熙†

1. ま　え　が　き

音声に含まれる個人性情報を用いて，誰の声であるかを自動的に判定することを，話者認識（Speaker Recognition）という．これが実現できれば，電話によるバンキング，買い物，および情報案内サービス，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，音声によって誰であるかが確認できるので，極めて便利になると期待されている．最近公開されたアメリカ映画「スニーカーズ」では，話者認識装置が建物への非登録者の侵入防止に用いられている．話者認識の詳しい原理や基本的方法などに関しては，すでに種々の解説があるので[1)~4)]，本解説では，2～3節で基本的原理を説明した後，4節以降で最近の技術動向に絞って紹介したい．

2. 話者認識の原理

話者認識の形態は，話者識別（Speaker Identification）と話者照合（Speaker Verification）に分けられる．話者識別とは，入力音声が，あらかじめ登録されている誰の声であるかを判定することである．一方，話者照合では，入力音声と同時に自分が誰であるかを名乗り，その音声が本当にその話者の声であるか否かを判定する．いずれの場合も，入力音声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，その値によって判定を行う．話者識別の場合は，多数の登録話者の中から最も類似度の高い話者を選び，話者照合の場合は，類似度が一定の閾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する．音声を本人確認の一種の鍵として用いる多くの応用のほとんどは，話者照合に該当する．話者識別の性能は，登録話者の数が大きくなると低下するが，話者照合では，登録話者の数がある程度以上大きければ，性能はその数には依存しない．

† NTTヒューマンインタフェース研究所
"Speaker Recognition" by Sadaoki Furui (NTT Human Interface Laboratories, Tokyo)

テレビジョン学会誌 Vol. 47, No. 12, pp. 1600～1603 (1993)

話者認識の方法はさらに，あらかじめ決まっている言葉(キーワード)を発声しなければならない発声内容依存(限定)型と，任意の言葉を発声してよい発声内容独立型に分類できる．一般に前者の方が高い認識性能を得るのが比較的容易であるが，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る．また，一般に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る．このため，発声内容独立型やそれを基本とする方法が最近活発に研究されている．発声内容依存型の方法には，映画「スニーカーズ」で示されているように，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機械を容易にだますことができるという問題もある．このため我々は最近，5節で紹介する発声内容指定型の話者認識法を提案した．

3. 話者認識に用いる音声の特徴パラメータ

音声に含まれる情報は，音声生成のメカニズムから，音源に関連したピッチ(基本周波数)，発声レベルなどの情報と，声道に関連したスペクトル包絡情報に分けることができる．話者認識には，このうちの後者，あるいは両者が組合せて用いられる．スペクトル包絡を表現する方法には種々のものがあるが，最近では音声認識の場合と同様に，ケプストラム（対数スペクトルの逆フーリエ変換）とその線形回帰係数（Δケプストラム）が用いられることが多い．これらの特徴パラメータは，指紋と違って，同じ人の同じ言葉でも発声のたびに変化する．数カ月も時間が経ったり電話を通ったりすると大きく変化することもある．このため，これらの変化を吸収するようなパラメータ変換や，統計的処理，パターン認識処理などを施すことが必要である．

4. 発声内容独立型話者認識

発声内容独立型の最近の代表的な方法には，ベクトル量子化による方法とHMM（Hidden Markov Model；隠れマルコフモデル）による方法がある．その構成の例を図1に示す[5][6]．ベクトル量子化による方法では，各登録話者について，任意の文章を発声した学習用音声から短時間ごとに，スペクトル包絡を表す特徴ベクトル，あるいはそれにピッチを組合せたベクトルを抽出し，それらをクラスタ化して符号帳を作成する．各話者の違いは符号帳の違いによって表現される．ピッチは有声音区間からしか抽出されないので，それを用いる場合は図1に示すように，各話者について有声音区間用と無声音区間用の符号帳を別々に作成する．認識時には，入力音声の有声音区間，無声音区間それぞれについて各登録話者の符号帳でベクトル量子化し，入力音声全体にわたる平均量子化歪みを求める．その値によって話者識別あるいは話者照合の判定を行う．

HMMによる場合は，ベクトル量子化の符号帳の代わりに各登録話者の学習用音声からHMMを作成する．各話者の特徴は，モデルの違い，具体的にはHMMの状態間の遷移確率と，各状態におけるパラメータの出現確率分布(複数のガウス分布の混合)の違いによって表現される．入力音声を各登録話者のモデルに与えたときの音声区間全体の平均尤度によって，話者認識の判定を行う．

実験の結果，ベクトル量子化による方法とHMMによる方法では，ほぼ同程度の認識性能が得られること，後者の場合，HMMの状態間の遷移情報は個人差の表現にはほとんど寄与しないため，状態数の増加と状態内の混合分布数の増加は，同程度の認識性能向

