論文 / 著書情報
Article / Book Information

論題	音声による話者の自動認識技術の動向
著者	古井貞熙
Author	SADAOKI FURUI
掲載誌/書名	ハイ・セキュリティ情報ネットワーク研究会，，，
発行日	1984,

音声による話者の自動認識技術の動向

Technical Trends of Automatic Speaker Recognition from Their Voices

古井 貞熙
Sadaoki FURUI

日本電信電話公社 武蔵野電気通信研究所
Musashino Electrical Communication Laboratory,N.T.T.

1．話者認識の原理

音声波に含まれる第一義的な情報は、言葉の意味内容に関連した音韻性情報であるが、第二義的情報として重要なものに、誰が話しているかという個人性情報がある。音声波から音韻性情報を抽出して言葉を判定することを音声認識、個人性情報を抽出して誰の声かを判定することを話者認識という[1～5]。音声が初めて犯人の割り出しの手掛かりとして用いられたのは、1660年の英国チャールズ一世の死に関する裁判の時であるといわれている[6]。その後時代を経て、電話が距離を克服し、録音手段が時を克服するに伴って、個人性の分析に対する関心が高まり、C．A．Lindbergh氏子息誘拐事件をきっかけとして、1937年に初めて音声の個人性に科学的なメスが加えられた。そして、1945年に米国のベル研究所のR．K．Potterによってサウンドスペクトログラムが発明され、いわゆる声紋(voice print)が自動的に描かれるようになった。サウンドスペクトログラムは、細かい時間毎に音声を周波数分析して、音声のスペクトルが時間とともにどのように変化しているかを図にあらわしたものである。例を図1に示す[7]。サウンドスペクトログラムに対して声紋の言葉を初めて用いて、これによる話者認識の可能性を初めて述べたのは、ベル研究所のL．G．Kerstaで、1962年である[8]。1966年の米国での裁判で、初めてこれが証拠として用いられている。

話者認識の研究には、サウンドスペクトログラムを人間が目で見て判定する手掛かりに関する研究[9]のほか、人間が耳で誰の声かを判定するときに、音声のどのような特徴を用いているかという研究[10]もあるが、近年電子計算機等の発達により、自動的な認識方法の研究が急速に発展して来ている。ここでは以後話者認識という言葉を、自動認識に限定して用いることにする。話者の自動認識の研究は、犯罪捜査のみならず、情報化社会の発達に伴って、個人のプライバシーを守り、バンキング等におけるセキュリティーに役立つ技術として、需要が増してくるものと考えられる。音声を用いて自動的に話者が認識できれば、カードや鍵を用いるよりも簡単であり、両手がふさがっていても使え、盗まれる心配がないなど、メリットが多い。一方欠点としては、本人の声でも発声のたび毎に変化し、回線雑音、マイクロホンの特性等によっても変化するため、この変動を許容できるようにすると、似た声で誤動作する可能性を生ずる点がある。このため実用的には、できるだけ他人が真似しにくく、伝送系の影響等を受けにくい特徴を用い、必要な場合は音声以外の簡易的な手段と組み合わせて用いるという方法が考えられる。

話者認識と音声認識の研究は、本来表裏一体をなすべきものであるが、音韻性と個人性情報は、複雑にからみあった形で音声波（スペクトル）に含まれており、この両者を明確に分離して抽出する技術はまだ確立されていない。このため、音声認識では個人を固定し、話者認識では発声内容を固定することによって、高い認識率を得ているのが現状である。

2．話者認識に用いる特徴

音声に含まれる個人性情報には、声道長、声道形状、声帯等の先天的な発声器官の個人差に起因するものと、語彙、なまり、速度、テンポ、アクセント等の後天的な発話習慣（話し方のくせ）に起因するものとがある。前者は、ホルマント周波数の高低、帯域幅の大小、平均基本周波数、スペクトル概形の傾斜等として現われ、後者は語彙を除き、基本周波数やホルマント周波数の時間パターン、単語の時間長等として現われるが、両者を明確に分離して抽出するのは難しい。このため、両者の特徴が同時に含まれる特徴パラメータが用いられることが多い。

実際の話者認識の方法としては、スペクトル包絡、基本周波数、パワー等の時間パターンをそのまま用いるものと、その時間パターンから抽出した統計的特徴を用いるものとがある。前者の方法は、単語音声認識でしばしば用いられる方法と類似しているが、安定した個人の特徴を強調

