

論文 / 著書情報
Article / Book Information

論題	統計的特徴と動的特徴による話者認識
著者	古井貞熙
掲載誌/書名	研究実用化報告, Vol. 30, No. 8, pp. 2045-2056
発行日	1981,

543.78 : 534.322 : 165 : 519.24 : 531.3

統計的特徴と動的特徴による話者認識*

古 井 貞 熙**

あらまし 発声内容依存形の話者認識に関して、単語あるいは短文章音声のスペクトルの動的特徴、すなわちスペクトルの時間パターンを直接的に用いる方法と、統計的特徴に変換してから認識を行う方法の有効性の比較評価を、筆者らがこれまでに検討した4種類の方式について行った。その結果、統計的特徴、動的特徴のいずれを用いても、同程度の性能を得ることが可能で、認識に必要な計算量と記憶容量の点から前者の特徴の方が有利であることや、両者の特徴の組合せ効果、スペクトル等化処理の有効性等が明らかとなった。LPC ケプストラムの統計的特徴あるいは動的特徴を用いる、新しい改良形の方式によれば、マイクロホン音声のみならず、電話音声に対しても、極めて高い精度で話者認識を行うことができる。前者の方法によるマイクロホン音声に対する認識率は、各話者の学習期間が10カ月、入力音声までの時期差が3カ月のとき、99.9%程度である。

1 ま え が き

音声波にふくまれる個人性情報を用いて、その音声の話者を自動的に判定する話者認識の研究が国内・外で進められている。一般に話者認識(speaker recognition)は、話者照合(speaker verification)と話者識別(speaker identification)にわけることができる^{(1)~(3)}。話者照合とは、入力された音声の名乗った本人のものであるか否かを、入力音声と登録されている音声との比較によって判定するものであり、話者識別とは、入力された音声、登録されている複数の話者のうちの誰の声であるかを、入力された音声と最も類似している登録音声を探し出すことによって判定するものである。話者照合あるいは話者識別を自動的に行うシステムが完成すれば、電話によるバンキングサービス等の種々の応用が考えられ、米国の銀行では、実際のサービスの試行が検討されている⁽⁴⁾。

話者認識の形態は、さらに別の観点から分類すること

ができる。すなわち、判定に用いる音声の発声内容をあらかじめ限定せずに、どんな言葉を発声してもよいとする方法(発声内容不依存形 text-independent)と、あらかじめ発声内容を限定しておく方法(発声内容依存形 text-dependent)である。当然ながら、前者の方法が確立されることが理想的であるが、このためには、音声波にふくまれる音韻性情報と個人性情報を明確に分離して抽出することが必要であり、現時点ではかなり難しい。

さらに、話者認識の方法は、判定に用いる特徴パラメータの種類によって、音声スペクトル(ピッチ・パワー等もふくむ)の時系列(動的特徴)をそのまま用いるものと、時間という変数を持たない統計的なパラメータ(統計的特徴)に変換してから、それを用いるものの二つに分類することができる。時系列を直接的に用いる方法を、動的特徴による方法と呼ぶのには異論の余地もあり、一方、後述する統計的特徴の中にも、スペクトルの動的な特徴が実際には反映されているが、本論文では二種類の方法を対比するために、このような名称を用いることとした。これまでの発声内容不依存形の認識では、必ずといってよいほど統計的特徴が用いられており、発声内容依存形の認識では、動的特徴が用いられる場合が多い。

筆者らはこれまでに、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

* Comparative Study of Speaker Recognition Methods Using Statistical or Dynamic Features of Speech Spectrum. By Sadaoki FURUI.

この研究は第四研究室で行われたものである。

** 武蔵野電気通信研究所研究専門調査役

を提案し、この方法の有効性を確認したが⁽¹¹⁾⁽⁶⁾、93%程度の認識率を得るために10秒長以上の音声を用いる必要があるという問題点があった。次に、筆者らは、発声内容依存形の認識方法として、単語音声の対数断面面積比および基本周波数の統計量を用いた認識方法を提案し、4単語から抽出した特徴を組合せて用いれば、99%程度の精度で話者の認識が可能であることを示した⁽⁷⁾⁽⁸⁾。筆者らはさらに、動的特徴を用いた発声内容依存形の認識方法として、対数断面面積比と基本周波数の時系列と、そのボタンマッチング時の時間軸正規化に伴って抽出される時間伸縮関数にふくまれる情報を用いた認識方法を提案し、この有効性を示した⁽⁹⁾⁽¹⁰⁾。この方法は、単に時系列のマッチングを行うのみならず、その時間軸非線形伸縮によって失われてしまう動的情報を積極的に抽出して、認識のための情報として用いようとしたものである。この方法はまだ改良の余地があるが、今後はこのような新しい動的情報の抽出法の検討が、重要な課題の一つとなると考えられる。

筆者らがこれまでに検討した上述の種々の方法には、それぞれ特徴があり、その優劣を決めるのは必ずしも容易ではない。ここでは発声内容依存形の方法に対象を絞り、筆者らがこれまでに検討したうちの代表的な二つの方法に、最近検討した二つの方法を加えた四つの方法について、その有効性の比較評価を行った結果を述べる。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、数か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動し、それが話者認識の精度に大きな影響を与える点にある⁽¹¹⁾⁽¹²⁾。この問

題は、音韻性情報を対象とした、いわゆる音声認識(単語音声認識)の場合でも必ずしも無視できる問題ではないが、その影響が現れる度合いは、話者認識の場合よりも小さい⁽¹³⁾。本論文でも、このような性質を考慮しながら、各方法の比較を行った。また、このようなスペクトルの長期間にわたる変動の影響を軽減する方法としてのスペクトル等化処理の有効性の評価や、統計的特徴と動的特徴を組合せる効果の検討も行った。

