論文 / 著書情報
Article / Book Information

論題(和文)	音声による個人識別の技術
Title	
著者(和文)	古井 貞熙
Authors	SADAOKI FURUI
出典 / Citation	システム/制御/情報, Vol. 35, No. 7, pp. 408-414
Citation(English)	, Vol. 35, No. 7, pp. 408-414
発行日 / Pub. date	1991, 9

システム/制御/情報, Vol. 35, No. 7, pp. 408～414, 1991

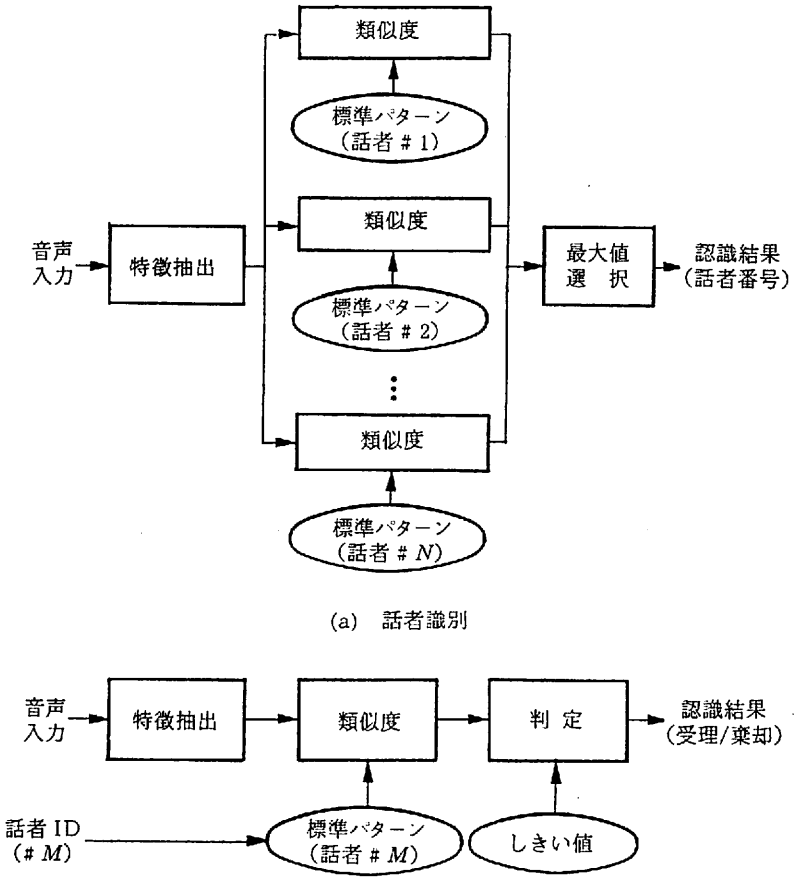
解　説

音声による個人識別の技術

古　井　貞　熙*

1. まえがき

音声の個人差，すなわち音声波に含まれる個人性情報を用いて，誰の声であるかを判定することを話者認識 (speaker recognition) という．話者認識の方法としては，人が声を耳で聞いて判断する方法，音声のスペクトルの時間パターンを表示したスペクトログラム（声紋）を目で見て判断する方法，音声に含まれる種々の物理的特徴を取り出し，コンピュータなどによって自動的に判断する方法，などがある．前2者の方法では，判断する人の主観によって判定が変わる問題があるが，自動的な方法では客観的な判断基準で判定を行なうことができる．最近では，コンピュータやLSIの進歩もあいまって自動的な方法が進歩し，一部では実際に用いられるようになってきている．現在の主な用途は，電話によるバンキングや買物サービス，音声・Faxメールなどの電話サービス，電話網を使ったコンピュータのリモートアクセス，高度な情報提供サービス，などにおけるアクセスコントロールである．家や特殊な場所への非登録者の侵入防止のための音声キーなども検討されている．ここでは自動的な方法に限定して，最近の話者認識，すなわち音声による個人識別の技術動向について述べる[1)~5)]．



