論文 / 著書情報
Article / Book Information

論題(和文)	ケプストラムの統計的特徴による話者認識
Title(English)	
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	電子情報通信学会論文誌, Vol. J65-A, No. 2, pp. 183-190
Citation(English)	, Vol. J65-A, No. 2, pp. 183-190
発行日 / Pub. date	1982, 2
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 1982 Institute of Electronics, Information and Communication Engineers.

UDC 534.784.001.36:[534.442.087.9:681.322.022:519.233.4-37]

ケプストラムの統計的特徴による話者認識

正員　古井　貞熙†

Speaker Recognition by Statistical Features of Cepstral Parameters

Sadaoki FURUI†, *Regular Member*

あらまし　単語音声スペクトルの統計的特徴を用いた新しい話者認識方法の原理と，電話音声によるオンライン実験を含む評価実験の結果について述べる．この認識方法では，各話者はあらかじめ定めた4種類の単語を発声し，その音声波はＬＰＣケプストラムと基本周波数の時系列に変換される．各単語の有声音区間における各パラメータの平均値，標準偏差，低次のフーリエ展開係数，およびパラメータ間の相互相関行列が計算され，これらの中から分散分析によって選択された特徴を用いて，話者の識別および照合が行われる．まず静かな部屋で発声された男性9名の長期間の音声を用いた実験を行ったところ，学習期間が3箇月おきの4時期，入力音声との時期差が3箇月の条件で，99.9％以上の認識率が得られた．次に実際の交換機と内線を通った電話音声による評価を行うため，ミニコンを用いたオンライン話者照合システムを作成し，男女計34名の話者による2箇月半にわたる実験を行った．このシステムでは，電話音声のひずみなどに対処するため，ピッチ抽出などに特別な配慮がされている．実験の結果，97～98％の認識率が得られ，この認識方法の有効性が確認された．

1. まえがき

音声波には，発せられた言葉の意味内容に関する情報のほかに，発声者個人の声の特徴，喜怒哀楽に関する情報などが含まれている．このうち発声者の声の特徴を自動的に抽出して，その話者が誰であるかを判定する話者認識の研究は，意味内容に関する情報を抽出し認識する（狭義の）音声認識の研究に比べて地味ではあるが，国内外の研究機関で地道に進められている．このような技術は，既に一部人間の能力を凌駕（りょうが）する程度にまで進んでおり，将来は，音声を印鑑や身分証明書代りに用いた，電話によるバンキング，クレジット電話，種々の場所や情報へのアクセスをコントロールするための音声による鍵などへの応用が考えられる．

話者認識の形態は，ある音声が，登録されている複数の話者のうちの誰の声に最も似ているかを判定する話者識別（speaker identification）と，ある音声が，名乗った本人の音声であるか否かを判定する話者照合（speaker verification）に分類できる．話者識別においては，入力音声から抽出した種々の特徴を，登録されている各話者の標準パターンと比較し，最も類似している標準パターンの話者による音声であると判定する．話者照合においては，入力音声から抽出した種々の特徴と，その発声者が名乗った人の標準パターンとの類似度が，あらかじめ定めたしきい値よりも大きければ，その音声は名乗った本人の音声であるとして受理し，そうでなければ他人の音声であるとして棄却する．話者照合の2種の誤り率，すなわち本人の音声を棄却する誤りと他人の音声を受理する誤りは，当然判定のしきい値によって変る．一般にはこのしきい値を，両者の誤り率が等しくなる値に事後的に設定し，このときの誤り率によって性能を評価する場合が多いが，実際の場合には，いかにしてしきい値を適切な値に設定するかが重要な課題となる．

一般に話者認識を行う上で最も難しい問題点の一つは，同じ人が同じ言葉を発声しても，数箇月程度以上の期間をおくと，そのつど音声のスペクトルが変化し，それが認識の精度に大きな影響を与える点にある．この問題に対処する方法として，筆者らは，1次系などの逆フィルタによるスペクトル等化処理が有効である

†電電公社武蔵野電気通信研究所，武蔵野市
Musashino Electrical Communication Laboratory, N.T.T., Musashino-shi, 180 Japan
論文番号：昭57-論62[A-23]

ことを明らかにしたが[1]，この処理を施した場合でも，スペクトル変動の影響を免れることはできない．このため話者認識系の評価には，長期間にわたって収集した音声サンプルを用いることが不可欠である．