ここで検討した四つの方法の概略を表1に示した⁽¹⁴⁾。方法1と方法2は統計的特徴を用いる方法、方法3と方法4は動的特徴を用いる方法である。発声内容としては、方法4は短文章音声を用いているが、他の方法は複数個の単語を発声した音声を用いている。スペクトルパラメータとしては、方法1と方法3では対数断面面積比と基本周波数、方法2ではLPCケプストラムと基本周波数、方法4ではLPCケプストラムと多項式展開係数が用いられている。方法2と方法4は、筆者が最近検討した方法であり、それぞれ方法1と方法3の改良になっている。それぞれの方法の詳細は後述する。

2 方法1と方法3の比較

2.1 音声分析

図1に、方法1と方法3を比較するための実験システム全体のブロック図を示す⁽¹⁵⁾⁽¹⁶⁾。実験には、男性9名が7年間の長期間にわたって繰り返し発声した/na₁mae/と/ba₁ngo:/の二種類の単語音声を用いた。音声波

表1 ここで検討した4種類の話者認識方法の概略

	特徴量	発声内容	認識方法と尺度
方法1	統計的特徴	複 数 単 語	対数断面面積比(LAR)と基本周波数の統計的特徴(平均値・標準偏差・相関行列)の重みつき距離
方法2	統計的特徴	複 数 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・相関行列・フーリエ展開係数)の重みつき距離
方法3	動的特徴	複 数 単 語	LAR と基本周波数の時系列の時間伸縮マッチング
方法4	動的特徴	短文章	LPC ケプストラムとその多項式展開係数の時系列の時間伸縮マッチング

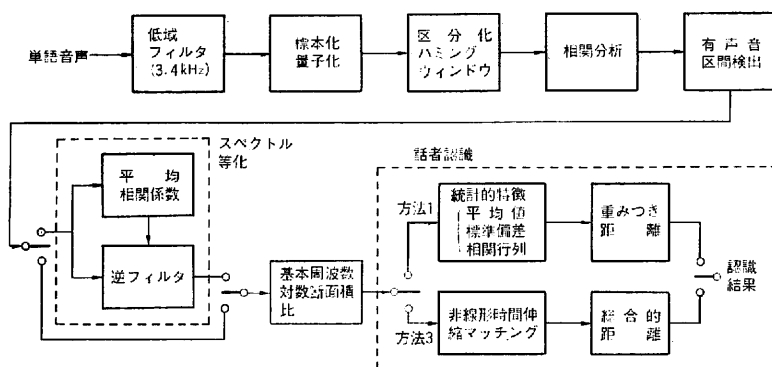


図1 方法1と方法3の比較実験システムのブロック図

は 8 kHz, 11 bit で A/D 変換され, 10 ms ごとに 32 ms の区間 (フレーム) をハミング窓を乗じて切り出す。各フレームごとに有聲/無聲の判定を行い, 有声音区間のみの相関係数の時系列を得る。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, この値を用いて二次の臨界制動系の逆フィルタを構成する。もとの各フレームの相関係数に畳み込みを行って, この逆フィルタによる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 12 次までの対数断面積比 (Log Area Ratio, LAR) $\{g_{it}\}_{i=1}^{12}$ からなる 13 次元ベクトル $\mathbf{X}_t$ の時系列に変換する。

$$\begin{aligned}\mathbf{X}_t &= (x_{1t}, x_{2t}, \dots, x_{13t}) \\ &= (g_{1t}, g_{2t}, \dots, g_{12t}, f_{0t})\end{aligned}\quad (1)$$

2.2 方法1の特徴抽出と認識方法

ベクトル $\mathbf{X}_t$ 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 $\boldsymbol{\mu}$, 標準偏差ベクトル $\boldsymbol{\sigma}$, 共分散行列 $\boldsymbol{\Sigma}$, および相互相関行列 $\mathbf{A}$ を計算する。

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

$$\begin{aligned}\boldsymbol{\Sigma} &= (\sigma_{ij}) = \left\{ \sum_{t=1}^N (\mathbf{X}_t - \boldsymbol{\mu})^i (\mathbf{X}_t - \boldsymbol{\mu})^j \right\} \\ &\quad / (N-1)\end{aligned}\quad (3)$$

$$\boldsymbol{\sigma} = (\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, \dots, \sqrt{\sigma_{13,13}}) \quad (4)$$

$$\mathbf{A} = (\lambda_{ij}), \lambda_{ij} = \sigma_{ij} / \sqrt{\sigma_{ii} \sigma_{jj}} \quad (5)$$

ここで N は, その単語音声における全有声音区間のフ

レーム数である。これらの統計量のうち, 標準偏差ベクトル $\boldsymbol{\sigma}$ および相互相関行列 $\mathbf{A}$ (以後, 両者をまとめて相関行列成分と呼ぶ) から, 話者を分離するうえで比較的有效性の大きい 20 成分を選択し, 平均値ベクトル $\boldsymbol{\mu}$ の全成分と組合せて認識のための特徴ベクトル $\mathbf{T}$ (33 次元) を構成する。この特徴選択は, 多数話者の学習用音声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 (F 比) にもとづいて行う。

話者認識の判定は, 入力音声の特徴ベクトルと各話者の標準パターンとの, 次のような重みつき距離によって行う。

$$D(\mathbf{T}, \mathbf{R}_k) = |\mathbf{V}|^{1/2} (\mathbf{T} - \mathbf{R}_k) \mathbf{V}^{-1} (\mathbf{T} - \mathbf{R}_k)^t \quad (6)$$