第1図　話者識別と話者照合の構成

2. 話者認識の基本的方法

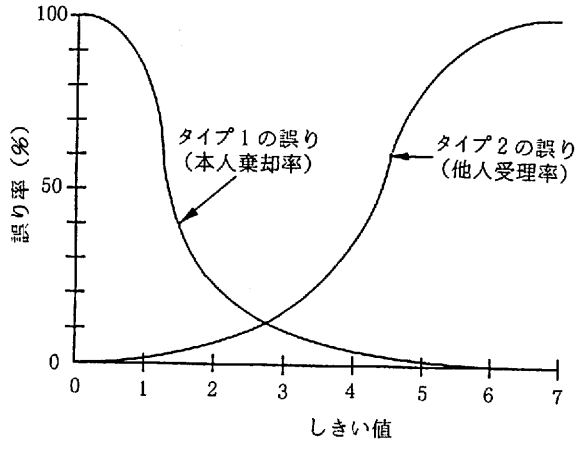
話者認識の形態は，話者識別 (speaker identification) と話者照合 (speaker verification) に分けることができる．第1図に示すように話者識別とは，入力音声が，あらかじめ登録されている誰の声であるかを判定するものであり，話者照合とは，入力音声と同時に自分が誰であるかを入力して，その音声が本当にその人の音声であるか否かを判定するものである．いずれの場合も，入力音声と登録されている話者の標準パターンとの類似度を調べ，その値に基づいて判定を行なう．話者識別の場合は，多数の登録話者の内から最も類似度の高い話者を選び，その話者の音声であると判定する．話者照合の場合は，本人の標準パターンとの類似度が，一定の

* NTT ヒューマンインタフェース研究所
音声情報研究部

Key Words : speaker recognition, speaker identification, speaker verification.

しきい値よりも大きければ本人の音声であると判定し，それ以外の場合は他人の音声であると判定する．1. で述べたような，音声を本人確認の一種の鍵として用いるケースは，話者照合に対応する．

話者識別の性能は，登録話者の内の本人以外の話者が選択される誤り率で評価する．当然ながら登録話者の数が多くなればそれだけ難しくなるので，話者識別の誤り率は，方法が一定でも登録話者の数が増えるにつれて単調に増加する．一方話者照合の誤り率は，本人が棄却される誤り率（タイプ1の誤り）と他人が受理される誤り率（タイプ2の誤り）で評価される．これら2種の誤り率と判定のしきい値との間には，一般に第2図のような関係がある（横軸は非類似度すなわち距離値で示している）．しきい値を大きくすればタイプ2の誤りが大きくなり，小さくすればタイプ1の誤りが大きくなる．このように両者の誤りにはトレードオフの関係があるので，一般にしきい値は，2種の誤りの相対的な重要性（タイプ1の誤りが多少大きくなっても，タイプ2の誤りがある一定値を越えないようにするなど）に従って設定される．しきい値をある値に固定したときのこれらの誤り率は，話者識別と異なり，登録話者の数には依存しない．話者照合システムの評価には，ROC 曲線 (receiver-operating-characteristics curve)[1] を用いる場合もある．



第2図　話者照合の判定のしきい値と2種の誤り率の関係

話者認識の方法はさらに，発声すべき言葉（キーワード）をあらかじめ決めておく発声内容依存型（限定型）と，任意の言葉を発声すればよい発声内容独立型に分類できる．発声内容依存型の場合は，標準パターンと入力音声に含まれる物理特徴を直接比較すればよいので，発声内容独立型よりも比較的容易であり，これまでに実用化されている方式はすべてこの型である．発声内容依存型において，キーワードを話者によって違えておけば，その違いも手がかりとして用いることができるので，それだけ認識性能を高めることができる．

話者認識，特に話者照合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，どのような言葉を発声しても話者を認識することができれば，応用の分野は一層広がると期待される．犯罪捜査などにおいては，必ずしも同じ言葉どうしを比較できず，発声内容独立型が要求される場合もある．人間は通常，発声内容に拘らず相手が誰であるかを判定することができるので，コンピュータを人間に近付けるという意味でも，発声内容独立型の研究の意味があると考えられる．ただし，音声に含まれる個人性情報は，一般に発声内容すなわち音素によって変化し，個人性情報と音韻性情報の間には交互作用があるため，両者を明確に分離するのは極めて難しい．このため，発声内容独立型で話者を認識するには種々の工夫が必要であり，この方法はまだ基礎研究段階にある．

一般に，発声内容依存型では認識に2〜3秒程度の音声を用い，独立型では，学習に10〜30秒程度の音声，認識に5〜30秒程度の音声を用いる場合が多い．依存型でも実際のシステムにおいては，詐称者（本人以外の話者）が録音した音声を用いることを防ぐため，数字や決められた単語を，認識のたびにランダムに並び変えた順序で発声させる形態をとることが多い．

3. 話者認識に用いる音声の物理特徴

音声による個人識別には，鍵，カード，指紋，サイン，ID番号などを用いる方法に比べて，忘れたり紛失したり，他人に盗まれる心配がない，電話などを用いて遠隔地からでも判定ができる，などの特徴がある．しかし，指紋などと決定的に違う点として，同じ人が同じ言葉を発声しても，音声の波形やスペクトルはそのつど変化してしまうということがある．このため，ある程度時期がたっても変化しにくい物理特徴を探索して用いたり，変化をキャンセルできるような方法を工夫して用いることが必要である．