筆者らはこれまでに，このような立場に立って種々の話者認識方法を提案し，有効性の比較を行ってきた．まず発声内容不依存形の認識方法として，長時間平均スペクトルのケプストラムを用いる方法を提案し，この有効性を確認したが，９３％程度の認識率を得るために１０秒長の音声を必要とするという問題点があった[2]．次に発声内容依存形の認識方法として，単語音声の対数断面積比および基本周波数の統計的特徴を用いた認識方法を提案し，４単語の特徴を組み合せて用いれば，９９％程度の認識率が得られることを示した[3]．一方，動的特徴を用いた発声内容依存形の認識方法としては，対数断面積比と基本周波数の時系列と，パターンマッチング時の時間伸縮関数に含まれる情報を用いた方法[4]と，ＬＰＣケプストラムとその多項式展開係数の時系列の時間正規化マッチングによる認識方法[5]を提案し，これらの方法の有効性を示した．動的特徴と統計的特徴を用いる方法には，細かい点では有効性に相違点があるが，どちらの方法を用いても同程度の性能を得ることが可能であり，一般に後者の特徴を用いる場合の方が計算量や記憶容量が少なくてすむという長所がある[6]．

本論文では，統計的特徴を用いる新しい方法の原理と，この方法を用いた話者認識実験の結果，更にこの方法に基づいて作成した，電話音声用のオンライン話者照合システムの構成とその実験結果について述べる．

2. ケプストラムの統計的特徴を用いた話者認識

単語音声スペクトルの統計的特徴を用いた発声内容依存形の認識方法の改良として，図１に示す方法について検討した．音声サンプルとしては種々の単語音声を用い，音声波に対してまず6.67 kHz，12 bitで標本化および量子化を行う．次に音声の始端と終端を検出したのち，10 msごとに30 msの区間（フレーム）を切り出して，各区間ごとに有声／無声の判定を行い，有声音区間のみの相関係数の時系列に変換する．この相関係数を，各単語の全有声音区間にわたって平均化し，この値を用いて１次系の逆フィルタを構成する．もとの相関係数に畳込みを行ってスペクトル等化処理を行い，こののちに基本周波数 f_{0t} と10次までの（ＬＰＣ）ケプストラム $\{c_{it}\}_{i=1}^{10}$ からなる11次元ベクトル $\boldsymbol{X}_t$ の時系列に変換する．

$$\boldsymbol{X}_t=(c_{1t},c_{2t},\cdots,c_{10t},f_{0t}) \tag{1}$$

この時系列から，各単語の全有声音区間における平均値ベクトル μ，標準偏差ベクトル σ，共分散行列 Σ，相互相関行列 Λ，及びフーリエ展開係数行列 $\boldsymbol{F}$ を計算する．ここでは計算の便利のために，展開区間を π としてコサイン展開のみを行った．

$$\mu=\left(\sum_{t=1}^{N}\boldsymbol{X}_t\right)\Big/N \tag{2}$$

$$\Sigma=\left(\sigma_{ij}\right)=\left\{\sum_{t=1}^{N}(\boldsymbol{X}_t-\mu)'(\boldsymbol{X}_t-\mu)\right\}\Big/(N-1) \tag{3}$$

$$\sigma=(\sqrt{\sigma_{11}},\sqrt{\sigma_{22}},\cdots\cdots,\sqrt{\sigma_{11,11}}) \tag{4}$$

$$\Lambda=\left(\lambda_{ij}\right),\ \lambda_{ij}=\sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}} \tag{5}$$

$$\boldsymbol{F}=\left(F_{il}\right),\ F_{il}=\left\{\sum_{t=1}^{N}x_{it}\cos\frac{\pi(t-1)l}{N}\right\}\Big/N,$$
$$:\ 1\leqq l\leqq 6 \tag{6}$$

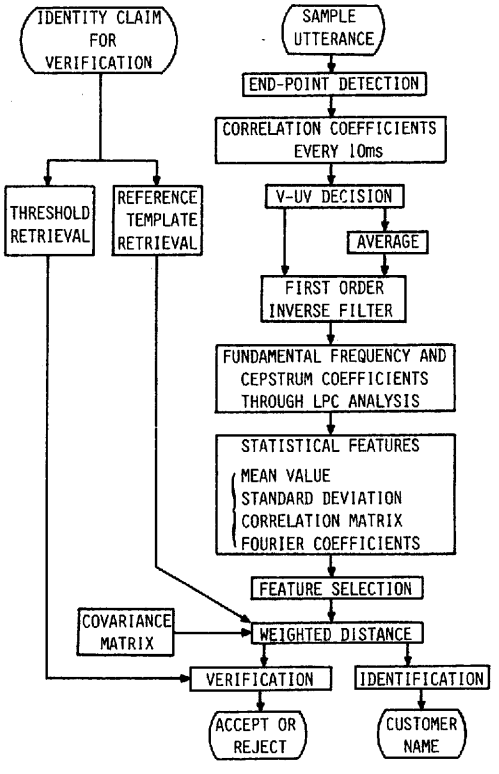

図１　話者識別と話者照合のブロック図
Fig. 1 - Principal operation of the speaker identification and verification system.