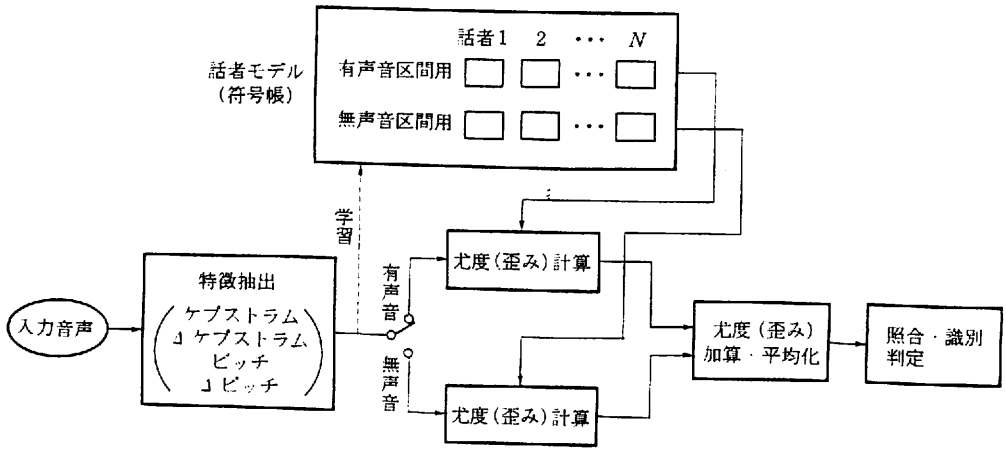

図 1 発声内容独立型話者認識システムの構成

上効果があり，性能はほぼ両者の積によって決まることがわかった．また，いずれの方法の場合も，歪みあるいは尤度を計算する際に，入力音声のすべての部分を用いずに符号帳あるいはモデルとの重なりのある部分のみを用いた方が，学習音声と入力音声に含まれる音韻の違い，時期による音声の変動などの影響を削減する効果があるため，より高い認識性能が得られることが確認されている．ピッチ情報に関しては，それだけでは低い認識性能しか得られないが，スペクトル包絡情報と組合せると大きな効果があることがわかっている[5][6]．

発声内容独立型話者認識に関して，最近，音声の動的特徴に含まれる個人性を表現する方法として，ケプストラムの多次元自己回帰（MAR：Multivariate Auto-Regressive）モデルを用いる方法が提案され[7]，類似度尺度を適切に選択すれば，かなり高い認識性能が得られることが確かめられている[8]．

5. 発声内容指定型話者認識

2節で述べた発声内容依存型の問題点，すなわち録音した音声で認識装置がだまされてしまうことを防ぐ方法として，数字や決められた単語の発声順序を，認識のたびに変えるという方法が試みられている．しかしこの方法でも，最近の電子化された記憶装置を用いれば，比較的容易に任意の単語系列を再生することができるので，万全ではない．そこで我々は，認識のたびに装置の側から新しいテキストをユーザに示し，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時のみ話者照合の受理判定をする方法，すなわち，発声内容指定型話者認識法を提案した．この方法のブロック図を図2に示す[9]．

この方法の場合，話者と発声内容の両方の判定を行う必要があり，しかもテキストが毎回変更されるので，各登録話者ごとに，その話者の声に適応した音韻モデル（話者音韻モデル）を作成しておく．認識時には，指定したテキストに応じて名乗った話者の音韻モデルを接続して，そのテキストの文章（音声）モデルを作成する．入力音声をその文章モデルと比較し，類似性が充分に大きいときのみ，その話者が正しいテキストを発声したとみなして受理する．

この方法の重要な技術的ポイントは，各登録話者の限られた量の学習音声から，いかにして話者の特徴と音韻の特徴を充分に表現した音韻モデルを作成するかという点と，判定の閾値をいかに適切に設定するかという点にある．筆者らは，多数話者の音声を用いて作成した不特定話者用の音韻HMMを，登録話者の学習音声を用いて，その話者に適応化する方法を試みた．その処理の流れを図3に示す．学習音声のテキストは既知であるので，それに応じて不特定話者用音韻モデルを接続して文章モデルを作り，それと学習音声を対応づける．その対応づけができるだけうまくいくように，HMMのパラメータ（特徴パラメータの出現確率分布関数の平均値と重み係数）を修正する．これを図に示すように何度か繰り返すことによって，HMMのパラメータが話者に適応化される．

話者照合と発声内容に関する判定の閾値に関しては，文章モデルと入力音声との類似性，すなわち尤度の値を事後確率におきかえ，テキストや時期による音声の変動を正規化することにより，閾値を安定化する方法を提案した．