ここで $\mathbf{T}$ は入力音声の特徴ベクトル, $\mathbf{R}_k$ は登録話者 k の標準パターンで, その話者の学習サンプルの平均値である。 $\mathbf{V}$ は特徴ベクトルの共分散行列で, 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。

2.3 方法3の認識方法

方法3においては, 上述の方法1の場合のように統計的特徴に変換するようなことをせず, ベクトル $\mathbf{X}_t$ の時系列をそのまま用いて, 各話者の標準パターン時系列との DP (動的計画法) による非線形時間伸縮マッチングを行う。ここでは, 図2に示すように, 二つの時系列が作る平面上の対角線の両側に 10 フレームずつの整合窓を設けて, その中で対応づけを行い, DP パスの進む方

向は、この平面上で縦・横・斜めの3方向のみとした。対応づけるフレームどうしの距離には、方法1の場合と同様に、各成分の変動量にもとづいた重みつき距離を用いる。こうして最適に対応づけられたときの、全フレームにおける平均距離によって認識の判定を行う。

2.4 認識実験の方法

音声スペクトルの長時間にわたる変動の影響を考慮するため、学習サンプルの収集期間を10日間(短期間学習)と10カ月間(長期間学習)の二種類に変え、学習サンプルと入力音声の時期差を2~3日から5年ま

で変えて、話者9名の音声による話者識別と話者照合の実験を行った。各話者の学習サンプル数は、短期間学習で時期差が2~3日の場合のみ9サンプルであるが、他の条件の場合は12サンプルである。

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ンの話者による音声と判定し、話者照合の場合は、標準パターンとの距離があるしきい値よりも小さい音声は本人のものと見なし、他は本人の音声ではないとして棄却する。このしきい値は、便宜的に本人の音声を棄却してしまう誤り率と、他人の音声を受理してしまう誤り率が等しくなる値に、事後的に設定する。

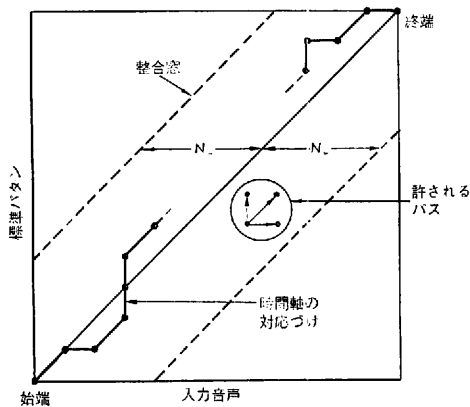
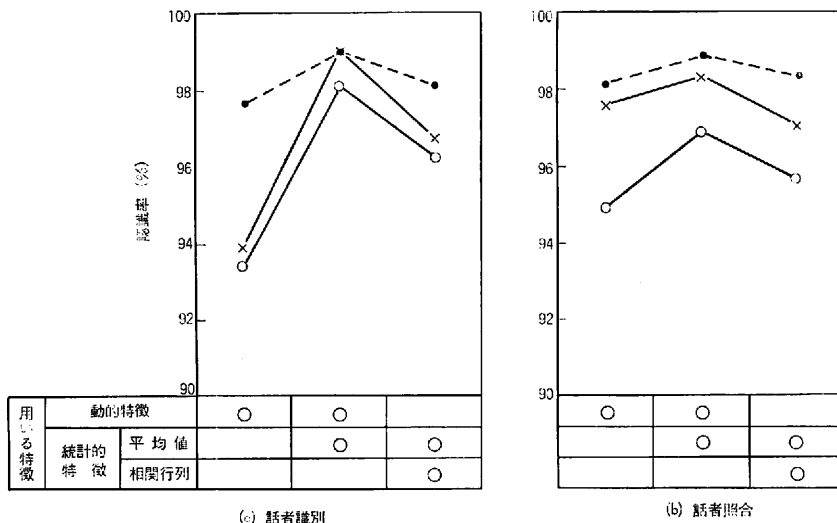


図2 方法3における入力音声と標準パタンの時間軸の対応づけ

2.5 認識実験の結果

2.5.1 長期間学習・時期差3カ月の場合

学習期間を長期間、学習サンプルと入力音声の時期差を3カ月としたときの、全話者の全入力音声に関する平均認識率を、図3に示す。この場合はスペクトル等化処理は行っていない。方法1(統計的特徴)と方法3(動的特徴)による場合のほか、統計的特徴の一部の平均値ベクトルの重みつき距離を、動的特徴による距離に組合せて用いる場合(重みつき平均)の結果も示した。処理が煩雑になるので、この組合せの場合は相関行列成分は除外した。また、図3には、認識の判定に単語



(a) 話者識別

(b) 話者照合

—○— /namae/ —×— /baNgo:/ ---●--- 2単語組合せ
図3 方法1(統計的特徴)、方法3(動的特徴)およびその組合せによる認識結果(長期間学習・時期差3カ月)

を用いる場合と、二単語を組合せて用いる場合の結果が、あわせて示してある。二単語を組合せる場合は、それぞれの単語による距離の平均値によって、認識の判定を行った。各実験条件における入力音声サンプルの総数は216である。

図3の結果から、次のようなことがわかる。

(1) 話者識別に関しては、方法3(動的特徴)よりも方法1(統計的特徴)の方が認識率が高いが、話者照合に関してはほとんど差がない。全体的には、両者の有効性の差は小さいが、方法1の方が方法3の1/10の程度の計算量と記憶容量ですむので有利であると言える。