ここで，Nは全有声音区間のフレーム数，式(3)の′は転置を意味する．高次のフーリエ展開係数は認識への寄与が少ないと考えられるので，ここでは各時系列に対して第1次から第6次までの展開係数を計算する（第0次の展開係数は平均値ベクトルμの成分と同じである）．これらの統計量のうち，標準偏差ベクトルσ，相互相関行列A，及びフーリエ展開係数Fから，話者の分離における有効性が比較的大きい60成分を選択し，平均値ベクトルμの全成分と組み合せて，71次元の特徴ベクトルTとする．選択する成分は，多数話者の学習サンプルを用いて計算した各成分の話者間／話者内分散比に基づいて決定しておく．

話者認識の判定は，入力音声の特徴ベクトルTと，各登録話者の標準パターンR_kとの，各成分の安定性を考慮した次のような重み付き距離によって行う．

$$d(T,R_k)=|V|^{1/M}(T-R_k)V^{-1}(T-R_k)' \quad (7)$$

ここで，kは登録話者の番号，Vは特徴ベクトルの共分散行列である．Vは，まず各話者ごとに学習サンプルを用いて計算し，その行列をすべての登録話者について平均したものを用いる．各登録話者の標準パターンは，各話者の学習サンプルの特徴ベクトルの平均値を用いる．$d(T,R_k)$の値は，認識の判定に用いる各単語音声について計算したのち，全単語（通常4種類の単語）について平均化し，この値に基づいて話者識別および話者照合の判定を行う．話者照合の判定のしきい値は，事後的に各登録話者ごとに，2種類の誤り率が等しくなる最適値に設定する場合と，事前にすべての登録話者に共通の値に設定しておく場合について検討した．

単語音声スペクトルの統計的特徴を用いる，筆者らの従来の方法との主たる相違点は，スペクトルパラメータを対数断面積比からケプストラムに変えた点と，特徴量にフーリエ展開係数を加えた点にある．

3. 認識実験

3.1 実験方法

実験には，男性9名の話者が3年間の長期間にわたって繰り返し発声した/namae/,/baNgoo/,/koogeN/,/bakuoN/の4単語の音声サンプルを用い，4単語の特徴を組み合せて認識の判定を行った．これらの音声はいずれも，十分にしゃ音された静かな部屋で収録した．

〔認識実験1〕

高品質マイクロホンによって収録した音声を用い，各話者ごとに，3箇月おきの4時期に収録した12サンプルを学習サンプルとして，その3箇月後の音声の話者を識別する実験を，多時期の音声の種々の組み合せに対して行った．全話者の入力音声の総数は216サンプルである．話者照合の詐称者（登録話者以外の棄却されるべき音声の話者）の音声としては，9名の登録話者のうち本人以外の話者が各時期に3回ずつ発声した音声と，それ以外の男性103名が1回ずつ発声した音声を用いた．

〔認識実験2〕

カーボン送話器を含む600形電話機および擬似加入者線（0.4 mmφ−7 dB）を通った音声と，同時に録音したマイクロホン音声を用いて実験を行った．学習は6箇月おきの3時期（9サンプル），入力音声との時期差は6箇月とし，詐称者の音声として，登録話者のうちの本人以外の音声と，それ以外の男性11名が3回ずつ発声した音声を用いた．全登録話者の入力音声の総数は54サンプルである．

3.2 特徴パラメータの有効性の評価

上述の認識実験1で用いた9名の初めの4時期の音声によって，各特徴パラメータの有効性の評価を行った．各特徴パラメータ（統計量）ごとに計算した話者間／話者内分散比（F比）の値を，4単語に関して平均して図2に示した．フーリエ展開係数Fの要素と，標準偏差σ及び相互相関行列Aの要素（従来の特徴パラメータ）に分け，前者は展開係数ごとに，後者は全要素のF比の値の平均値を示した．平均値μのF比の値は，第0次のフーリエ展開係数として示してある．以前の方法との比較のために，スペクトルパラメータ

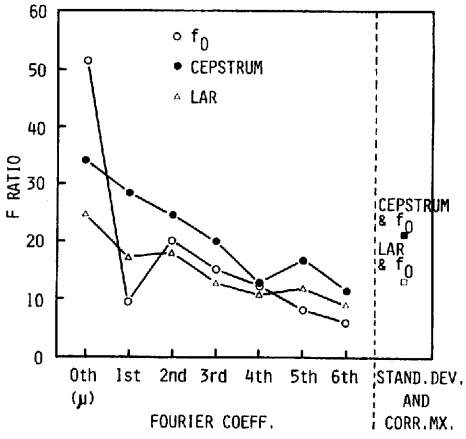

図2 各特徴パラメータの有効性（F比）
Fig.2 - Effectiveness of each feature parameter set.