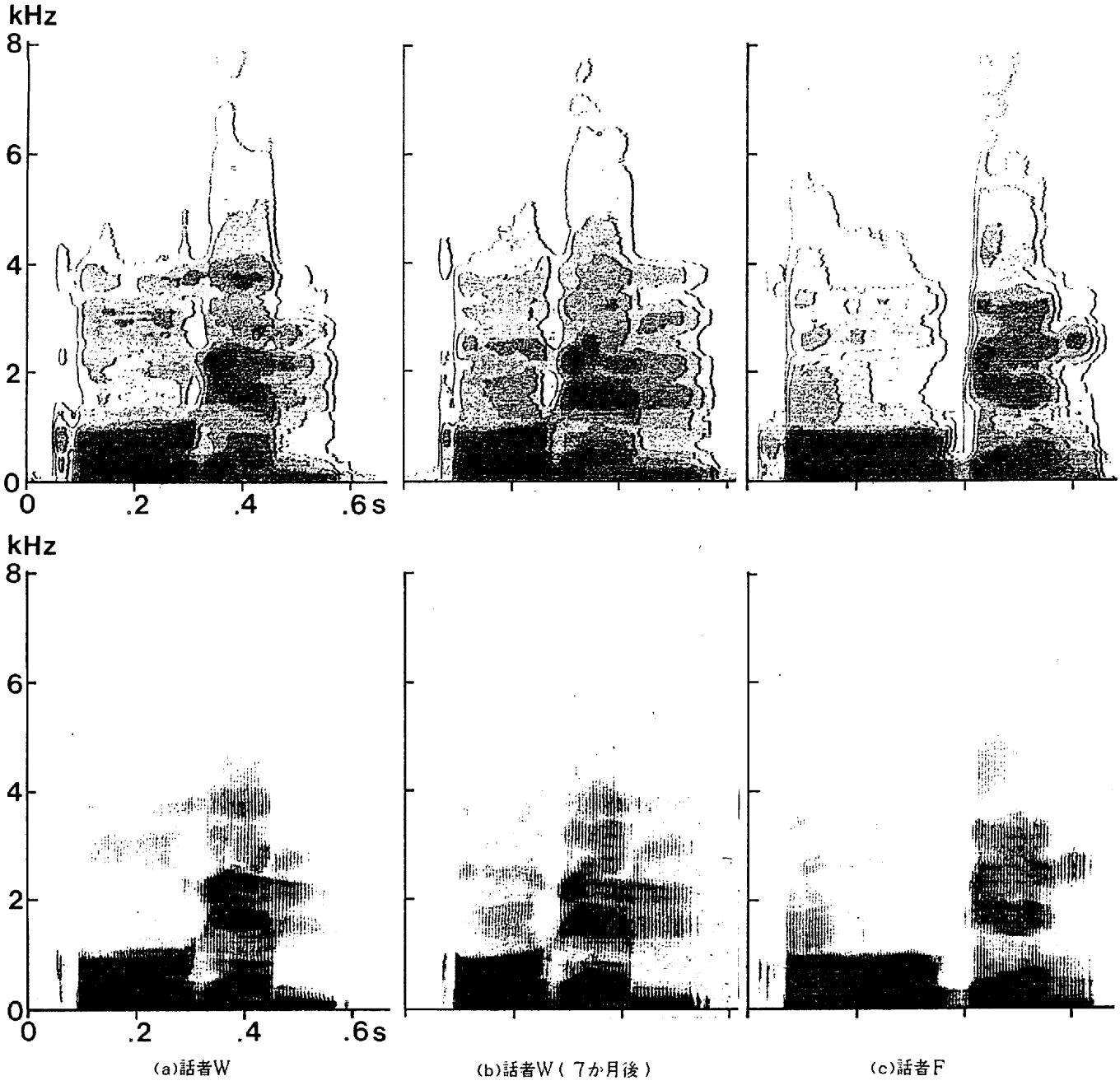

図1.「高原(こうげん)」と発声した音声のサウンドスペクトログラムの例 （上:等高線表示、下:濃淡表示）

するために、話者認識に特有の配慮が必要である。声帯模写では、後天的な発話習慣に関連する基本周波数やパワーの時間パターンを真似しているケースが多い。自動的な話者認識においても、これらの音源の特徴を基礎とする方法は作り声に弱く、声道の特徴に関連したスペクトル包絡情報を組み合わせた方が、作り声に対して強くなることが示されている[11]。

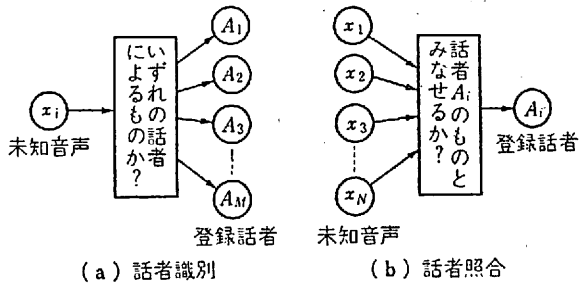

図2.話者識別と話者照合

3．話者認識の分類

話者認識は、話者識別（speaker identification）と話者照合（speaker verification）に分けることができる。図2に示すように、話者識別とは入力音声が、あらかじめ登録してあ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が、その発声者が名乗った本人の音声であるか否かを、標準パターンとの類似の度合いによって判定するものである。

話者照合の例としては、電話によるバンキングサービス（送金、振替、残高照会、振り込み通知等）、個人情報へのアクセス、機密保管場所等への入室者のチェック等において、本人か否かの鍵として音声を用いる場合がある。話者識別の例としては、犯罪捜査において、犯行時に録音

された音声が複数容疑者の内の誰の声であるかを判定する場合がある。ただしこの場合は、容疑者の中に真犯人が含まれていないこともありうるので、実際には話者照合と話者識別の組み合わせの形をとることになる。

話者認識の方法には、認識に用いる言葉をあらかじめ決めておく方法（発声内容依存形、text-dependent）と、どんな言葉を発声してもよい方法（発声内容独立形、text-independent）とがある。発声内容依存形の方法の中には、鼻音などの特定の音素の特徴に着目する方法もあるが、現在研究されている方法は、各話者がそれぞれに決めたキーワード（暗証番号、名前等）を一貫して発声すればよいものが多い。後者の場合は、実用的にはそのキーワードの違いも手掛かりとして用いることができるので、それだけ高い認識率を得るのが容易となる。研究段階の評価実験では、すべての話者に共通のキーワードを用いることが多い。

話者認識の応用の多くは、このように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、犯罪捜査の場合等では、必ずしも同じ言葉どうしを比較することができないケースもある。このような目的のためには、発声内容独立形の方法を開発する必要がある。話者認識の難しさは、発声者が自分の声が正しく認識されることを望んで好意的に発声するか否か、によって大きく異なる。一般に話者照合の応用の多くの場合は、発声者が発声速度や話し方をことさら変えずに、好意的に発声することが期待できるが、犯罪捜査のような場合はこれと全く異なる。このため後者の例の場合は、話者の認識が非常に難しい場合が多い。

4．話者認識系の構成

話者認識系の基本的構成を図３に示す。音声波から特徴抽出をしたのち、あらかじめ蓄えられている各登録話者の標準パターンとの距離あるいは類似度を調べ、その度合いにより認識の判定を行う。話者識別の場合は、入力音声との距離が最も小さい標準パターンの話者によるものと判定し、話者照合の場合は、入力音声と標準パターンの距離と、あらかじめ定めた判定基準（しきい値）との大小関係によって判定を行う。