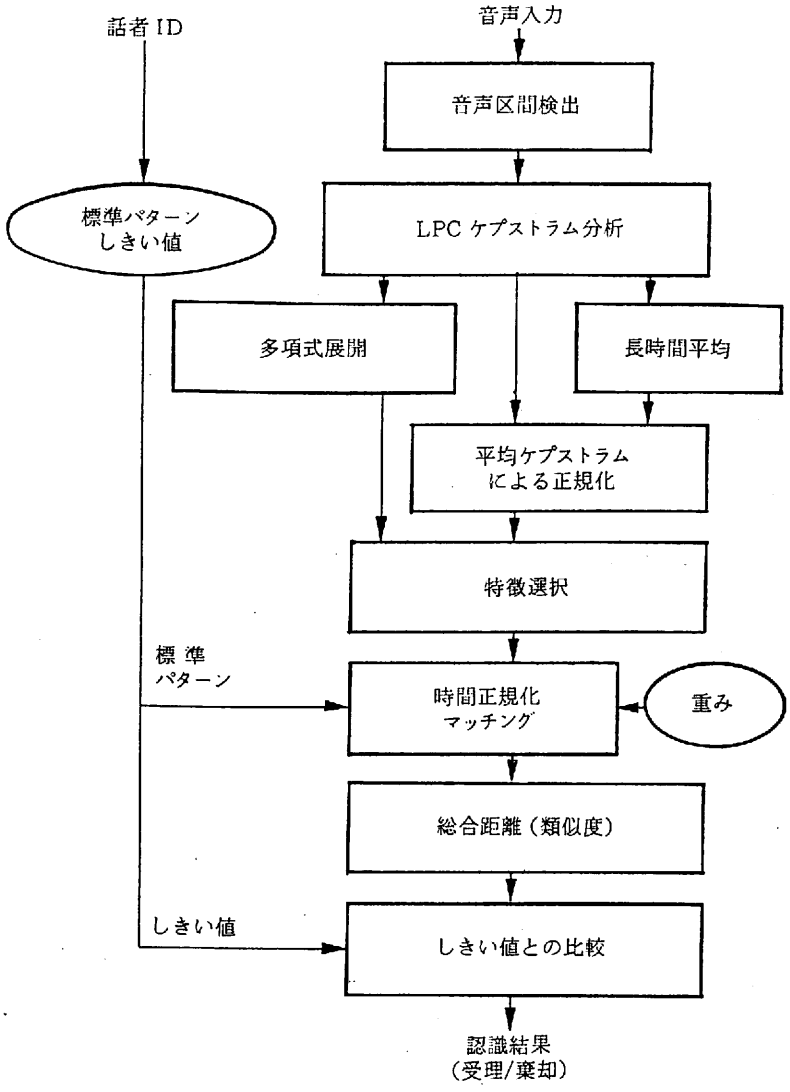
話者認識に用いる物理特徴の持つべきその他の条件としては，音声波から自動的に抽出しやすいこと，詐称者がまねしにくいこと，伝送系やマイクロホンの変動の影響を受けにくいことなどがある．実際に用いられる物理特徴としては，

- スペクトル包絡（スペクトル概形）… 声の質
- ピッチ（基本周波数）… 声の高さ，アクセント，イントネーション
- 発声レベル … 声の大きさ
- 発声速度 … 発声の速さ

などがあるが，主として用いられるのは，スペクトル包

絡とピッチである．ただしピッチは，有用な特徴ではあるが，正確な測定が難しい（特に雑音がある場合），他人がまねしやすい，ストレスなどによる話者内変動が大きいなどの問題がある．スペクトル包絡に関しては，他人がまねすることは極めて難しく，これによる問題は少ないことが確認されている．これらの特徴を用いる具体的方法としては，発声内容依存型の場合は，これらの特徴の時系列を直接用い，発声内容独立型の場合は，細かい時間ごとの特徴を独立に取り扱ったり，統計量に交換してから用いることが多い．

スペクトル包絡の話者内（長期間）変動を除去する方法としては，その全体的傾斜を除去する逆フィルタ（スペクトル等化）の有効性が高いことが確認されている[6]．この方法は，近似的にスペクトル包絡に含まれる声帯波情報を除去し，声道情報のみを用いることに相当する．さらに，類似度の計算において，各パラメータに長期間にわたる変動の分散の逆数の重みをかけ，安定なパラメータにより大きな重みを与えることがよく行なわれる．話者内の変動だけでなく，話者間の変動との相対関係を考慮した判別分析などによる重み付けを行なう場合もある[4]．また，話者照合において，受理された本人の音声とそれまでの標準パターンとの逐次平均化を繰り返して，標準パターンの更新を行なうのも，高い本人受理率を維持するうえで有効な方法であり，特に最初の10～15時期において有効であることが確かめられている[7]．