これらの方法を用いて認識実験を行った．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

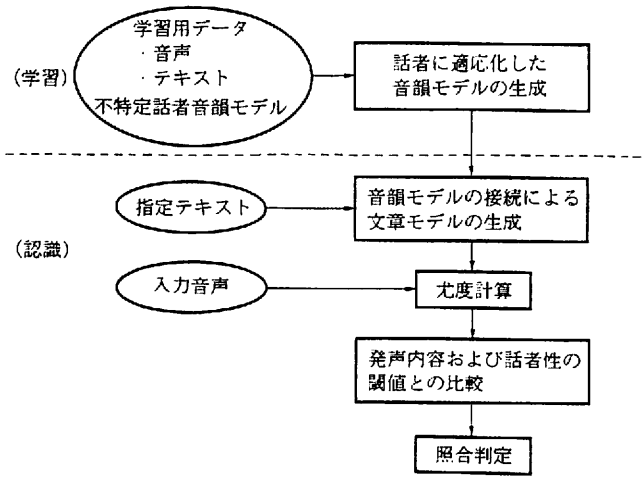

図2 発声内容指定型話者認識法のブロック図

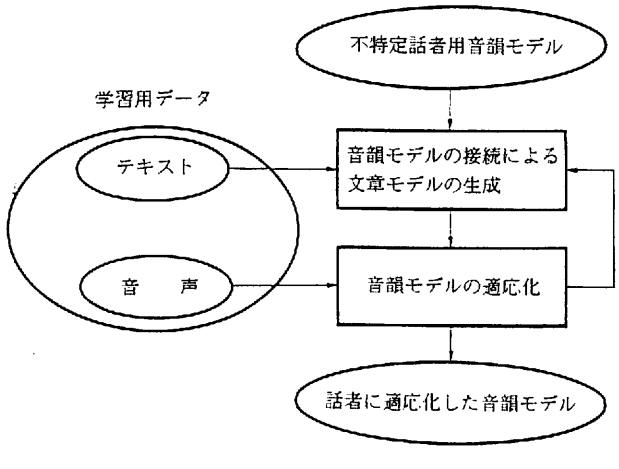

図3 話者に適応した音韻モデルの生成法

る3時期に発声した文章音声で，特徴パラメータとしてはケプストラムを用いた．学習には，話者ごとに，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，テストには，それと異なる2時期に発声した異なる内容の5文章を1文章ずつ用いた．その結果，100%の話者照合率が得られ，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た[9]．

6. むすび

ここでは，最近の話者認識技術の動向について解説した．音声認識における最も重要な課題のひとつに，不特定話者の音声認識精度の向上があり，そのための方法として，音声認識システムを話者に適応化させる研究が行われている．話者認識の研究は，この話者適応の研究と密接に結びついている．前節で紹介した発声内容指定型話者認識における音韻モデルの生成過程は，音声認識における話者適応過程と同じである．今後は，話者認識と音声認識がかなり一体となった形で研究が進められて行くことであろう．

話者認識技術の実際のフィールドにおける有効性を評価するため，大量な電話音声データベースの収集と，それを用いた大規模実験が計画されている．話者認識システムの良否を決定する最も大きな要因は，その認識性能であるが，ユーザへの指示の出し方，認識誤りを生じた場合の対処の仕方など，ユーザインタフェースの観点からの研究も極めて重要である．最近，話者認識技術の発展として，2人以上の人が対話をしている一連の音声から，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行われている[10][11]．今後は，このような多角的な研究が活発に進められるようになるであろう．

(1993年6月7日受付)

〔参考文献〕

1) 古井："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 (July 1991)
2) 古井："音声の個人性情報と話者認識"，信学誌，75，6，pp. 631-632 (June 1992)
3) 古井："音響・音声工学"，pp. 211-219，近代科学社 (1992)
4) 古井："ディジタル音声処理"，pp. 193-206，東海大学出版会 (1985)
5) 松井，古井："声道・音源特徴を用いたテキスト独立型話者認識"，信学論，J75-A，4，pp. 703-709 (Apr. 1992)
6) 松井，古井："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"，信学技報，SP91-89 (1991)
7) C. Montacie, P. Deleglise, F. Bimbot and M.-J. Caraty: "Cinematic Techniques for Speech Processing: Temporal Decomposition and Multivariate Linear Prediction", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., San Francisco, I-153-156 (1993)
8) グリフィン，松井，古井："MARモデルによるテキスト独立形話者認識のための距離尺度"，信学技報，SP93-14 (1993)
9) 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)
10) 渡辺，村上，杉山："未知・複数信号源クラスタリング問題—未知話者クラスタリングへの応用—"，信学技報，SP92-45 (1992)
11) G. Yu and H. Gish: "Identification of Speakers Engaged in Dialog", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Minneapolis, II-383-386 (1993)

古井（ふるい） 貞熙（さだおき） 昭和45年，東京大学大学院修士課程修了．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．昭和53年～54年，米国ベル研究所客員研究員．昭和61年，NTT基礎研究所第4研究室長，平成元年，NTTヒューマンインタフェース研究所音声情報研究部長．平成3年より，古井特別研究室長となり，現在に至る．工学博士．