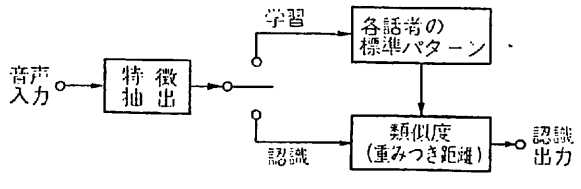

図3. 話者認識系の基本的構成

話者照合システムの評価には、精神物理学におけるＲＯＣ曲線（receiver-operating-characteristic curve）を用いることができる。話者照合においては、未知音声が真に本人のものである状態（ｓ）と、そうでない状態（ｎ）、それを本人の音声と受理する場合（Ｓ）と、そうでないと棄却する場合（Ｎ）の四つの組み合わせがあり、表１に示す４種類の確率が定義できる。Ｐ（Ｓ｜ｓ）は正しい受理、Ｐ（Ｓ｜ｎ）は他人（詐称者imposter）の音声を受理する誤り（ＦＡ；false acceptance）、Ｐ（Ｎ｜ｓ）は本人の音声を棄却する誤り（ＦＲ；false rejection）、Ｐ（Ｎ｜ｎ）は正しい棄却の確率である。このとき、

$$P(S|s)+P(N|s)=1$$
$$P(S|n)+P(N|n)=1$$

表１　話者照合の四つの場合

		状態	
		s（本人）	n（他人）
判定	S（受理）	Ｐ（Ｓ｜ｓ）	Ｐ（Ｓ｜ｎ）
	N（棄却）	Ｐ（Ｎ｜ｓ）	Ｐ（Ｎ｜ｎ）

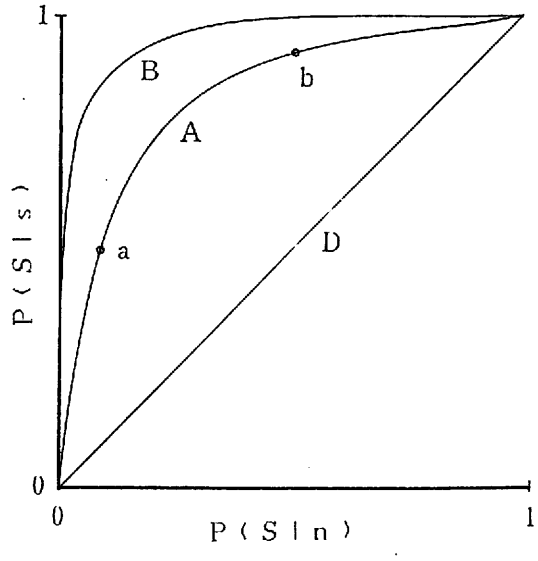

図4. 話者照合におけるROC曲線

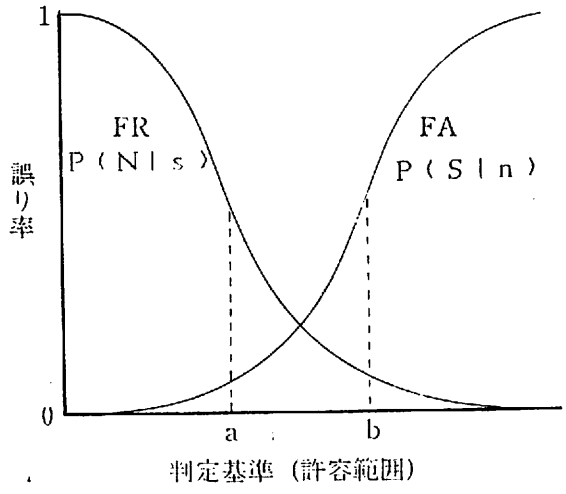

図5. 話者照合の判定基準と誤り率の関係

の関係があるので、P(SIs)とP(SIn)の二つを用いれば、その認識系を評価することができる。この両者を縦軸と横軸にとり、本人の音声と判断する判定基準(許容範囲)を変えると、各認識系について図4に示すようなROC曲線を得ることができる。この図で方法Bは一貫して方法Aより優れ、Dは認識能力がない場合に相当する。

一方、判定基準と二種の誤り率の関係は図5のようになり、図4と図5でaの点は許容範囲が狭く厳しい場合、bの点は許容範囲が広く緩い場合に対応する。実際の許容範囲は、判定誤りが生ずる影響の大きさに応じて設定するのがよく、事前確率及びコストとROC曲線の接線勾配から決定することができる。実験では二種の誤り率が等しくなる値(図中のc)に判定基準を事後的に設定し、この時の誤り率によって評価することが多い。

特徴パラメータの有効性の評価には、次のような方法を用いることができる。a)各パラメータの組み合わせによる認識実験を行う。b)パラメータ毎に、F比(話者間/話者内分散比)を計算する。c)F比を多次元に拡張したdivergenceを用いる。d)認識誤り率に基づくknock out法を用いる。効率的に情報圧縮しパラメータ数を減らす方法としては、判別分析によりF比を最大にする空間へ射影する方法がしばしば用いられる。

5. 認識誤り率と話者数の関係

登録話者N人の集合をZ_N、音声データ(n次元特徴ベクトル)を$\mathbf{X}'=(x_1, x_2, \cdots, x_n)$、話者i($i \in Z_N$)の$\mathbf{X}$の確率密度関数を$P_i(\mathbf{X})$とすると、$Z_N$に含まれる話者の$\mathbf{X}$の出現確率密度関数は、

$$P_Z(X)=\underset{i\in Z_N}{E}[P_i(X)]$$
$$=\sum_i P_i(X)P_r[i],\ i\in Z_N$$

となる。Pr[i]は、話者iの出現確率である[12]。

話者照合の場合、話者i本人の音声として受理すべき$\mathbf{X}$の領域は、

$$R_{Vi}=\{X|P_i(X)>C_iP_Z(X)\}$$

である。ここでC_iは前述の二種類の誤り率を適当なバランスに保つように選ぶ。今、Z_Nが話者のランダム集合で$Pr[i]=1/N$なら、Nが大きくなるに従って、$P_Z(\mathbf{X})$はZ_Nに無関係な密度関数に近づく。従って、二種類の誤り率は、Nが大きくなってもその影響を受けない。なお、通常$P_Z(\mathbf{X})$を正確に推定することは困難なので、一定値を仮定し、

$$R_{Vi}=\{X|P_i(X)>K_i\}$$

とすることが多い。

一方話者識別の場合は、話者iの音声と判定する領域は、

$$R_{Ii}=\{X|P_i(X)>P_j(X),\ \forall j\neq i\}$$

である。このときの誤り率は、

$$P_{Ei}=1-\prod_{\substack{k=1\\k\neq i}}^{N} P_r(P_i(X)>P_k(X))$$

であるから、Z_Nが話者のランダム集合なら、

$$\underset{Z_N}{E}[P_{Ei}]=1-\underset{Z_N}{E}\{\prod_{\substack{k=1\\k\neq i}}^{N} P_r(P_i(X)>P_k(X))\}$$
$$=1-\underset{i}{E}\{\prod_{\substack{k=1\\k\neq i}}^{N} \underset{k}{E}[P_r(P_i(X)>P_k(X))]\}$$
$$=1-\underset{i}{E}\{P_{Ai}{}^{N-1}\}$$