第3図　ケプストラムの静的および動的特徴と動的計画法を用いた発声内容依存型話者照合システムの構成

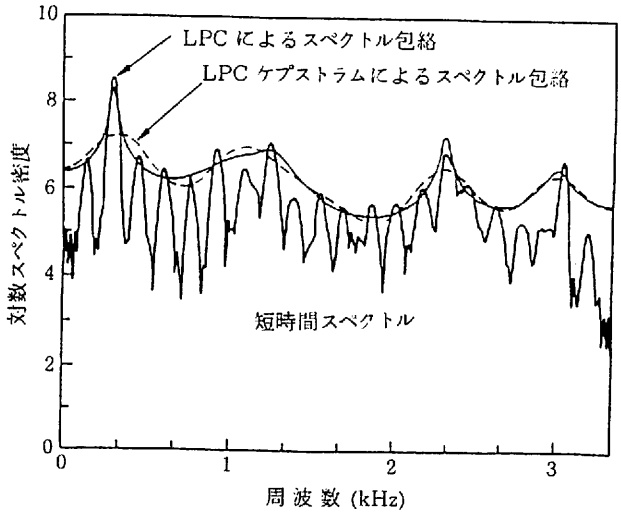
4. 発声内容依存型話者認識

4.1 動的計画法を用いる方法

第3図に，筆者らによる典型的な発声内容依存型話者照合システムの構成を示す[8]．発声する内容は，1.5秒長程度の決まった短文章である．音声波から，10 msごとに1～10次のLPCケプストラム（係数）を抽出する．LPCケプストラムは，LPC分析（線形予測分析）法によるスペクトル包絡を対数変換した後，さらにフーリエ逆変換したもので，実際の計算では，10 msごとに音声から抽出した線形予測係数を用いて，直接的に再帰式によって求めることができる．LPCケプストラムはその定義から，低次の係数はスペクトルの概形を表わし，高次の係数はスペクトルの微細構造を表わす．このため1～10次程度の比較的低次の係数に限定することによって，第4図に例を示すように，スペクトル包絡を効率的かつ安定に表現できる特徴がある．

つぎに，各次数のLPCケプストラムの時系列から，動的特徴として，10 msごとに90 msの区間における1次（傾斜）と2次（曲率）の多項式展開係数を計算する．これらの展開係数の内，話者認識での有効性の比較的高い係数を，多数話者の音声を用いてあらかじめ調べておき，これらの10 msごとの時系列をLPCケプストラムの時系列と組み合わせて，認識に用いる．

LPCケプストラムについては，各次数ごとに，短文章の音声区間全体での平均値を計算し，この平均値を10 msごとの実際の値から差し引くことによって逆フィルタ処理を行なっている．ケプストラム上での減算は，

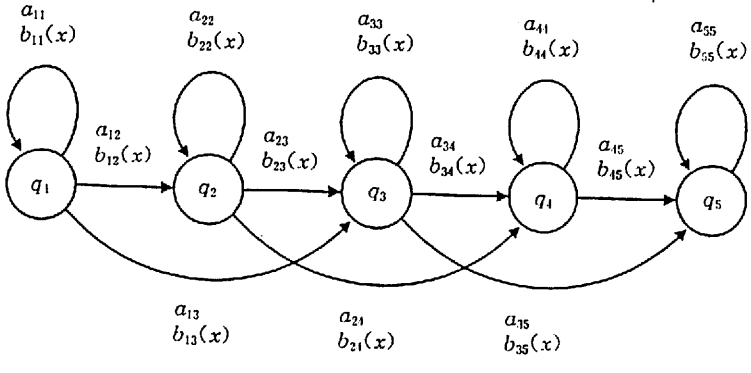


第4図　LPC ケプストラムによるスペクトル包絡表現の例

スペクトル上での除算に対応するので，この処理により，伝送歪と音声の時期的変動を正規化することができる．多項式展開係数は，その定義から，定常的な歪の影響を受けないので，正規化処理を適用する必要はない．

短文章を発声したときの音声は，発声のたびに時間的に非線形に伸縮するので，入力音声と標準パターンの類似度を計算する際，同じ音素どうしが対応するように，一方の時間軸を非線形に伸縮する必要がある．この処理は，よく知られているように，動的計画法 (DP) を用いて容易に行なうことができる[1]．こうして音声全区間にわたる入力音声と標準パターンの平均的類似度を計算し，これを各話者の変動許容値（しきい値）と比較して，話者照合の判定を行なう．なお，類似度の計算において，ケプストラムおよび多項式展開係数の各パラメータに対して，分散の逆数による重み付けを行なっている．

筆者らがこの方法を用いて，研究所内の交換機を通った男女計 120 名の電話音声によるオンライン実験を約 6 か月にわたって行なった結果，97～98 % の良好な認識率が得られている[2),3)]．