として対数断面積比（ＬＡＲ）を用いる場合の結果を合せて示した．この結果から，ＬＡＲよりもケプストラムの方が有効性が高いこと，フーリエ展開係数の有効性は次数とともに低下するが，第３次以下の係数は，従来からの特徴である標準偏差および相互相関行列成分と同程度，あるいはそれ以上の有効性を有することが分かる．

3.3 認識実験の結果

認識実験１と認識実験２の結果をそれぞれ表１と表２に示す．これらの表には，全登録話者に関する平均誤認識率を示した．なお，話者照合で判定のしきい値を事前に設定する場合については，２種の誤り率（本人を棄却する誤り率と他人を受理する誤り率）の平均値を示した．表１には，特徴パラメータとしてＬＡＲを用いる場合の結果もあわせて示してある．ＬＡＲを用いて，フーリエ展開係数を除く従来の統計量のみを用いたときの平均誤認識率は，話者識別で 0.9 %，事後的にしきい値を最適に設定したときの話者照合で 0.7 %[3] であるので，スペクトルパラメータをＬＡＲのままにしても，フーリエ展開係数を特徴量に加えることによって，誤り率が約４割程度低下するが，スペクトルパラメータをＬＡＲからケプストラムに変更したことによる改善の割合は，更に大きいことが分かる．スペクトルパラメータとしてケプストラムを用い，フーリエ展開係数を特徴量に加えたときの平均認識率は，話者識別で 100 %，事後的なしきい値による話者照合で 99.98 %，事前にすべての話者に共通のしきい値を設定する話者照合で 99.90 %である．

表１ 認識実験１の結果

スペクトルパラメータ	認識の種類		平均誤認識率
ケプストラム	話者識別		0 %
	話者照合	事前しきい値	0.10 %
		事後しきい値	0.02 %
LAR	話者識別		0.46 %
	話者照合	事前しきい値	1.21 %
		事後しきい値	0.51 %

表２ 認識実験２の結果

入力系	認識の種類		平均誤認識率
電話系	話者識別		0 %
	話者照合	事前しきい値	2.44 %
		事後しきい値	0.49 %
マイクロホン	話者識別		0 %
	話者照合	事前しきい値	2.44 %
		事後しきい値	1.36 %

表２には，ケプストラムを用いた場合の結果のみが示してあるが，認識実験２は認識実験１よりも学習サンプル数が少ない上に，学習サンプルと入力音声との時期差が大きく，実験条件が厳しいので，誤り率が表１の結果よりもやや大きくなっている．表２の結果から，ここで検討している方法が，電話音声に対してもマイクロホン音声とほぼ同程度の性能で動作することが分かる．

4. 電話音声によるオンライン話者照合システム

4.1 システム構成法と実験方法

前章で述べた実験により，ここで検討している新しい話者認識方法の有効性が確認されたので，次にこの認識方法を用いて，実際の構内回線と交換機を通った電話音声によるオンラインの話者照合実験を行った．システムのハードウェア構成は，図３に示すとおりである．交換機は研究所の C-400 交換機で，話者照合システムは通常の電話端末の一つとして接続されている．

話者照合実験を行う各話者は，カーボン送話機を内蔵した通常の電話機からシステムを呼び出し，音声応答による指示に従って，自分の識別番号をダイヤル入力したのち，キーワードを発声する．このキーワードはすべての話者に共通に，前章までの実験に用いたものと同じ４種類の単語を用いた．キーワードの音声が入力されると，実時間で相関関数と基本周期の時系列が抽出され，音声区間の検出ののち，ケプストラムと基本周波数の時系列に変換される．次に各統計的パラメータが計算され，識別番号で検索された話者の標準パターンとの距離が計算される．この距離が，その話者についてあらかじめ定めてあるしきい値よりも小さければ，その入力音声は本人の音声であると判断し，そうでなければ本人の音声ではないと判断する．後者の場合は，念のためにもう一度試みるよう話者に要請する．