と書ける。ここでP_{Ai}は、iと他の任意の話者との間で誤りを生じない確率の期待値である。従って、誤り率の期待値は、Nの増加とともに指数関数的に増大し、極限では1となる。

この結果は、無限個の分布は有限パラメータの空間では別々に分離していることはできないということから、当然ともいえる。すなわち、話者数が大きくなれば、話者の集合中の二人あるいはそれ以上の話者のデータの分布が互いに非常に近い可能性が大きくなるからである。このため話者識別系の評価は、登録話者数の影響を十分考慮して行う必要がある。図6に、話者識別と話者照合における登録話者数と誤り率の関係について、単語音声スペクトルの統計的特徴を用いた認識系によって、実験的に調べた結果を示す[13]。

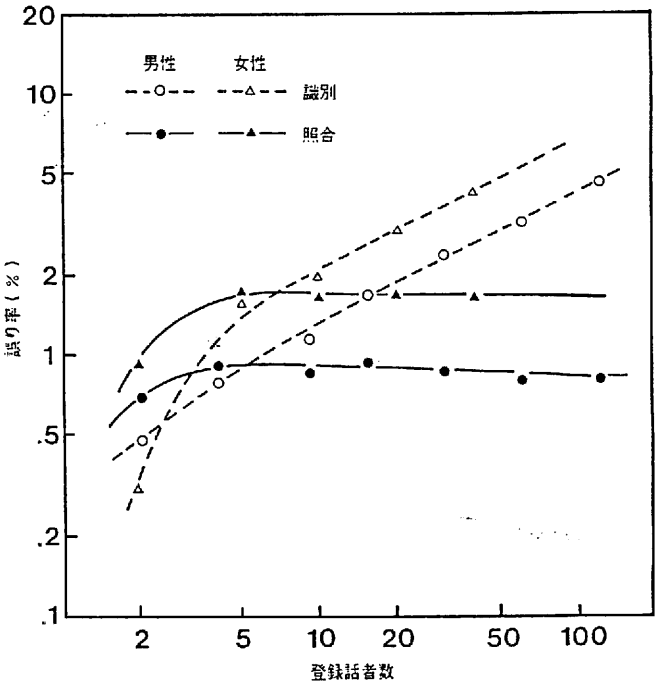

図6. 話者識別と話者照合における登録話者数と誤り率の関係

6. 特徴パラメータの長期間変動

話者認識系を構成するときに考慮すべきこととして、特徴パラメータの長期間にわたる変動の影響がある。即ち図1の例にも示したように、同じ人が同じ言葉を繰り返し発声しても、ある程度時期をおくと、発声の仕方が変わって音声のスペクトルが変化し、認識系の性能に影響を及ぼす。音声認識においてはこのようなことは通常問題にならないが、話者認識においては、標準パターン作成のための学習サンプルと未知入力音声が発声される時期に必ず隔た

りがあり、しかも個人性情報は音韻性情報に比べて細かい特徴であるため、この時期差による変動の影響を受けやすい。このように本質的に変動する性質を有する点が、指紋との音声の相違点である。

特徴パラメータの時期的変動の影響を軽減し、高い認識率を得るためには、次のような方法が有効であることが確認されている[7,14]。

（１）音声波にスペクトル等化処理を行う。即ち、単語又は短文章の時間平均スペクトルを１次系又は２次臨界制動系で近似し、音声をこの逆フィルタに通す。この処理は近似的に、音声中の声帯波情報を除去し、声道情報のみを保存することに相当している。

（２）長期間にわたる音声データにもとづく統計的評価により、安定した特徴パラメータを選択するとともに、異なる複数の単語等から抽出した特徴パラメータを組み合わせて用いる。

（３）長期間にわたる学習サンプルにもとづいて、標準パターンと距離尺度を定める。

（４）標準パターンを適切な周期で更新する。

図７は、単語音声スペクトルの統計的特徴を用いた話者認識系によって、スペクトル等化処理の有効性を実験的に調べた結果を示したものである。各話者の学習サンプルとして、１０日間程度の短期間の間の４日間に発声した１２サンプルを用い、未知入力音声までの時期差を２～３日から１０か月まで変えて、話者照合の誤り率の変化を調べた。その結果、逆フィルタを適用すれば６週間～１０か月後の音声に対する誤り率を１／３程度に低下させることができることが確認された[15]。

7．話者認識系の研究動向

話者認識システムはまだ実用化されるに至っていないが、米国Texas Instruments社における、男女計３００名の話者による長期間にわたるコンピュータ室の入室管理システムの試行実験[16,17]、ベル研究所における男女計１００名の電話音声による話者照合実験[18]など、大規模な実験も行われるようになってきた。基礎的な研究も国内外で着実に行われている。

この１０年間の主な研究例を表２に示す[4]。認識率の欄には報告されている値をそのまま載せたが、表中の各条件のほか、学習サンプルの収集期間、入力音声までの時期差等の条件も各実験で異なっているので、この数値だけで単純に優劣を比較することはできない。全般的には、発声内容独立形の方式はまだ基礎研究段階にあるが、発声内容依存形の話者照合は実用化に近づいていると言える。

8．発声内容依存形の認識系

上述のＴＩ社とベル研究所における実験で用いられている方法は、いずれもスペクトルパラメータの時系列を直接的に用いる発声内容依存形の方法である。ＴＩ社のシステムでは、計算センターの入口にブースが作られており、