第5図　HMM（隠れマルコフモデル）の例
（a_{ij}：遷移確率，$b_{ij}(x)$：出現確率）

4.2　HMM による方法

音声認識において，隠れマルコフモデル (HMM, Hidden Markov Model) による方法が，現在広く用いられている．これは，たとえば第5図に示すようなマルコフモデルを用いて，左端の状態から右端の状態に至る確率的な状態遷移と，その時の確率的なスペクトル出力によって，単語や音素のスペクトル時系列を表現する方法である．この方法によれば，従来のスペクトル時系列をそのまま標準パターンとして蓄える方法に比べて，発声のたびに生ずる変動を，大量の学習サンプルに基づいて，確率的なモデルとしてより柔軟に表現できる特徴がある．このため，HMM を話者認識に用いる試みが，最近いくつか行なわれている[9),10)]．

Naik らの実験によれば，1 単語を発声した音声を用いる話者照合システムにおいて，単語のテンプレートと DP による従来法を用いたときの誤り率が 6.2 % であったのに対し，HMM を用いれば 4.6 % に低下することが確認されている[10)]．ただし，HMM による方法では一般に，DP による方法よりも多量の学習用音声を必要とするため，話者ごとにモデルを構成しなければならない話者認識の場合には，実用的な意味でやや問題がある．

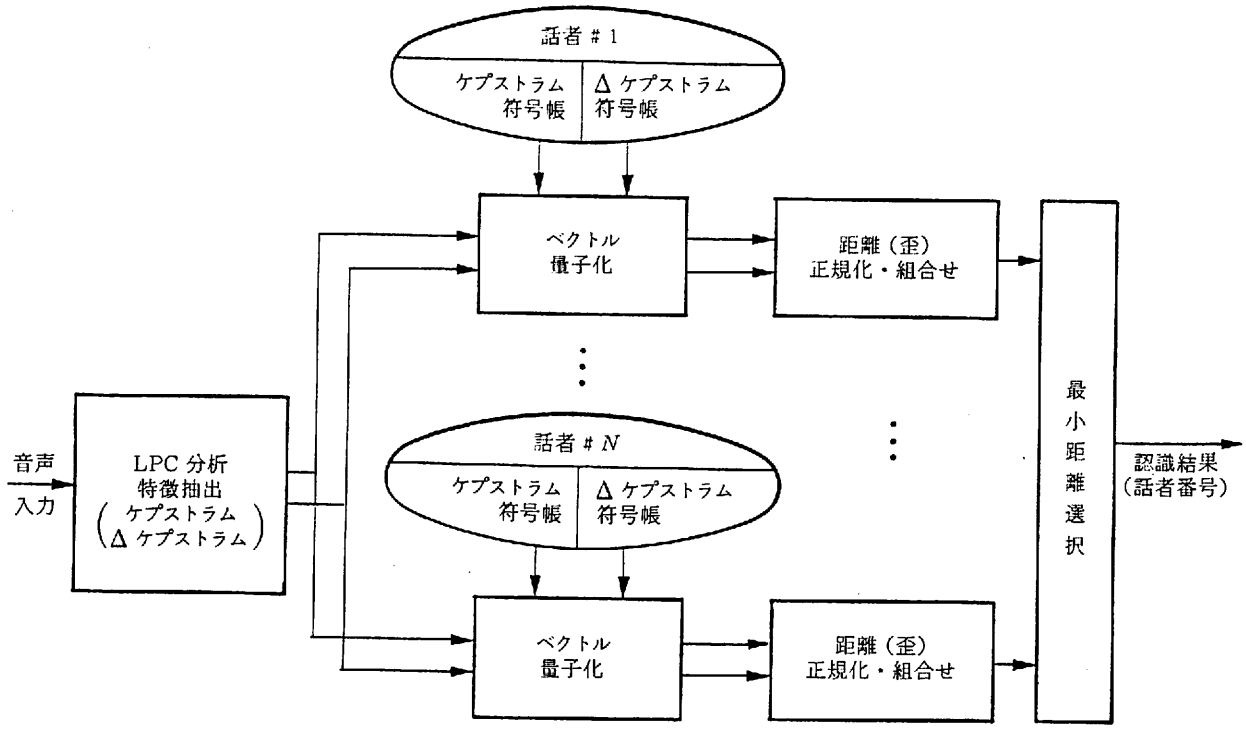
5.　発声内容独立型話者認識

5.1　長時間平均スペクトルによる方法

発声内容独立型の典型的な方法は，長時間平均スペクトルを用いる方法である[1)]．これは，短時間ごとの音声スペクトルを，発声内容による違いが無視できる程度に長時間にわたって平均し，その平均スペクトルの個人差によって話者を認識する方法である．上で述べた発声内容依存型の方法と同様に，時期的な変動や発声内容の影響を軽減するために，重み付きのケプストラムに変換する方法などが検討されている．2. で述べたように，音声スペクトルの個人差は，その発声内容に依存する部分が多く，長時間平均スペクトルではこのような特徴が平均化されて失われてしまうため，認識性能に限界がある．