(2) 統計的特徴と動的特徴を組合せると、誤り率が単独の特徴の場合の約1/2に低下する。

(3) 二種類の単語を組合せて用いると、誤り率が一単語の場合の約1/2に低下する。

(4) 二種類の単語の二種類の特徴を組合せて用いると、誤り率が一単語の単独の特徴の場合の約1/4に低下する。

2.5.2 短期間/長期間学習・時期差が10カ月までのとき

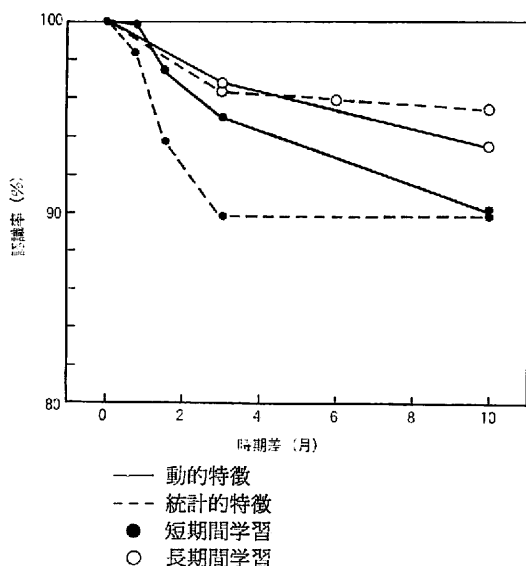


図4 方法1(統計的特徴)と方法3(動的特徴)による話者照合における時期差と認識率の関係(スペクトル等化なし、1単語による認識)

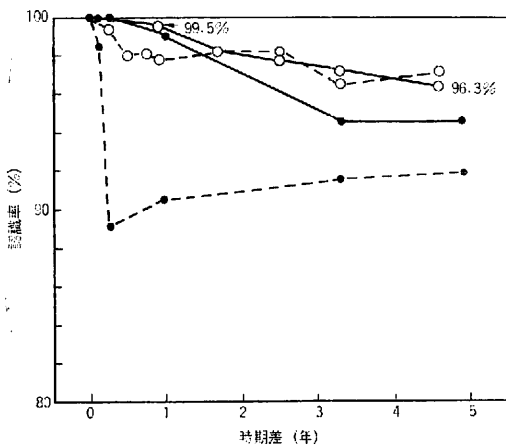
短期間学習と長期間学習のそれぞれの場合について、学習サンプルと入力音声の時期差を10カ月まで変えたときの、一単語による話者照合の結果を図4に示す。
/namae/ と /baNgo:/ のそれぞれの単語による結果を平均した値を示した。この場合もスペクトル等化処理は施していない。この結果から、短期間学習、長期間学習いずれの場合も、認識率が時期差とともに低下する傾向があるが、短期間学習の場合に特に、その傾向が著しいことがわかる。短期間学習の場合について、統計的特徴と動的特徴の結果を比較すると、時期差が3カ月程度までの範囲では、統計的特徴の場合の方が、時期差による認識率の低下が著しいが、時期差が10カ月程度になると両者の特徴による差は小さくなる。逆に言いかえると、統計的特徴の場合は、時期差が3カ月以上になると認識率がほぼ安定化するが、動的特徴の場合は認識率が時期差とともに単調に低下する。また、統計的特徴よりも動的特徴を用いる場合の方が、学習期間の長さによる認識率の変化が比較的少ない。前項の(1)でも述べたように、長期間学習の場合は、統計的特徴と動的特徴はほぼ同程度の認識能力を有する。

2.5.3 短期間/長期間学習・時期差が5年までのとき

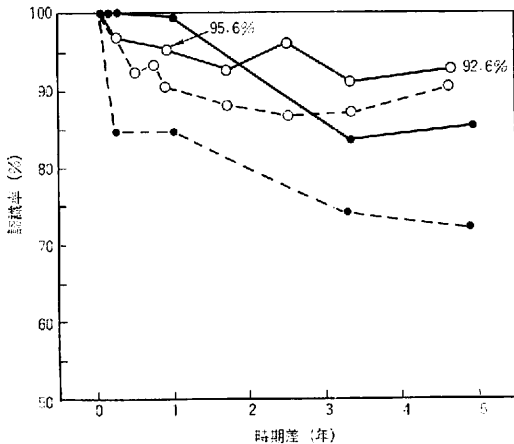
方法1で二単語の特徴を組合せて用いる場合について、学習サンプルと入力音声の時期差を5年まで延長して、認識実験を行った。短期間学習と長期間学習の両者の場合について、スペクトル等化処理を行った場合と、行わなかった場合の比較を行った。

実験の結果は図5に示すとおりで、学習期間が短期間・長期間いずれの場合も、スペクトル等化処理は、時期差による認識率の低下をさきえるうえで有効であり、特に短期間学習の場合にその効果が大きい。時期差が1年半程度以内の場合は、スペクトル等化処理により、短期間学習でも長期間学習のときと同程度の認識率を得ることができる。時期差がこれ以上長い場合は、長期間学習を行ってスペクトル等化処理を施すのが有効で、時期差が5年程度にまで長くなっても、この方法により、話者照合で96.3%、話者識別で92.6%の認識率を得ることができる。