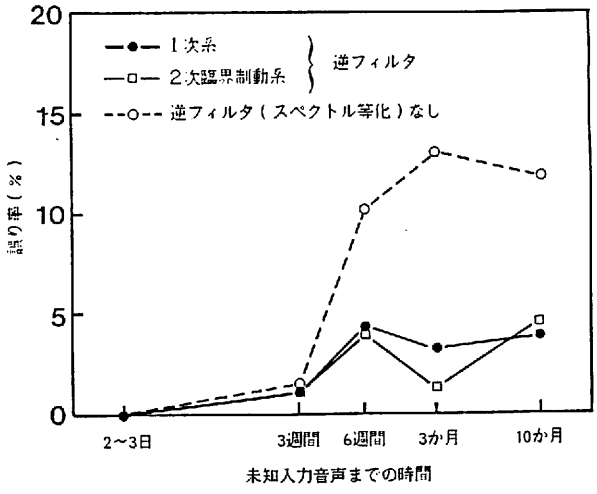

図7．スペクトル等化処理の有効性(学習サンプルを10日間程度の短期間に収集した場合の、10か月後までの音声の話者照合結果)

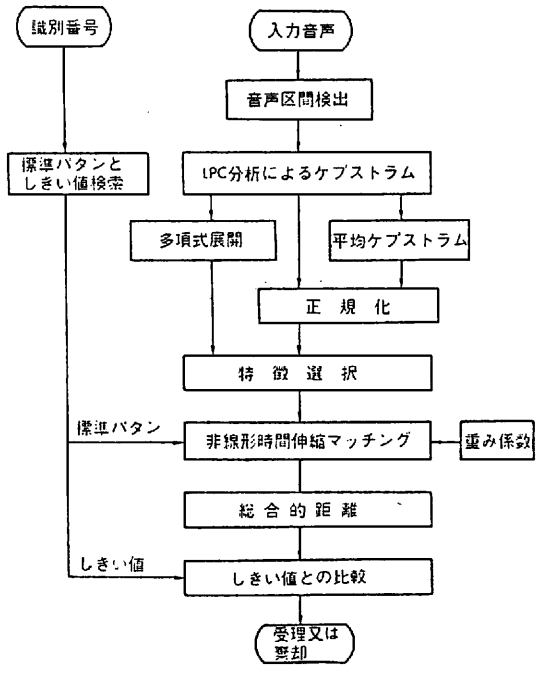

図8．短文章音声のケプストラムとその動的特徴を用いた話者照合のブロック図

そこに入って自分のＩＤカードを指定の場所に差し込んでから、合成音で指示された言葉をマイクロホンに向かって復唱する。その言葉の母音中央部のスペクトルマッチングによって、ＩＤカードの本人か否かの話者照合を行い、本人と認識されればロックがはずれるしくみになっている。実験では、マイクロホン入力による高品質音声が対象であるが、９９％の高い認識率が得られている。

ベル研究所における方法のブロック図を図８に示す。この方法では、単に時系列の時間伸縮マッチングを行うだけでなく、スペクトルの動的な特徴を積極的に抽出して用いている。音声サンプルとしては短文章音声を用い、１０ｍｓ毎に１０次までのＬＰＣケプストラムを抽出する。音声区間全体にわたって平均化したケプストラムの値を、各時点のケプストラムの値から減ずることにより、スペクトル等化処理を行う。これは、伝送歪やスペクトルの話者内

表 2　最近の話者認識の主な研究例

	研究機関	研究者（年代）	発声内容	特徴パラメータ・方法	識別(I)照合(V)	登録話者数（）：照合の詐称者数	認識率(%)
発声内容依存形	電電公社武蔵野通研	古井・板倉 (1973)	1〜4単語	対数断面積比と基本周波数の統計的特徴（平均値・標準偏差・相互相関係数）の重みつき距離	I	男 9	96.8〜99.1
					V	男 9 (111)	97.3〜99.3
		古井 (1982)	4単語	LPC ケプストラムと基本周波数の統計的特徴（平均値・標準偏差・相互相関係数・フーリエ展開係数）の重みつき距離	I	男 9	100
					V	男 9 (111)	99.9
	日立中研	市川・中島・中田 (1979)	1単語	Cosh 尺度を用いた DP マッチングと平均基本周波数，帯域選択自己平均逆フィルタ法による伝送ひずみの除去，電話音声	I	男 5	91
	Bell Laboratories (米国)	Rosenberg (1976)	短文章（2秒）	基本周波数とパワーの DP マッチングの後，種々の距離尺度を適用，電話音声	V	男・女 100 (同)	適応後 FR 4% FA 5%
		Furui (1981)	同上	LPC ケプストラムとその多項式展開係数の DP マッチング，スペクトル等化処理	V	男 10 (49) 女 10 (49) 男 21 (75)	99.8 } 電話 99.6 } 音声 99.2
		Atal (1974)	短文章（0.5 秒〜）	種々の線形予測パラメータを直交線形変換し有効性を評価，LPC ケプストラムが最も有効	V	男 10 (9)	98 (1秒)
					I	男 10	98 (0.5 秒)
	Texas Instruments (米国)	Doddington (1976)	CVC 形 1〜4 単語	母音中央部 (100 ms) のスペクトルマッチング，スペクトル等化処理，計算機室の入室管理に適用	V	男 180 (同)	4 単語 FR 0.3% FA 1%
発声内容独立形	電電公社武蔵野通研	古井・板倉・斎藤 (1972)	短文章（10〜30 秒）	長時間平均スペクトルのケプストラムの重みつき距離，時期差の効果の分析	I	男 9	90.8
					V	男 9 (26)	93.5
	東北大	松本・曽根・二村 (1977)	短文章（0.5 秒〜）	ケプストラムと基本周波数の断片的正準判別分析	I	男 10	96
	Bell Laboratories (米国)	Sambur (1976)	短文章（2秒）	種々の線形予測パラメータを直交線形変換し，固有値の小さい成分を使用，音韻性と個人性を分離	I	男 21	94
	SCRL (米国)	Markel Davis (1979)	会話音声（39 秒〜）	PARCOR 係数と基本周波数の統計量（平均値・標準偏差）の重みつき距離	V	男 11 女 6 (同)	96
					I	男 11 女 6	98
	Philips 研究所 (西独)	Bunge (1977)	短文章（11 秒）	長時間平均スペクトルに対して種々の距離尺度を適用	V I	男 41 女 9 (同)	98〜99
	ITT (米国)	Li Wrench (1983)	会話音声（3〜10 秒）	LPC 分析した各話者の音声をそれぞれ多次元空間内の 40 点にクラスタ化して登録．未知音声フレームとこれらとの距離の統計的分布を使用	I	男 11	79〜96