各話者の標準パターンは，まず１回目と２回目の音声から抽出した特徴パラメータの平均値として作成し，以後はその安定化をはかると共に声質の変化に対応するために，新しい音声が本人の音声として受理されるごとに，平均化を行って更新してゆくようにした．各話者の照合判定のしきい値も，同時に更新した．各話者には，発声の速度，レベル，長さなどについては特に指示は行わなかったが，意図的に声を変えることはしないように要請した．

実験に使われた電話機は 600 形または 601 形が多い

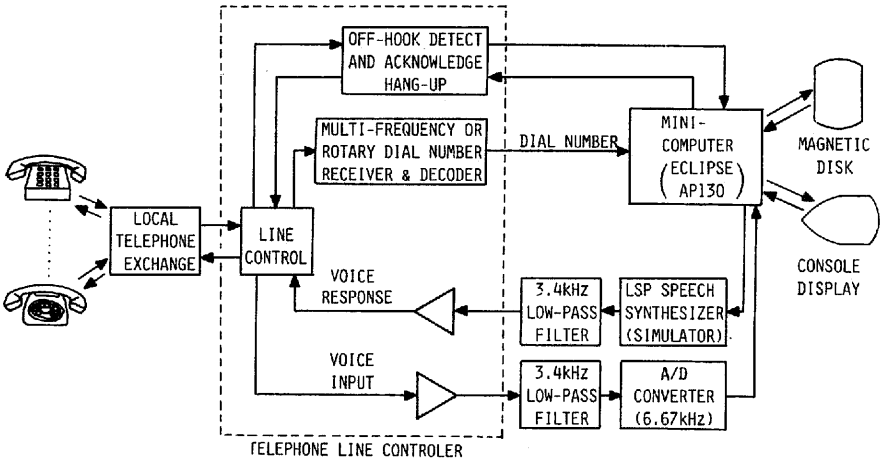

図3　オンライン話者照合システムのハードウェア構成
Fig.3 - Experimental setup block diagram of the online speaker verification system.

が，若干650形などの他の機種も含まれている．各話者は自分の机上の電話機から実験を行うことを原則としたが，そのときの都合により，実験室や他の部屋の電話機から実験を行ってもよいことにし，電話機や回線は特に固定しないことにした．但し，外線からの実験は行っていない．音声応答は，ＬＳＰ方式による分析合成音によって行った．各話者からシステムへの着信の検出，回線の接続，ダイヤル入力による各話者の識別番号の読み取り，回線の切断などは専用のハードウェア（回線制御装置）によって行った．

システムの全体的な制御や，ＬＰＣ分析，音声区間検出，ピッチ抽出，統計的パラメータの抽出，標準パターンとの距離計算，標準パターンとしきい値の更新などのすべての演算処理は，アレープロセッサを内蔵したミニコンピュータ（ECLIPSE-AP130）によって行った．但し，実時間による相関分析とピッチ抽出を可能とするために，フレーム周期は，前章までのオフラインの実験では10msであったが，このオンライン実験では15msに変更した．同じ理由で，オフラインの実験ではピッチ抽出に変形相関関数を用いていたが，ここでは相関関数のピーク位置による方法に変更した．

相関関数と基本周期が実時間で抽出された後の処理の所要時間は，音声区間検出，統計的特徴の抽出（ピッチの自動修正を含む），及び認識のための距離計算がそれぞれ，約1.5秒，4.5秒，3秒であり，これにモニタ用の表示とオーバヘッドの3秒を加えて，システムの応答時間は約12秒である．演算部のハードウェア化を行えば，容易に全部の処理を実時間に近い時間で終了できると考えられる．

4.2　ピッチ抽出と音声区間検出

オンライン話者照合システムにおける特徴抽出と認識のアルゴリズムは，前章のオフラインの実験の場合とほぼ同じであるが，前節で述べたようにフレーム周期とピッチ抽出法が異なるほか，雑音レベルと音声レベルの変動に対処するため，音声区間の検出に特別の配慮を払っている．

音声区間の検出法に関しては，各話者からの着信があると，すぐに雑音レベルを自動的に測定し，この値に従ってパワーのしきい値を設定して，パワーがこのしきい値を越える区間を音声区間の候補とする．これだけでは雑音の区間も音声区間の候補としてしまうことがあるので，候補区間の長さや，他の候補区間との時間的関係などを考慮して，音声区間の決定を行う．

