氏名	二矢田勝行
授与学位	工学博士
学位授与年月日	昭和63年3月11日
学位授与の根拠法規	学位規則第5条第2項
最終学歴	昭和47年3月 東北大学大学院工学研究科電気及通信工学専攻 修士課程修了
学位論文題目	**高速処理を目指した不特定話者・多数語用音声認識法の研究**
論文審査委員	東北大学教授 城戸 健一　東北大学教授 木村 正行 東北大学教授 高木 相　東北大学助教授 牧野 正三

論文内容要旨

音声認識装置に関する研究は古くから行なわれており，また社会からも期待されているが，まだ本格的な利用には程遠い状態である。その主な理由は，現状の装置が「使用上の制限があまりにも大きい」ためであると考えられる。現在実用化されている装置は，登録した人の声のみしか認識できない特定話者用がほとんどである。不特定話者用の装置は，認識できる語彙数が16～32と少ない。また語彙の変更が容易でなく，柔軟なシステム構成が困難である。このように，現状の認識装置は人間にとって決して使いやすい装置であるとは言えない。

本研究は，これらの制限を緩和して，より使いやすい装置の実現を目的とし，次の具体的な目標を掲げて行なった。

①不特定話者，200単語に対して95%以上の認識性能を有し，3000単語程度の大語彙を取り扱い得るようにすること。

②認識対象単語を容易に追加，変更できるようにすることによって，タスク構成上の柔軟性を持たせ，音声認識装置を幅広く使えるようにする。

③騒音条件など環境変動に対して影響を受けにくに方式を採用する。

④小型・高速化が可能なアルゴリズムを用いる。

これらの目標を達成するための基本方式として，音素を認識単位とする方式を採用した。

第2章では，音素認識法の開発用およびアルゴリズムの評価用として用いる音声データベースについて述べた。前者については，音韻バランスを考慮して選定した212単語集団のデータを男100

名，女50名分収録した。後者は，駅名3138単語を男女各30名分収録した。このように，本研究は十分なデータベースを用いて行なわれている。

第3章では，パターンマッチングを基本とする音声認識法において最も重要な事項である，特徴パラメータと距離尺度について検討を行なった。

(1) 特徴パラメータとして，自己相関係数，PARCOR係数，LPCケプストラム係数，フォルマント周波数，バンドパスフィルタ出力を選び，距離尺度として，ユークリッド距離，重み付きケプストラム距離，相関距離，ベイズ判定，マハラノビス距離，簡易型マハラノビス距離（共分散行列を全カテゴリーに対して共通とし，マハラノビス距離を一次判別式に展開した距離），LPC距離（WLR）および判別図を用いて，これらの特徴パラメータと距離尺度を組合せて，識別実験を行なった。その結果，LPCケプストラム係数とベイズ判定の組合わせが最も良く，212単語中の母音区間全フレームに対して，男声85.2%，女声83.1%の識別率を得た。また，簡易型マハラノビス距離を用いると，識別率が1.5～2%ほど低下するが，計算量は大幅に削減できる。パラメータの次元数が大きい場合，高速化の観点から，簡易型マハラノビス距離は有効な距離尺度である。

(2) 白色ノイズおよび－6dB/octの特性を有するノイズを混入したデータに対しても，相対的にはLPCケプストラム係数とベイズ判定または簡易型マハラノビス距離の組合せが良好な結果となることがわかった。しかし，ノイズが識別率に与える影響は大きく，特に白色ノイズの影響は著しい。その対策として，ノイズ特性をシミュレートして原音声に加えたデータを用いて，標準パターンを作成する方法を提案した。

(3) 不特定話者用の標準パターン作成に要する人数を見極めるために，標準パターン作成用の人数を徐々に増しながら，話者に関してオープンとクローズの識別実験を行ない，識別率と標準偏差の変化を求めた。その結果，実用的には50～60名以上の人数が必要であると判明した。

(4) 特徴パラメータの個数を減らすことによって計算量の削減を図るために，LPCケプストラム係数の打切り次数と識別率の関係を評価した。その結果，10次でほぼ飽和し，13次で完全に飽和することが判明した。

(5) 母音スペクトルの男女差は他の音素群に較べて大きい。これに対処するには，男女のデータを混合して単一の標準パターンを作成するよりも，男女各々の標準パターンを作成して，マルチテンプレートとして用いたほうが良い結果が得られることがわかった。

第4章では，子音のセグメンテーション法と大分類法の検討を行ない，セグメンテーションと大分類を同時に行なうことができる，高速化に適した簡易な方法を提案した。

(1) 騒音状態，発声のしかたなどの変動要因の影響を軽減するために，セグメンテーションと大分類用には，音素の識別に用いるパラメータとは異なる，単純なパラメータを用いる。具体的には，低域（250～600Hz），高域（1500～4000 Hz）の帯域フィルタの対数パワーの時間変化を用いる。

(2) 低域および高域のパワーディップの大きさを求め，これらを2次元の判別図に適用することによって，語中子音のセグメンテーションと大分類を同時にしかも極めて簡単に行ない得ることを

示した。この方法を主として用い，それに鼻音性および無音性情報を併用した，語中子音のセグメンテーションと大分類法を開発した。評価の結果，セグメンテーション正解率96.3%を得た。そして，無声破裂音群98.7%，有声破裂音群85.2%，鼻音群96.9%，無声摩擦音群88.3%，全平均で94.1%の大分類率を得た。