SCRL : Speech Communications Research Laboratory　FR : 本人棄却率　FA : 詐称者受理率

変動に対処するための処理である。一方、各ケプストラムの時系列を１０ｍｓ毎に９０ｍｓの区間を用いて直交多項式展開し、その１次と２次の係数の時系列を得る。これらの展開係数と、平均値を減じて正規化されたケプストラムのうち、話者を分離する上で有効性の高い１８種類の成分の時系列を取り出し、標準パターン時系列との非線形時間伸縮マッチングを行う。最適なマッチングが行われたときの標準パターンと入力音声の距離を、あらかじめ学習サンプルを用いて各話者毎に設定してあるしきい値と比較して、話者照合の判定を行う。話者照合の判定のしきい値と各話者の標準パターンは、２週間に１回程度の間隔で定期的に更新している。この方法によれば、学習サンプルと未知入力音声（いずれも電話音声）が、ＡＤＰＣＭ、ＬＰＣボコーダ等の異なる伝送系を通った場合でも、高い認識率が得られることが確認されている。

統計的特徴を用いる発声内容依存形の代表的方法に、単語音声スペクトルの統計量を用いる方法がある[19]。この方法の構成は図９に示す通りで、各話者はあらかじめ定められた四種類の単語を区切って発声する。音声波は、各単語区間での平均相関係数を用いた１次系の逆フィルタによるスペクトル等化が施されたのち、ＬＰＣ分析により、（ＬＰＣ）ケプストラムと基本周波数の１０ｍｓ毎の時系列に変換される。各単語の有声音区間におけるこれらのパラメータの平均値、標準偏差、フーリエ展開係数、及び相互相関係数を計算し、これらの統計量のうち話者の分離に有効性の高い成分を選択する。これらの特徴量を用いて、話者識別と話者照合を行う。この方式は、マイクロホン音声と電話音声を用いた実験により、高い認識率の得られる

ことが確認されている。実際の交換機を通った男女計３４名の電話音声を用いた約６か月にわたるオンライン話者照合実験の結果では、実験の初期段階においては各話者の標準パターンが安定していないので誤認識が比較的多いが、実験回数とともに次第に認識率が向上し、安定した段階では９７～９８％の認識率が得られている。

９．発声内容独立形の話者認識系

発声内容独立形の方法には、図１０に示す二つの種類がある。（ａ）の方法は、声帯波スペクトルに近似的に対応していると考えられる長時間平均スペクトルを用いる方法に代表されるように、音声波から短時間毎に抽出したパラメータを音声区間全体にわたって平均化することによって、発声内容の影響を除去し、この平均化された特徴によって認識を行うものである。一方（ｂ）の方法は、各話者毎に種々の音素（音韻）に対応した標準パターンを用意しておき、音声波から短時間毎に抽出したパラメータとこれらのすべての標準パターンとの距離を計算する。このうち最も小さな距離を有する音素がその瞬間に発声されているものと見なし、この距離を全音声区間について平均化した値により認識を行う。一般に（ａ）の方法よりも（ｂ）の方法の方が、同じ認識率を得るのに短い音声ですむと言われているが、いずれの方法でもまだ十分な性能は得られていない。

（ａ）の方法の具体例としては、１０～３０秒長の長時間平均スペクトルのケプストラムを用いる方法がある[20]。この方法には、ケプストラムを用いてスペクトルパターンを展開することにより、長期間にわたって比較的安定した特性を選択的に重視することができるという特徴がある。図１１に男性話者２名の約１０秒長の短文章音声の長時間平均スペクトルの測定例を示す。２ｋＨｚより上の周波数領域の個人差が比較的大きい。同じく（ａ）に属す方法に、ＬＰＣ分析パラメータを直交変換した特徴パラメータの平均値と分散を用いるものがある[21]。この方法では、ＬＰＣ分析パラメータの共分散行列の、比較的大きい固有値を有する固有ベクトルへの射影は音韻性に、比較的小さい固有値を有する固有ベクトルへの射影は個人性に関与するという仮定に立ち、後者のみを話者認識に用いている。

（ｂ）の方法による代表的試みとしては、ケプストラムを直交変換したパラメータを用いる方法[22]や、クラスタ化した標準パターンを用いる方法[23]などがある。後者の方法の概要は次の通りである。まず、各話者毎に１００秒の会話音声を学習サンプルとして、短時間毎にＬＰＣ分析し、そのスペクトルパターンの集合をベクトル量子化の手法によりクラスタ化して、代表的な４０種類のパターンを用意する。未知音声の短時間毎のスペクトルと、各登録話者の４０パターンとの最小距離を計算し、未知音声全体におけるその分布を各登録話者対応に求める。この距離の値は図１２に示すように、標準パターンと未知音声の話者が同じときに比較的小さな値を示すので、この分布が最も