なお、動的特徴の場合におけるスペクトル等化処理の



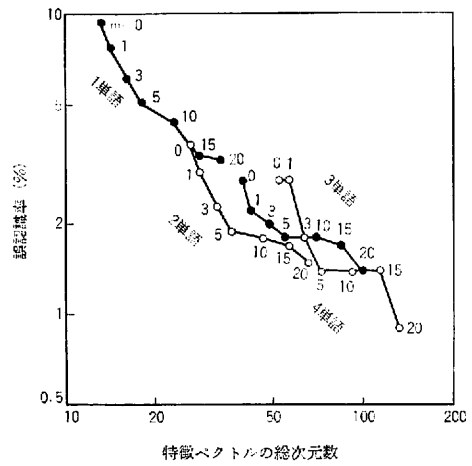
(a) 話者照合



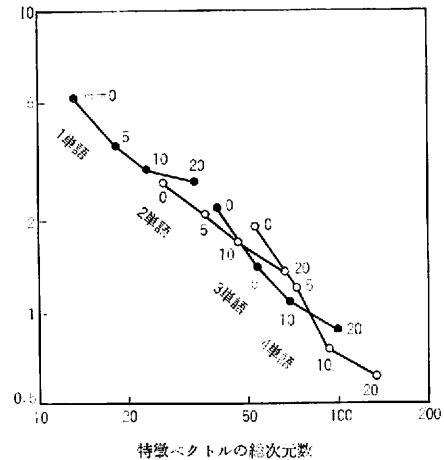
(b) 話者識別

- スペクトル等化あり
 --- スペクトル等化なし
 ● 短期間学習
 ○ 長期間学習

図5 方法1(統計的特徴)による話者認識における時期差と認識率の関係(2単語組合せによる認識)



(a) 話者識別



(b) 話者照合

m : 相関行列から選択して用いる成分の数

図6 方法1(統計的特徴)における特徴ベクトルの総次元数と誤認識率の関係(長期間学習・時期差3カ月・詐称者103名)

効果に関する検討は行っていないが、スペクトル等化処理を施さない場合に関しては、時期差が5年程度までの範囲で、統計的特徴(スペクトル等化処理なし)の場合と同様の結果が得られている。

2.6 方法1で4単語を組合せた場合

これまでの実験に用いた二単語に、/ko:geN/ と /bakuoN/ の二単語を加えて、認識の判定に用いる単語数を一単語から四単語まで変え、相関行列成分から

選択して特徴ベクトルにふくむ成分の数を0から20まで変えて、方法1による認識実験を行った。学習期間は長期間(10カ月間)、学習サンプルと入力音声との時期差は3カ月とし、詐称者(棄却すべき音声の話者)として、9名の登録話者以外の男性103名の音声を用いた。登録話者の入力音声の総サンプル数は216である。

認識の判定に用いる特徴ベクトルの総次元数と誤認識率の関係を図6に示す。両者の間には、両対数平

面上では線形な関係があり、相関行列成分から20成分を選択して用い、4単語を組合せれば、話者照合で99.3%、話者識別で99.1%の認識率を得ることができる。

3 方法4による話者照合

3.1 特徴抽出と認識方法

音声スペクトルの動的特徴を用いる認識方法(方法3)の改良として、単に音声スペクトル時系列の非線形時間伸縮マッチングを行うだけでなく、動的特徴を積極的に抽出して用いる次のような方法を検討した。本方式は特に、電話音声に対しても良好に動作することを目的として構成されている^{(17)~(20)}。

音声サンプルとしては、発声内容を固定した短文章音声を用い、図7に示すように、音声の始端と終端を検出したのち、高域強調を施し、LPC分析(線形予測分析)をへて、ケプストラムの時系列に変換する。標準化周波数は6.67 kHz、フレーム長は30 ms、フレーム周期は10 ms、分析次数は10である。音声区間全体にわたって各ケプストラムの値を平均化し、この値を

各時点のケプストラムの値から減ずることによって、伝送歪を除去し、長期間にわたるスペクトルの話者内変動の影響を低下させる。この操作は、前章で述べたスペクトル等化処理と類似した効果を有している。

一方、各ケプストラムの時系列を、10 msごとに90 ms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。平均値を除去して正規化したケプストラムと1次と2次の展開係数のうち、話者を分離するうえで有効性の高い18種の成分の時系列をとり出し、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。選択する成分は、多数話者の学習サンプルを用いた分散分析によって決定する。なお、上述の正規化処理は、1次以降の多項式展開係数には影響をおよぼさないで、多項式展開は正規化処理を施す前の時系列に適用している。

時間伸縮マッチングには、板倉⁽²¹⁾によって提案された非対称形のDPマッチング法を、始端と終端を開放した形に変形した方法を用いた。音声区間の検出誤りの影響を低下させ、マッチングの精度を良くするため、マッチングを行うごとに、入力音声と標準パターの全長(総フレーム数)を比較し、全長の短い方を基準(Guideline)側として用いた。この背景については文献⁽²⁰⁾を参照されたい。図8に示すように、二つの時系列が作る平面の対角線の両側に20フレームずつの整合窓を設け、斜線を施した範囲内でDPマッチングを行う。このとき入力音