ピッチ抽出に関しては，カーボン送話器と実回線を通った音声から相関関数のピーク位置のみによって正しいピッチ軌跡を得ることは難しいので，ここでは，図4に示す方法で自動修正を行った．この方法では，まず各単語の全有声音区間における基本周波数のヒストグラムを作成する．そして最もひん度の高い周波数に近い範囲が真の基本周波数の存在範囲（信頼範囲）であるとし，この範囲からはずれている点の値を自動的に修正する．なお，相関関数のピークによる抽出法では2倍の周波数の値が抽出されることが比較的多いので，信頼範囲を決定する際にこの点を考慮する．こうして信頼範囲からはずれた値を修正したのち，軌跡の不連続性が大きい孤立点を修正し，全体の軌跡に対して移動平均による平滑化を行って，これを最終的な

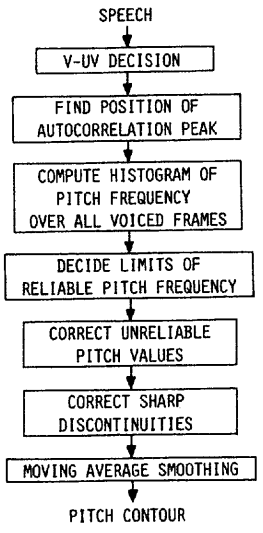

図4 ピッチ抽出と自動修正の方法
Fig.4 - Block diagram of pitch extraction and its automatic correction.

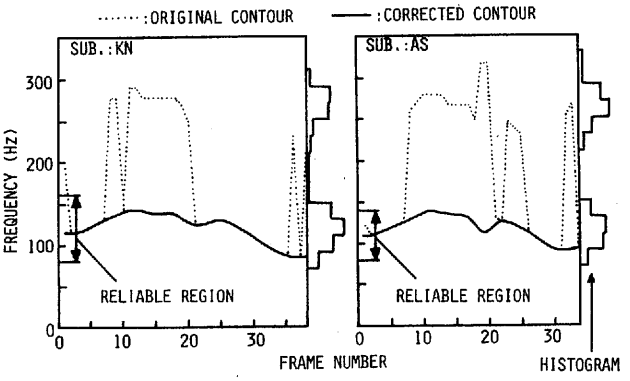

図5 ピッチの原軌跡と自動修正後の軌跡の例（男性話者の/baNgoo/）
Fig.5 - Examples of original and automatically corrected pitch contours (word/baNgoo/ uttered by male speakers).

ピッチ軌跡とする．この修正アルゴリズムでは，信頼範囲を誤ると致命的となるが，ヒストグラムという，全フレームを横断的に見た安定した特徴に基づいているので，このようなことはほとんどない．そして正しい値はそのままにして，明らかに異常とみなされる点のみの修正を行うので，初めから平滑化を適用することができないような，抽出誤りが多い場合にも適用できる．

図5に，相関関数のピークによって直接抽出したピッチ軌跡と，上述のアルゴリズムによって自動的に修正した後の軌跡の例を示す．ここで用いた方法が，かなり誤りの多い軌跡に対しても，正しい値を変えてしまうことなく，誤った値を適切に修正する能力を有することが分かる．この方法は，女性の音声に対しても同様に有効に動作することが確認されている．

4.3 標準パターンとしきい値の更新方法

i 番目の入力音声の特徴ベクトルを $\boldsymbol{T}_i$，そのときの話者 k の標準パターンを $\boldsymbol{R}_{ki}$ とするとき，$\boldsymbol{T}_i$ が話者 k の音声として受理された場合に限り，次式に従って標準パターンを更新する．

$$\boldsymbol{R}_{k,i+1}=\{(i-1)\boldsymbol{R}_{ki}+\boldsymbol{T}_i\}/i \quad (3\leqq i\leqq 10) \tag{8}$$

$$\boldsymbol{R}_{k,i+1}=(9\boldsymbol{R}_{ki}+\boldsymbol{T}_i)/10 \quad (11\leqq i) \tag{9}$$

この方法では，過去の値がいつまでもその影響を及ぼす（徐々にその影響は小さくなるが）という問題はあるが，過去の入力を蓄えておく必要がないという長所がある．11番目以降の入力に対して式(9)を用いることにより，それらの入力が標準パターンの更新に寄与する度合いが小さくなり過ぎるのを防いでいる．

話者照合の判定のしきい値の更新には次式を用いた．

$$\theta_{i+1}=\theta_0 \quad (i=2) \tag{10}$$

$$\theta_{i+1}=\mathrm{Min}\{\{(a-1)\theta_i+d_i+b\}/a,\theta_0\} \quad (3\leqq i) \tag{11}$$