5.2　スペクトルのベクトル量子化歪による方法

このような長時間平均スペクトルの問題に対処するために考え出されたのが，ベクトル量子化[1)]による歪を用いる方法である．複数のパラメータをそれぞれ独立に，有限の値によって量子化するスカラー量子化に対して，ベクトル量子化においては，パラメータの組み合わせをセットして，あらかじめ用意された有限のセッ



第6図　ケプストラムの静的および動的特徴からなる符号帳によるベクトル量子化に基づく発声内容独立型話者識別システムの構成

トの組（符号帳）によって量子化を行なう．すなわち，入力パラメータセットを符号帳の中の各セットと比較し，最も距離(量子化歪）の小さいセットの符号（インデックス）によって表現する．

この方法では，各登録話者について，任意の文章を発声したときの短時間スペクトルの集合をクラスタ化し，各クラスタの重心（セントロイド）を符号帳の要素として蓄える．未知の入力音声に対しては，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を求める．話者識別の場合は，最も平均量子化歪の小さい符号帳の話者を選択し，話者照合の場合は，平均量子化歪をしきい値と比較して，受理するか棄却するかの判定を行なう．

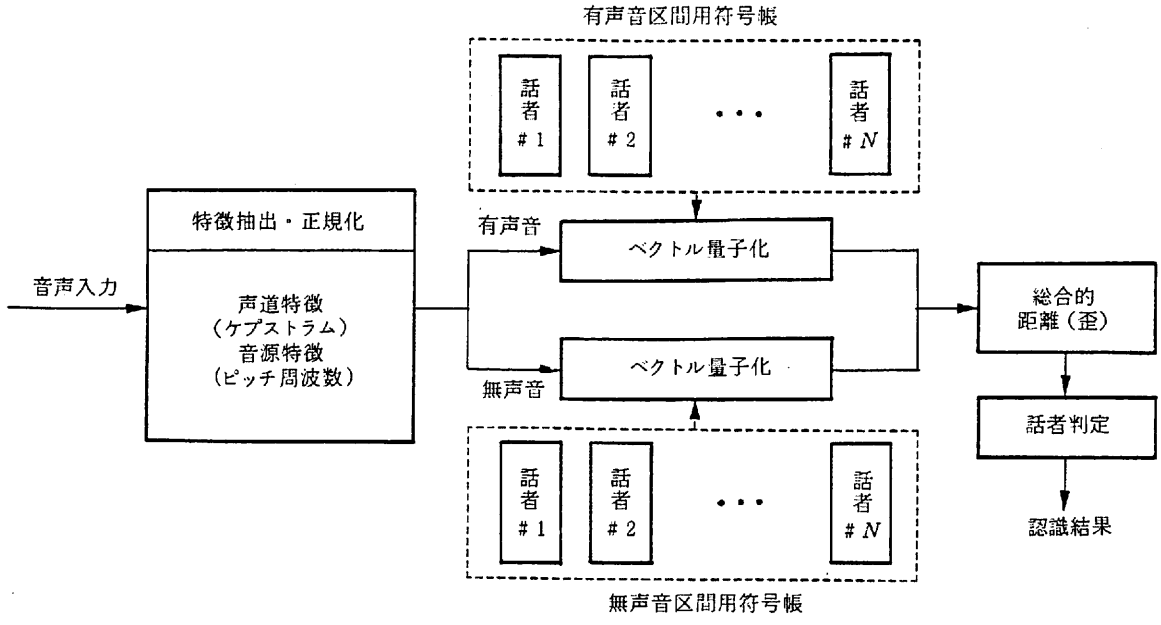
Soongらは，短時間スペクトルをケプストラムとその1次の多項式展開係数すなわち線形回帰係数（Δケプストラム）を用いて表わし，第6図に示す話者識別系を構成して実験を行なった[11]．量子化歪（距離）の計算においては，各パラメータについて，その分散の逆数で重み付けを行なっている．音声サンプルとしては，英語の0から9の10数字の電話音声を，男女各5名が2か月にわたる五つの時期に各単語20回ずつ発声したサンプル（各話者計200サンプル）を用いた．前半の100サンプルを各話者の符号帳作成に用い，残りの100サンプルをテストに用いた．その結果，ケプストラムのみを用いる場合に比べて，動的特徴であるΔケプストラムを組み合わせることによって大きな改善が得られ，任意の1数字を識別に用いた場合の平均識別誤り率は7.2％であった．