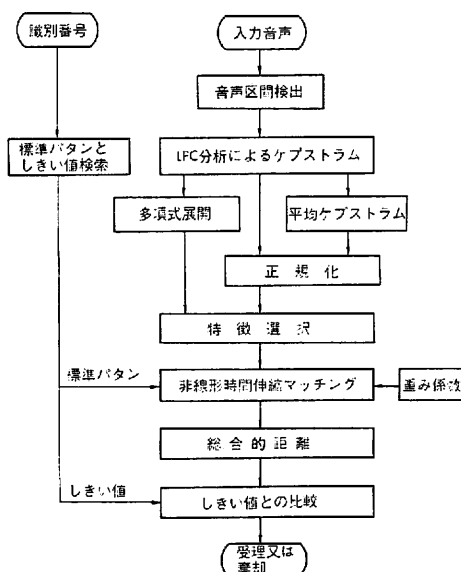


図7 方法4(動的特徴)による話者照合のブロック図

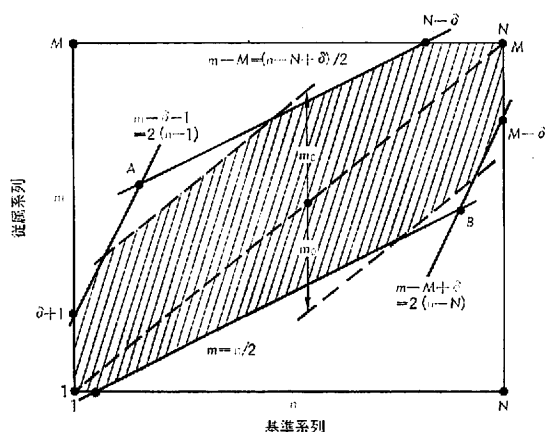


図8 方法4(動的特徴)の時間伸縮マッチングにおけるDPパス(時間軸対応づけ関数)の許容範囲

声と標準バタンのフレーム間の距離としては、方法3と同様に、各成分ごとにその安定性にもとづいた重みづけを施した距離を用いる。

こうして音声区間全体にわたる平均距離をもとめ、この値をあらかじめ各話者ごと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。このしきい値は、各話者の学習サンプルを用いて、異話者間の距離の分布を調べ、次のような実験式によって設定しておく。

$$\theta(k) = 0.6(\hat{\mu}_{DB}(k) - \hat{\sigma}_{DB}(k)) + 7 \quad (7)$$

ここで、 $\theta(k)$ は話者 k のしきい値、 $\hat{\mu}_{DB}$ は、話者 k の標準ボタンと異なる話者の音声のパラメータ時系列との距離の平均値、 $\hat{\sigma}_{DB}$ はその標準偏差である。(図9参照)

方法4の認識方法に関しては、話者照合実験のみを行い、話者識別実験は行わなかった。

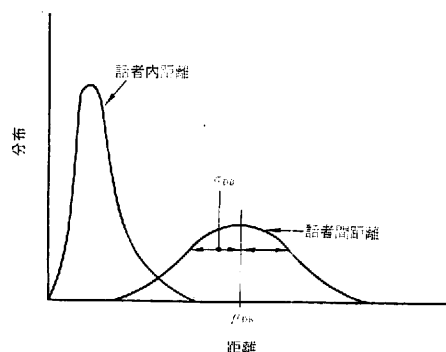


図9 話者内・話者間距離の分布の例

3.2 認識実験の方法

実験には、表2に示す三種類の音声サンプルセットを用いた。発声者はいずれも生まれたがらのアメリカ人で、男性は“‘We were away a year ago’”, 女性は“‘I know when my lawyer is due’”と発声した。音声サンプルセット1と2では、各話者は各文章を約2カ月にわたって1日2回(午前と午後)、計50回発声し、音声サンプルセット3では、各話者は、同じく約2カ月にわたって1日に1回ずつ、計26回発声した。各詐称者の発声回数はいずれも1回である。

各話者の標準ボタンは、図10に示すように、まず初めの5サンプルを相互に時間伸縮マッチングしたのち平均化して作成し、その後7回発声することにより、その直前の5回の音声を用いて更新した。話者照合の判定のしきい値も、前節のアルゴリズム(式(7))にもとづいて、同

表2 3種類の音声サンプルセットと誤認識率

音声サンプルセット		No. 1	No. 2	No. 3
登録話者数 + 詐称者数		男性10名 + 男性40名	女性10名 + 女性40名	男性21名 + 男性55名
録音条件		電話系	電話系	高品質 マイクロホン
しきい値 設定方法	固定しきい値	0.30%	0.42%	1.03%
	推定しきい値	0.19%	0.36%	0.77%
	最適しきい値	0%	0.06%	0.64%

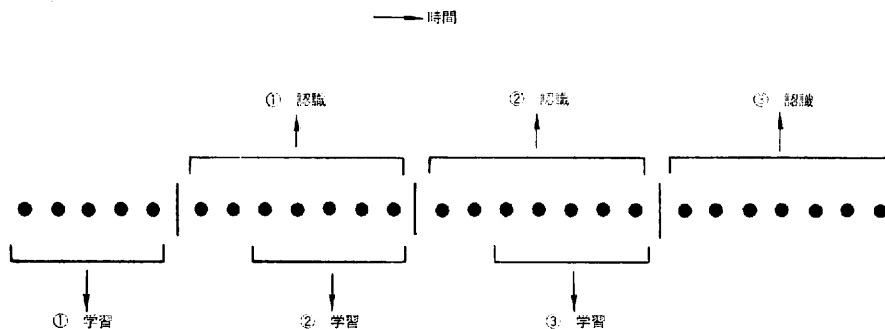


図10 方法4(動的特徴)による認識における標準ボタンと話者照合の判定のしきい値の更新方法

時に更新した。このことにより、声質の時期的変動による認識率の低下を防いでいる。

3.3 認識実験の結果

正規化したケプストラムと多項式展開係数の各時系列について、音声サンプルセット1の一部を用いて有効性の評価を行った結果を図11に示す。ケプストラムの時系列について一次の展開係数が有効であり、二次の展開係数の有効性はやや小さい。

各音声サンプルセットを用いて行った認識実験の結果を表2に示す。比較のために、話者照合の判定のしきい値を、事後的に最適値に設定する場合や、すべての話者に共通の値にあらかじめ設定しておく場合についても実験を行った。各条件について、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の平均値を示した。表2の結果から、三種類の音声サンプルセットに共通して、式(7)によってしきい値を話者ごとに推定すれば、事前に設定しても、1%以下の低い誤り率を得ることができることがわかる。しきい値をすべての話者に共通の値に設定すると、誤り率がやや増加する。

これまでの実験条件では、学習サンプルと入力音声の時期差は平均6口程度であるが、この時期差を6週間程度までの間で変えて認識実験を行ったところ、この範囲内では時期差による誤り率の増加はないことがわかった。さらに時期差を延長した場合については、実験の必要があろう。