ここで，θ_i は i 番目の入力に対するしきい値，d_i は i 番目の入力と標準パターンとの距離，θ_0 はしきい値の初期値，a は過去の値との重み付き平均における重みを与える定数，b は距離の値の変動を考慮したマージンである．θ_0, a, 及び b の値は，予備実験によって決定した．

5. オンライン話者照合実験

5.1 実験方法

前章で述べたシステムにより，オンラインの話者照合実験を行った．実験に用いた話者は成人男性29名・女性5名の計34名である．まず初めに各話者ごとに4種類の各単語を2回発声させて標準パターンを作成し，続けて最初のテストサンプル（入力音声）を発声してもらった．以後は各話者に，日を変えて週に2～3回程度システムを呼び出して，約2箇月半にわたって実験を行ってもらった．照合の結果棄却された場合，SN比が異常に低い場合，発声間違いの場合などには再発声を依頼した．オンライン実験で入力された音声と，異なる話者の標準パターンとの照合実験は，オンライン実験とは別に，それらの入力音声を用いて定期的にまとめて行い，正しく棄却されるか否かを調べた．

特徴選択と重み付き距離の設定のための F 比などの

統計的分析は，実験の初期の段階では3.のオフライン実験による結果をそのまま用い，オンラインの入力音声が各話者について7サンプル程度集積された段階で，それらの入力音声を用いた分析を行って，以後はこの結果に基づく特徴選択と重み付けを行った．

5.2 実験結果

図6に，話者照合の2種の誤り率の時間的推移を示す．3番目のサンプルが最初のテストサンプルであるが，標準パターンを作成するための学習サンプルと同じときに発声された音声なので，本人の棄却誤り（FR，False Rejection）は小さい．4番目のサンプルは，初めて学習サンプルと異なる日の音声が入力されたので，FRがかなり大きい．しかしこの場合でも，しきい値が比較的大きく設定されているので，繰り返し発声すれば受理されることが多く，この受理された音声によって標準パターンが更新される．

各話者の初めの3テストサンプルを用いた異話者間の照合実験の結果が，5番目の入力時点に示されている他人の受理誤り（FA，False Acceptance）の値である．なお異話者間の音声を比較したときに，男女間で同じ話者の音声とみなす誤りを生ずることはないので，このような組み合せの実験は行っていない．初めのうちはしきい値が大きく設定してあるのでFAがかなり大きいが，6番目のサンプル以降になると標準パターンが徐々に安定化し，各話者が発声に慣れてくるということもあって，話者内における距離が徐々に低下してくるので，しきい値も徐々に低下し，FAはFRと共に低下してくる．

FRはほぼ9番目の入力以降安定しており，FAも20番目前後から安定している．安定後の平均誤認識率は，FRが約3％，FAが約2％である．なお，ひどい風邪のため本人の声が2回以上続けて棄却された場合が，全体の実験を通じて，2人の話者について1回ずつあり，この結果は図6には含めていない．話者によっては1週間以上の間隔をおいて発声したり，ときどき計算機室のような騒音レベルの高い（65～70 dB(A)）環境から実験を行った者もあるが，このような場合にも特に問題はなかった．34名の各話者についてみると，9番目の発声以降のFRの回数は，0回が20名，1回が10名，2回が2名，3回が2名で，FRの全くない話者が最も多い．

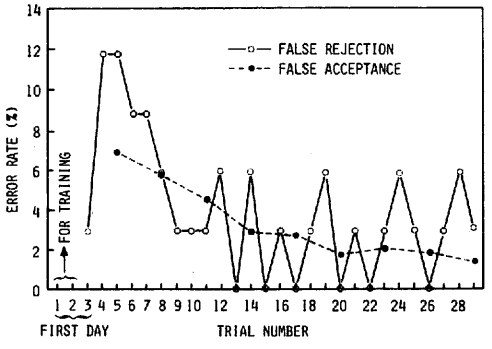

図6 オンライン話者照合実験における2種の誤り率の推移

Fig.6 - Two kinds of error rate vs. trial number in the online speaker verification experiment.