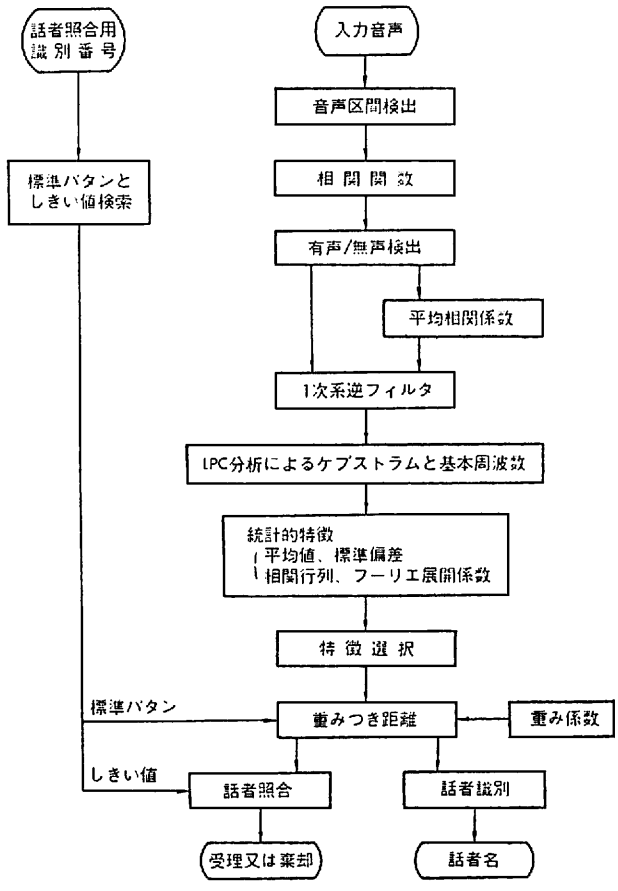

図9.単語音声のケプストラムの統計量を用いた話者識別と話者照合のブロック図

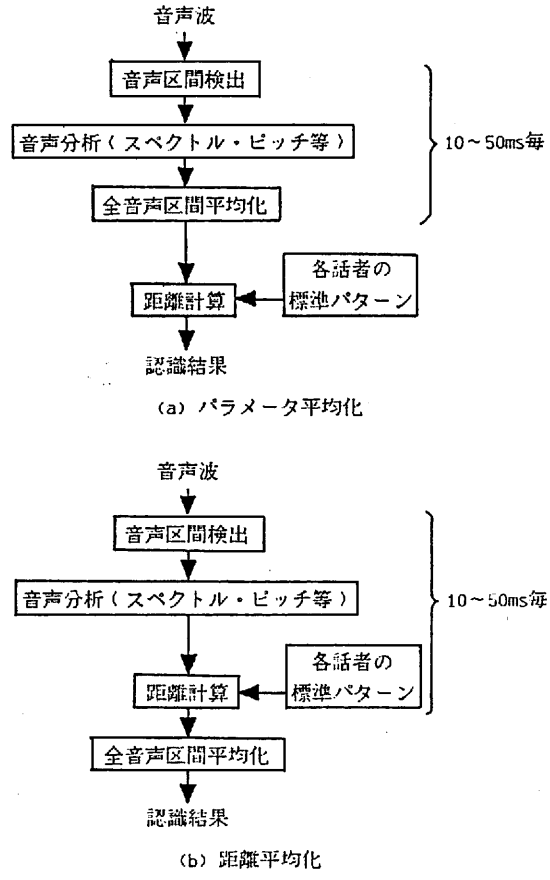

図10.発声内容独立形の話者認識系の構成

小さい値に片寄っている登録話者を選んで、その話者による音声であると判定する。発声内容を限定しない１０、５、３秒の未知音声に対して、男性１１名の話者識別でそれぞれ９６％、８７％、７９％の認識率が得られており、学習音声と未知音声に伝送路の違いがある場合についても検討がおこなわれている。

１０．" Sheep and Goats" 現象

多数の話者を用いた実験における話者毎の誤り率を調べてみると、話者によって片寄りを示すことが多い。これは音声認識の場合でもよく見られる現象であり、" Sheep and Goats" 現象と呼ばれる。誤り率の低い大部分の話者を" Sheep" 、人数的にはわずかな割合いではあるが誤り率の高い話者を" Goats" と呼び、システムの性能は主としてこの少数のGoatsによって決まる。図１３は、上述のＴＩ社の話者照合システムの累積誤り率と、その近似直線を表わしたものである[17]。平均誤り率（ＦＲ）は０．９％で、この点が近似直線の屈曲点にもあたっているので、この点をSheepとGoatsの境界とすると、全話者の７５％がSheepとなる。次の屈曲点は２．５％の点にあり、これより誤りの大きい話者を" Super Goats" と呼ぶと、この割合いは５％であるが、これらの話者の中には誤り率のかなり大きい話者も含まれ、これらへの対処策が実用上重要な課題である。