ここで検討した方法4の認識方法は、伝送系の変動にも強いことが実験的に確認されている⁽²⁰⁾。本方法は、

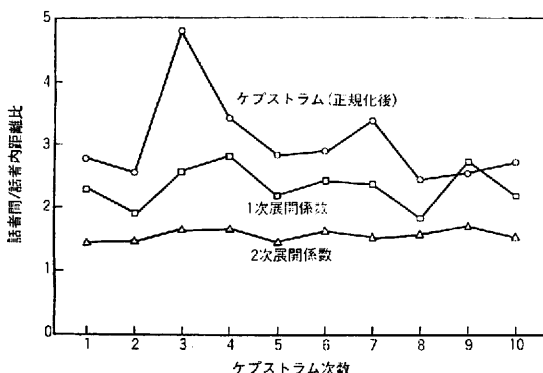


図11 各パラメータ時系列の有効性の評価結果

男女計120名の実際の構内回線を通った電話音声を用いた。6カ月にわたるオンラインの評価実験によって、有効性と安定性が確認されているが、この詳細は別の機会に報告する。

4 方法2による認識

4.1 特徴抽出と認識方法

音声スペクトルの統計的特徴を用いた認識方法(方法1)の改良として、図12に示すように、分析法をLARからLPCケプストラムにおきかえ、方法1で用いた統計的特徴に、各単語の有声音区間における各パラメータ時系列のフーリエ・コサイン展開係数 F_u を加えて、その効果を調べた。

$$F_u = \frac{1}{N} \sum_{t=1}^N x_{it} \cos \frac{\pi(t-1)l}{N} \quad (8)$$

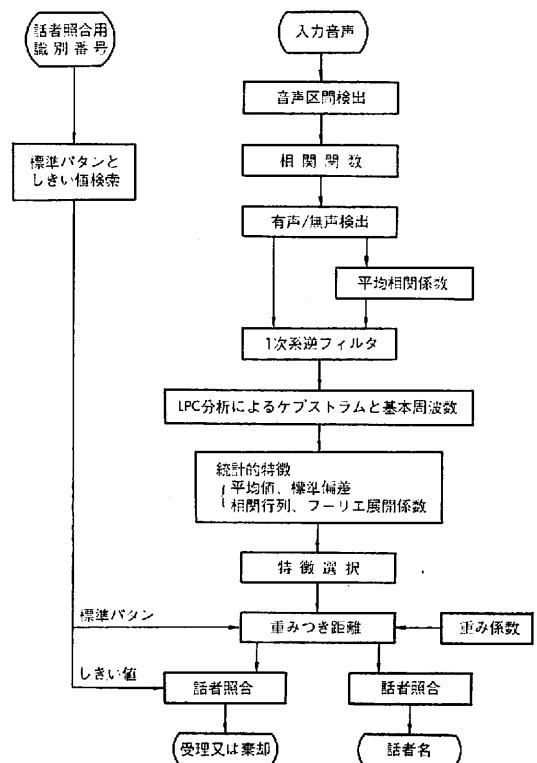


図12 方法2(統計的特徴)による話者認識のブロック図

ここで、 x_{it} は時刻 t ($1 \leq t \leq N$) における i 番目の特徴パラメータ (1~10 次の LPC ケプストラムまたは基本周波数)、 N は総フレーム数である。LPC 分析における標本化周波数は 6.67 kHz、フレーム長は 30 ms、フレーム周期は 10 ms、スペクトル等化用逆フィルタは一次系とした。

フーリエ展開係数は 1 次から 6 次まで計算し ($1 \leq l \leq 6$, 0 次は μ 成分と同じ)、 μ 成分 (平均値) および相関行列成分とあわせて、認識のための特徴パラメータとして用いた。方法 1 と同様に、フーリエ展開係数および相関行列成分のそれぞれについて話者要因に関する分散分析を行い、両者から計 60 成分を選択して、 μ 成分と組合せた。話者照合の判定のしきい値に関しては、事後的に最適値に設定する場合と、事前にすべての話者に共通の値に設定する場合の検討を行った。

4.2 認識実験の結果

節 2.6 と同様に、長期間学習、時期差 3 カ月、詐称者 103 名の条件で、4 単語を組合せて用いたときの認識誤り率を、LAR を用いる場合 (方法 1 に LAR のフーリエ展開係数を組合せた場合) の結果とあわせて、表 3 に示す。この結果から、LAR より LPC ケプストラムの方が有効で、ここで検討した方法により、99.9% 以上の精度で話者認識を行うことができることがわかる。

4.3 電話音声による認識実験

ここで検討した方法を、カーボン送話器をふくむ電話機と 0.4 mmφ-7 dB の擬似加入者線を通った音声に適用して、有効性を調べた。学習は 6 カ月おきの 3 時期の音声サンプルを用いて行い、入力音声との時期差は 6 カ月とした。登録話者は男性 9 名とし、詐称者のサンプルとして、他の 11 名が 3 回ずつ発声したものをを用いた。実験に用いた登録話者の入力音声サンプルの総数は 54 である。

同時に録音したマイクロホン音声を用いた同じ条件による実験の結果とあわせて、認識誤り率を表 4 に示す。学習サンプル数と入力音声との時期差の点で、節 4.2 の実験よりも実験条件が厳しいので、誤り率が全般的にやや大きくなっているが、本方式が、電話音声に対して

表 3 マイクロホン音声を用いた認識実験の結果
(長期間学習・時期差 3 カ月・4 単語組合せ)