最近，符号帳の要素を，ある瞬間のパラメータのセットであるベクトルから，音素程度の区間におけるベクトルをまとめたセグメント（マトリックス）に拡張し，各セグメントをHMMで表現する試みが行なわれて，認識性能の向上が確かめられている[12),13)]．

入力音声に対して，話者ごとの符号帳を用いたときの平均量子化歪ではなく，すべての話者に共通の符号帳を用いてベクトル量子化したときの，符号帳の各要素が選ばれる頻度分布の形の個人差を，話者性の尺度に用いる方法についても検討が行なわれている[14)]．

5.3　スペクトルとピッチのベクトル量子化歪による方法

上のSoongらの方法では，スペクトル包絡に関する情報のみを用いているが，これにピッチに関する情報をうまく組み合わせることによって，認識性能を向上させる研究が，松井らによって行なわれている[15)]．3.で述べたように，ピッチは話者を判定するうえで有用な情報であるが，正確な抽出が難しく，無声音区間のように元々ピッチが存在しない音声区間があるなどの問題がある．そこで，ピッチが信頼度高く抽出される区間と，そうでない区間を分け，改めて前者を有声音区間，後者を無声音区間と定義して，第7図に示すようにそれらに対応す



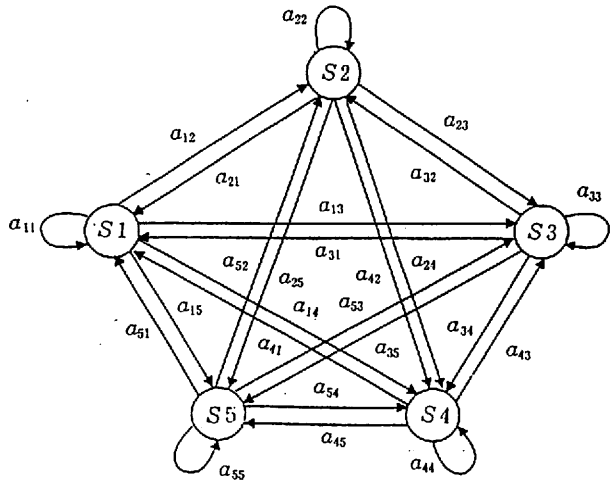
第7図　ピッチを含む静的および動的特徴の符号帳によるベクトル量子化に基づく発声内容独立型話者識別システムの構成

る別々の符号帳を用意する．有声音区間の特徴ベクトルの要素は，ピッチ，1〜16次のケプストラム，および両者の線形回帰係数であり，無声音区間の特徴ベクトルの要素は，これらからピッチ関連量を除いたものである．未知入力音声に対しては，有声音/無声音の判定を行なった後，それぞれに対応する符号帳を用いてベクトル量子化を行ない，最後に両者の歪を加え合わせて最終的な判定を行なう．

この方法ではさらに，パラメータの話者内変動を抑圧し，話者間変動を強調する新しいパラメータの変換方法と，符号帳と入力音声の特徴ベクトルの分布の重なりを考慮した距離尺度を導入することによって，認識性能の向上を図っている．男性9名が3年間にわたって発声した音声を用い，この内1時期の単語や文章を発声した音声から符号帳を作成して，異なる時期に発声した，学習音声とは異なる内容の30秒程度の文章音声を認識に用いる実験を行なった．その結果，話者識別で99.0％，話者照合で98.7％の平均認識率が得られ，ケプストラムへのピッチの組み合わせの効果とともに，新しいパラメータ変換方法および距離尺度の効果が確認された．

5.4　HMMによる方法

各話者の任意の発声内容の音声を，第8図に示すようなエルゴディックなHMMで表現し，このモデルから入力音声のスペクトル系列が出現する尤度を計算して話者を認識する方法が研究されている[16),17)]．エルゴディックなHMMの場合は，任意の状態からスタートして，任意の状態に遷移し，任意の状態を終端とすることができる．HMMの学習は，各話者ごとに学習用の文章や



第8図　エルゴディックなHMMの例

多数の単語を発声した音声を用いて自動的に行なうことができるが，第8図のHMMを用いたSavicらの実験では，五つの状態がほぼ，鼻音，摩擦音，破裂音，および2種類の有声音に対応するように設定されることが確かめられている．状態間の遷移は発声内容によって変化するので，発声内容独立の場合は遷移確率は評価に含めず，出力確率のみを用いて計算した尤度，あるいは類似度によって話者の判定を行なう場合が多い．