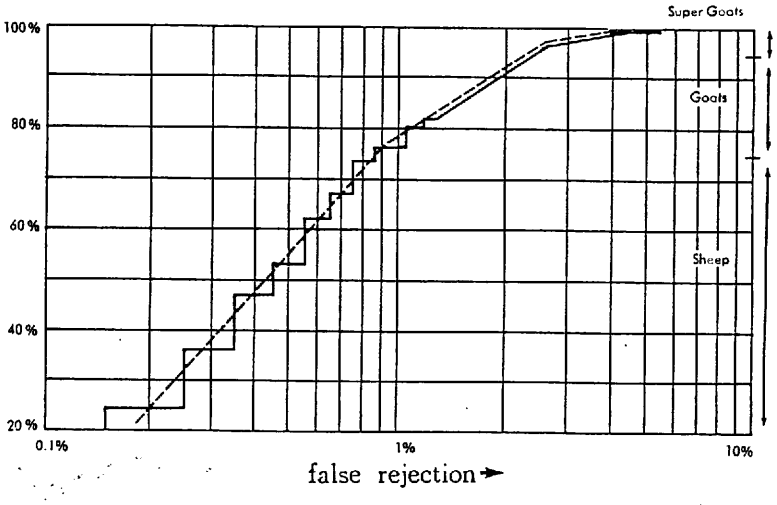

図13. 話者照合システムにおける登録話者 に関する累積誤り率の例

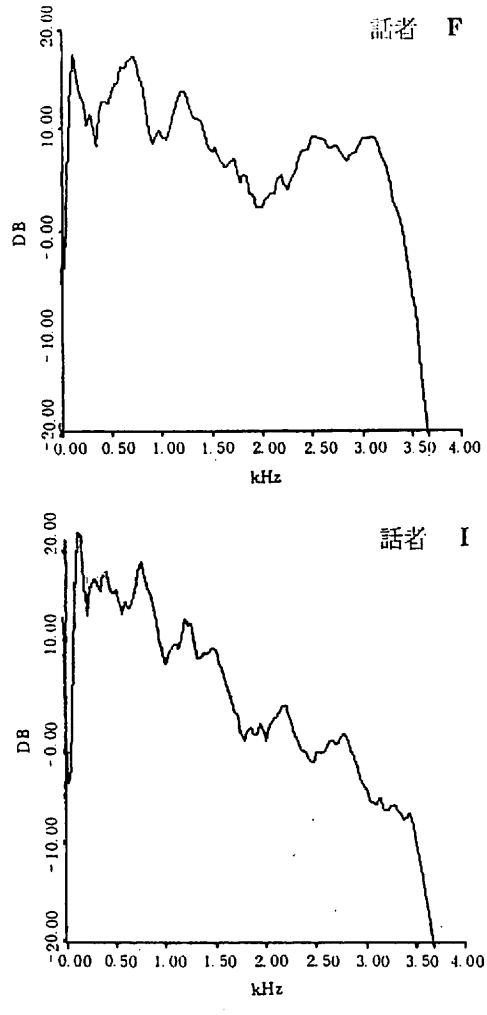

図11. 短文章音声の長時間平均スペクトルの例

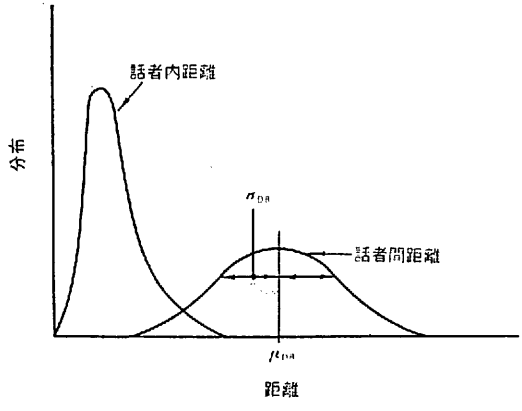

図12. 同一話者内および異話者間における標準パターンと入力音声の距離の分布

１１．話者認識の今後の課題と展望

米国空軍の委託を受けたMitre社の調査[24]によれば、音声は指紋や筆跡よりも、誤り率・処理時間のいずれの点でも優れていることが示されている。話者認識に対する将来のニーズに答えるために、今後研究しなければならない課題としては、次のようなものがある。

(1)長期間にわたって安定で、しかも作り声や風邪などに強い特徴パラメータの開発。

(2)電話機、伝送路等による歪や雑音を除去する対策の開発。

(3)標準パターンのための記憶容量や、分析・認識のための処理量の削減による経済化。

(4)数千～数万の大量の被験者による長期間の評価実験。

(5)発声内容独立形の認識方法の開発。

上記(5)の課題は、音声波に含まれる個人性情報と音韻性情報の分離という意味で、不特定話者の音声認識の研究と表裏一体をなしており、難しい課題であるが、今後の地道な努力による進展が期待される。

話者認識で用いる音声分析技術の多くは、単語音声認

識等で用いられるものと共通であるから、音声分析用のＬＳＩの開発が進めば、発声内容依存形の話者照合を用いた、一般家庭用あるいは車や金庫等の音声キーの実現は、そう遠い将来のことではないかもしれない。

参考文献

[1]Atal,B.S.:"Automatic recognition of speakers from their voices",Proc.IEEE,64,4,pp.460-475 (1976)

[2]Rosenberg,A.E.:"Automatic speaker verification:A review",Proc.IEEE,64,4,pp.475-487 (1976)

[3]斎藤収三:"声質情報の認識"、講座・音声認識III、信学誌、65、3、pp.262-271 (昭57)

[4]古井貞熙:"話者認識の現状と展望"、信学誌、67、5、pp.101-107 (昭59)

[5]古井貞熙:"音声の個人性情報"、数理科学、252、June、pp.33-41 (昭59)

[6]"On the theory and practice of voice identification",The National Research Council,Washington,D.C. (1979)

[7]古井貞熙:"音声波に含まれる個人性情報"、東京大学学位論文 (昭53)

[8]Kersta,L.G.:"Voiceprint identification",Nature,196,pp.1253-1257 (1962)

[9]Tosi,O. Oyer,H. Lashbrook,W. Pedrey,C. Nicol,J. and Nash,E.:"Experiment on voice identification",J.Acoust.Soc.Amer.,51,6(pt.2),pp.2030-2043 (1972)

[10]伊藤、斎藤:"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、1、pp.101-108 (昭57)

[11]Rosenberg,A.E. and Sambur,M.R.:"New techniques for automatic speaker verification",IEEE Trans.Acoust.,Speech,Signal Processing,ASSP-23,2,pp.169-176 (1975)

[12]Doddington,G.R.:"Speaker verification",Rome Air Development Center,Tech Rep.,RADC 74-179 (1974)

[13]古井貞熙:"話者認識における登録話者数の効果"、信学会情報全大、228 (昭52)

[14]Furui,S.:"Comparison of speaker recognition methods using statistical features and dynamic features",IEEE Trans.Acoust.,Speech,Signal Processing,ASSP-29,3,pp.342-350 (1981)

[15]古井貞熙:"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、57-A、12、pp.880-887 (昭49)

[16]Doddington,G.R.:"Personal identity verification using voice",Proc.ELECTRO-76,22-4,pp.1-5 (1976)

[17]Doddington,G.R.:"Voice authentication gets the go-ahead for security systems",Speech Technology,2,1,pp.14-23 (1983)

[18]Furui,S.:"Cepstral analysis technique for automatic speaker verification",IEEE Trans.Acoust.,Speech,Signal Processing,ASSP-29,2,pp.254-272 (1981)

[19]古井貞熙:"ケプストラムの統計的特徴による話者認識"、信学論、J65-A、2、pp.183-190 (昭57)

[20]古井、板倉、斎藤:"長時間平均スペクトルによる話者認識"、信学論、55-A、10、pp.549-556 (昭47)

[21]Sambur,M.R.:"Speaker recognition using orthogonal linear prediction",IEEE Trans.Acoust.,Speech,Signal Processing,ASSP-24,4,pp.283-289 (1976)

[22]Atal,B.S.:"Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification",J.Acoust.Soc.Amer.,55,6,pp.1304-1312 (1974)

[23]Li,K.P. and Wrench,Jr.,E.H.:"An approach to text-independent speaker recognition with short utterances",Proc.IEEE Int.Conf.Acoust.,Speech,Signal Processing,Boston,MA,12,9,pp.555-558 (1983)

[24]"Test results-Advanced development models of BISS identity verification equipment",MITRE Technical Rep.1,Rev.1,MTR-3442 (1977)