分析方法	認識条件		誤認識率
LPC ケプストラム	話者識別		0 %
	話者照合	固定しきい値	0.10%
		最適しきい値	0.02%
LAR	話者識別		0.46%
	話者照合	固定しきい値	1.21%
		最適しきい値	0.51%

表 4 電話・マイクロホン音声を用いた認識実験の結果

(長期間学習*・時期差 6 カ月・4 単語組合せ)
(分析方法: LPC ケプストラム)

録音条件	認識条件		誤認識率
電話系	話者識別		0 %
	話者照合	固定しきい値	2.44%
		最適しきい値	0.49%
高品質 マイクロホン	話者照合		0 %
	話者識別	固定しきい値	2.44%
		最適しきい値	1.36%

* 6 カ月おき 3 時期

もマイクロホン音声とほぼ同等の性能で動作することがわかる。

本方式は、方法 4 と同様に、構内交換機と回線を通った実際の電話音声を用いたオンライン実験により、その有効性が確認されているが、この詳細は別稿にゆずる。

5 おわりに

発声内容を限定した話者認識方法として、筆者らがこれまでに検討した、動的特徴と統計的特徴を用いる四種類の方法的比較評価を行った。まず初めに、基本的方法として、統計的特徴を用いる方法 1 と動的特徴を用いる方法 3 の比較を行った結果、両者は同程度の性

能を有し、計算量と記憶容量の点から方法1の方が有利であることがわかった。この方法を用い、スペクトル等化処理を施せば、学習サンプルと入力音声の時期差が大きい場合でも、高い認識率を得ることができる。また、計算量と記憶容量はふえるが、統計的特徴と動的特徴の組合せが認識率の向上に有効であること、方法1、方法3のいずれの場合も、複数の単語にふくまれる特徴の組合せが有効であること等が確認された。

次に、動的特徴を用いる改良された方法4と、統計的特徴を用いる改良された方法2の有効性を調べた結果、いずれも方法3および方法1よりも大幅に高い認識率を示し、電話音声に対しても良好に動作することが確認された。方法2と方法4は共通の条件で実験を行っていないので厳密な比較はできないが、ほぼ同程度の性能を有すると言えよう。認識に必要な計算量と標準パタンの蓄積に必要な記憶容量には、方法2と方法4の間に、方法1と方法3の比較の場合と同様に、ほぼ1:10程度の差があるので、性能が同程度であるなら、方法2の方が有利であると言える。

今後は、さらに話者の数を増やして長期間にわたる実験を行うとともに、市外回線系や雑音の影響という点から、方法2と方法4の有効性の比較を行い、さらに方法の改良を進めて行く予定である。

謝 辞

本研究に関し御指導、ごべんたつをいただいた、東京農工大学野田教授(もと基礎研究部長)、基礎研究部畔柳部長、日本通信技術株式会社小池氏(もと第四研究室長)、第四研究室板倉室長ならびに東京大学斉藤教授(もと特別研究室長)に感謝する。

また、章3の研究に関し御指導いただいた、アメリカベル研究所 Flanagan, Rosenberg 両博士に感謝する。

日ごろ熱心に討論していただく、第四研究室の諸氏に感謝する。

文 献

- (1) 古井：話者認識，電四連大，33-6，p.5-129，1980。
- (2) B. S. Atal：Automatic Recognition of Speakers from Their Voices，Proc. IEEE，64，No.4，p.460，1976。
- (3) A. E. Rosenberg：Automatic Speaker Verification；A Review，Proc. IEEE，64，No.4，p.475，1976。
- (4) Just a Phone Call Will Transfer Funds，Electronics，Jan. 27，p.53，1981。
- (5) 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，No.10，p.549，1972。
- (6) 古井・板倉・斎藤：音声の長時間平均スペクトルに含まれる個人性情報，通研実報，23，No.6，p.1199，1974。
- (7) 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，56-A，No.11，p.717，1973。
- (8) 古井・板倉：音声スペクトルの統計的パラメータに含まれる個人性情報，通研実報，24，No.5，p.959，1975。
- (9) S. Saito and S. Furui：Personal Information in Dynamic Characteristics of Speech Spectra，Proc. 4th IJCPR，p.1014，1978。
- (10) 古井・斎藤：音声スペクトルの動的特徴を用いた話者認識，通研実報，29，No.7，p.1263，1980。
- (11) 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，57-A，No.12，p.880，1974。
- (12) 古井：音声の個人性パラメータの時期的変動の解析，通研実報，25，No.2，p.425，1976。
- (13) 古井：話者認識と音声認識におけるスペクトル変動の影響の比較，音学春季講論，3-2-7，1977。
- (14) 古井：統計的特徴と動的特徴による話者認識の比較，音学音声研資，S80-23，1980。
- (15) S. Furui：Effects of Long-term Spectral Variability on Speaker Recognition，96th ASA，NNN-28，1978。
- (16) 古井：統計的特徴と動的特徴による話者認識の比較，音学春季講論，1-7-2，1980。
- (17) S. Furui：New Techniques for Automatic Speaker Verification Using Telephone Speech，98th ASA，R-4，1979。
- (18) S. Furui and A. E. Rosenberg：Experimental Studies in a New Automatic Speaker Verification System Using Telephone Speech，ICASSP 80，p.1060，1980。
- (19) 古井：ケプストラムを用いた電話音声用話者照合システム，音学春季講論，1-7-1，1980。

-
- (20) S. Furui : Cepstral Analysis Technique for Automatic Speaker Verification, IEEE Trans. **ASSP-29**, No. 2, p. 254, 1981.
- (21) F. Itakura : Minimum Prediction Residual Principle applied to Speech Recognition, IEEE Trans., **ASSP-23**, No. 1, p. 67, 1975.

(1981, 4, 3 受付)