ここで得られた安定化後の誤り率，すなわちFR≒2％，FA≒3％という値は，これまでベル研究所において，同様に研究所内の内線を用いて行った実験の結果[7]と比べると，ほぼ1/2程度に小さい．実験の規模がやや異なるので詳細な比較はできないが，ここで検討した方法の有効性を示しているといえよう．なおこの実験の終了後，1箇月おきに各話者に実験を行ってもらっているが，特に誤りが増えるという傾向は見られていない．

6. むすび

本論文では，単語音声スペクトルの統計的特徴を用いた新しい話者認識方法の原理と，電話音声によるオンライン実験を含む評価実験の結果について述べた．この認識方法では，各話者はあらかじめ定めた4種類の単語を発声する．音声波はLPCケプストラムと基本周波数の時系列に変換されたのち，各単語の有声音区間における，各パラメータの平均値，標準偏差，低次のフーリエ展開係数，およびパラメータ間の相互相関行列が計算され，これらの中から選択した特徴を用いて，話者の識別および照合が行われる．

まず，静かな部屋でマイクロホンによって収録した音声を用いた認識実験を行ったところ，学習期間が3箇月おきの4時期，入力音声との時期差が3箇月の条件で，話者9名の話者識別で100％，しきい値を事後的に設定した話者照合で99.98％，しきい値をすべての話者に共通の値に設定した話者照合で99.90％の平均認識率が得られた．次に，カーボン送話器を含む電話機と擬似加入者線を通して収録した音声と，同時にマイクロホンによって収録した音声を用いた実験をそれぞれ行って，両者の比較を行ったところ，どちらの音声の場合でもほぼ同程度の認識率が得られることが分かった．

そこで上述の方法を用いて，実際の交換機と内線を通った電話音声を対象とするオンラインの話者照合システムを作成し，男女計34名の話者によるほぼ2箇月半にわたるオンライン実験を行った．その結果，実験の初期の段階では，各話者の標準パターンが偏っているため誤り率が大きいが，標準パターンが安定した後には，本人の音声を棄却する誤り率と他人の音声を受理する誤り率として，それぞれ約3％と2％の値が得られた．オンラインの条件では音声スペクトルの変動要因が非常に多く，雑音などの影響も受けるため，これらの値はオフライン実験における値よりも大きいが，これまでに報告されている他の研究機関における実験結果と比べると，1/2程度である．以上の結果は，この話者認識方式が，高品質の音声のみならず，実際の電話音声に対しても高い有効性を有することを示している．

オンライン実験におけるスペクトルの変動要因としては，入力レベル（SN比）の変動，電話機と回線の特性のばらつき，送話器と口との距離や角度の変動，慣れによる雑な発声，周囲雑音の重畳，風邪などの発声者の体調などが上げられる．又，分析フレーム周期を10 msから15 msに延長した影響も変動要因に加わっていると考えられる．実際のデータを分析してみると，本人の音声が棄却されるケースは，入力音声のSN比が低い場合に多く，SN比（有声音区間の平均値）が25 dB程度以下になると，急激に棄却率が大きくなる傾向が見られる．

ここで用いている認識方法では，一人の話者の標準パターンを蓄えるのに必要な記憶容量は，71次元×4単語＝284語であり，音声の全フレームデータを蓄えるような方式と比べると，1/7程度（1単語のフレーム数が45程度のとき）に少ない．話者照合システムをバンキングサービスなどに導入する際には，かなり大量の顧客の標準パターンを蓄える必要が生ずると考えられるので，記憶容量が少なくてすむのは大きな長所となると考えられる．

今後はSN比が低い音声に対する対策などを検討するとともに，更に多数の話者の長期間の音声を用いた実験，そして市外回線を通った音声による実験などを通じて，問題点の解明と方式の改良を進めて行きたい．

謝辞　日ごろ御指導ごべんたつを頂く，当所畔柳基礎研究部長，橋本統括役，板倉第四研究室長，ならびに熱心に討論して頂く第四研究室の諸氏に感謝する．

文　献

(1) 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(A)，57-A，12，pp.880-887（昭49-12）．
(2) 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(A)，55-A，10，pp.549-556（昭47-10）．
(3) 古井，板倉：“単語の統計的パラメータによる話者認識”，信学論(A)，56-A，11，pp.717-724（昭48-11）．
(4) Saito, S. and Furui, S.: “Personal information in dynamic characteristics of speech spectra”, 4th IJCPR, pp.1014-1018 (Nov.1978).
(5) Furui, S.: “Cepstral analysis technique for automatic speaker verification”, IEEE Trans. Acoust., Speech & Signal Process, ASSP-29, 2, pp.254-272 (April 1981).
(6) Furui, S.: “Comparison of speaker recognition methods using statistical features and dynamic features”, IEEE Trans. Acoust., Speech & Signal Process., ASSP-29, 3, pp.342-350 (June 1981)
(7) Rosenberg, A.E.: “Evaluation of an automatic speaker-verification system over telephone lines”, Bell Syst. Tech. J., 55, 6, pp.723-744 (July-Aug.1976).

（昭和56年7月6日受付，8月26日再受付）