6. むすび

話者認識の基本的技術と，発声内容依存型と独立型に分けた最近の技術の進歩について解説した．この内，発声内容独立型の方法のほとんどは，入力音声と標準パタ

ーンあるいはモデルとの時間整合を必要としない簡易型の方法として，発声内容依存型の話者認識でも用いることができる[18]．

近年，通信網におけるセキュリティや，情報管理の厳格化などに対する関心が高まるにつれて，音声を個人識別に使いたいという要求が高まりつつあり，主として米国において，種々の会社 (Voxtron, Inc., ECCO Industries, Alpha Microsystems, Voice Prints, Inc. など) が話者照合の商品の開発・販売の実績を上げつつある．今後の研究開発により，一層の性能向上，性能の安定化なども期待される．野田らによる，男性523名が発声した単語および単独母音を用いた大規模話者認識実験の結果，発声内容依存型で行なえば，現在の自動話者認識技術は，法科学分野の話者照合や容疑者の絞り込みの手段として実用可能であるとの結論が得られている[19]．

話者認識が個人識別の手段として広く受け入れられるか否かは，その基本的性能のみならず，ユーザフレンドリネスに依存する面も大きい．たとえば，本人の音声が棄却されたときにオペレータなどによるバックアップが準備されているか，風邪をひいたりして声が大きく変わってしまったときはどうするか，システムからどのような指示で全体の流れを制御するか，などは実システムの使い勝手を決定する重要な要素である．これらの面を考慮しながら，ID番号など音声以外の個人識別手段と組み合わせたシステムを構成すれば，役に立つ話者認識システムが実現できる時代が到来しつつあるように思われる．

(1990年11月29日受付)

参 考 文 献

1) 古井：ディジタル音声処理，東海大学出版会 (1985)
2) 古井：話者識別技術；計測と制御，Vol. 25, No. 8, pp. 688～693 (1986)
3) S. Furui : Research on Individuality Features in Speech Waves and Automatic Speaker Recognition Techniques ; Speech Communication, Vol. 5, No. 2, pp. 183～197 (1986)
4) J. M. Naik : Speaker Verification : A Tutorial ; IEEE Communications Magazine, January, Vol. 28, No. 1, pp. 42～48 (1990)
5) S. Furui : Speaker-Dependent-Feature Extraction, Recognition and Processing Techniques ; Proc. Tutorial and Research Workshop on Speaker Characterization in Speech Technology, Edinburgh, Scottland, pp. 10～27 (1990)
6) 古井：音声の個人性パラメータの時期的変動と話者認識；信学論，Vol. J 57-A, No. 12, pp. 880～887 (1974)
7) 古井：ケプストラムの統計的特徴による話者認識；信学論，Vol. J 65-A No. 2, pp. 183～190 (1982)
8) S. Furui : Cepstral Analysis Technique for Automatic Speaker Verification ; IEEE Trans. ASSP, Vol. 29, No. 2, pp. 254～272 (1981)
9) Y. -C. Zheng and B. -Z. Yuan : Text-Dependent Speaker Identification Using Circular Hidden Markov Models ; Proc. ICASSP 88, New York. S 13.3 (1988)
10) J. M. Naik, L. P. Netsch and G. R. Doddington : Speaker Verification over Long Distance Telephone Lines ; Proc. ICASSP 89, Glasgow, Scottland, S 10 b. 3 (1989)
11) F. K. Soong and A. E. Rosenberg : On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition ; IEEE Trans. ASSP, Vol. 36, No. 6, pp. 871～879 (1988)
12) A. E. Rosenberg, C. -H. Lee and F. K. Soong : Sub-Word Unit Talker Verification Using Hidden Markov Models ; Proc. ICASSP 90, S 5.3 (1990)
13) B. -H. Juang and F. K. Soong : Speaker Recognition Based on Source Coding Approaches ; Proc. ICASSP 90, S 5.4 (1990)
14) 白井，間野，石毛：ベクトル量子化スペクトルの頻度分布による話者識別；信学論，Vol. J 70-D, No. 6, pp. 1181～1188 (1987)
15) T. Matsui and S. Furui : Text-Independent Speaker Recognition Using Vocal Tract and Pitch Information ; Trans. ICSLP 90, Kobe, 5. 3 (1990)
16) A. B. Poritz : Linear Predictive Hidden Markov Models ; Proc. ICASSP 82, Paris, S 11.5 (1982)
17) M. Savic and S. K. Gupta : Variable Parameter Speaker Verification System Based on Hidden Markov Modeling ; Proc. ICASSP 90, S 5.7 (1990)
18) D. K. Burton : Text-Dependent Speaker Verification Using Vector Quantization Source Coding ; IEEE Trans. ASSP, Vol. 35, No. 2, pp. 133～143 (1987)
19) 野田，長内：話者認識における話者数と認識率の信頼性との関係：信学論，Vol. J 73-A, No. 4, pp. 717～724 (1990)