(3) 無声性情報，鼻音性情報，パワー変化，スペクトル変化を用いた語頭子音のセグメンテーションと大分類法を提案し，評価結果を示した。正確率は94.6%，母音に対する付加率は33.4%である。

第5章では，パターンマッチングに基づく子音識別法（細分類法）の検討を行ない，識別条件を最適化した。そして，自動認識に適した実用的な子音識別法を提案した。

(1) 判別効率を用いた検討によって，子音の識別に有効な情報が存在する時間的位置を調べた。

(2) スペクトルの動的な特徴を用いた子音識別法を7種提示し，比較検討を行なった。その結果，LPCケプストラム係数の時系列情報を用い，統計的距離尺度によって標準パターンとの類似度を計算する方法が最良であることを示した。語中の有声子音（/m/,/n/,/η/,/b/,/d/,/r/,/z/）に対して，平均識別率77.7%が得られた。

(3) 無声破裂音（/p/,/t/,/k/,/c/），有声破裂音（/b/,/d/,/g/），鼻音（/m/,/n/）の識別方法の基礎的な検討を行い，子音識別に有効なスペクトルおよび時間的な特徴について述べた。

これらの音素群の特徴は，破裂時点付近の3フレーム程度の区間にあり，その部分のスペクトルの大局的特徴（LPCケプストラム係数の低次の係数）が識別に寄与することなどを示した。最適条件において，無声破裂音88.1%，有声破裂音86.4%，鼻音83.1%の識別率を得た。

(4) 第4章の方法で大分類された各音素群の細分類条件の最適化実験を行なった。語中子音の識別には，フレーム数：12程度，パラメータの打切り次第：8次（C_0～C_8）程度が必要であり，語頭子音では，無声子音群：11フレーム，有声子音群：6フレーム程度の時間長が必要である。

(5) 簡易型マハラノビス距離を事後確率化することによって，子音区間の検出ずれに対して強く，しかも計算量の少ない距離尺度を提案した。

パワーディップを用いて子音の概略の位置を検出し，その前後数フレームの間で上記の距離尺度を用いた類似度計算を行ない，類似度が最大となる音素を求める方法によって，子音の自動認識ができる。

第6章では，第3～第5章で確立した要素技術を中心として，それに他の構成要素を加え，高速処理に適した音声認識アルゴリズムを構成して評価を行なった。

(1) 本アルゴリズムはボトムアップ処理を基本とした構成である。まず音声区間の始端を検出した後，母音，半母音・拗音，語中子音，語頭子音を個々に認識する。音声区間の終端を検出すると，日本語の音形規則に基づいて，音素認識結果を音素系列に変換する。そしてこの音素系列と辞書中の各項目を照合比較し，最も類似度の高い辞書項目を結果とする。

このアルゴリズムは時間的な後戻り処理が無く，高速処理を要する部分と，複雑な処理の部分が分離されているなど，ハードウェア化に適した構成になっている。

(2) 上記のアルゴリズムを評価した。その結果，音素認識率81.3％（母音：90.6％，子音：71.9％，半母音・拗音：78.0％，脱落率：4.1％）を得た。また単語認識率は，212単語：95.7％，274単語：95.6％の結果を得ることができ，200単語95％以上という本研究の目標を達成した。話者に対する依存性や単語集団に対する依存性は小さく，本アルゴリズムが不特定話者・多数語用として適していることを示した。そして，他のシステムの結果と較べても認識率が高いことを述べた。

(3) 複数の発声を用いて，階層的に認識する方法による，多数語彙への対処法を示した。そして，3000単語程度からなる駅名を2階層で認識する具体的な方法を述べた。

第7章では，第6章で述べたアルゴリズムを用いて，小型・高速の音声認識装置を実現し，本アルゴリズムが小型化・高速化に適していることを実証した。また，モデルタスクを設定して，音声認識システムを構成し，音素を基本単位とする方式の柔軟性など，使用感について言及した。

(1) 計算量の削減を目的として，サンプリング周波数と認識率の関係を評価した。その結果，12 kHz の場合に較べて，10 kHzでは音素認識率0.7％，単語認識率0.2％の低下に留るが，8 kHzでは，それぞれ2.5％および1％ほど低下する。この結果に基づき，10kHzサンプリングの装置を作成することにした。

(2) 2つのDSPと1つのマイクロプロセッサを用いて，小型・高速化を実現する手法の概要を述べた。音響処理と類似度計算にDSPを1つずつ割当て，その他の処理はマイクロプロセッサで，パイプライン制御によって行なっている。

(3) ハード化した装置の評価を行なった結果，音素認識結果はほぼリアルタイムで出力され，単語認識結果は発声終了後0.8秒で出力されることがわかった（200単語の辞書の場合）。そして，音素認識率は第6章のシミュレーションシステムとほぼ同等であった。したがって，速度，認識率ともに，目標仕様を満足していることを確認した。

(4) 作成した認識装置を用いてモデルタスクを構成し，使いやすさの面からの評価を行なった。辞書内容の追加・変更の容易さが本方式の最大の長所であるという実感を得た。

定常騒音に対しては比較的頑強であるが，非定常な背景雑音や話者自身が発する雑音による音声区間検出の誤動作があり，今後の課題である。

本研究の意義は，第1に，実用をイメージできる本格的な不特定話者・多数語用のアルゴリズムおよび装置を，初めて実現したことである。これまでに，不特定話者・多数語の認識を目指した要素技術の研究や，シミュレーションシステムは幾つか発売されているが，高い認識率を確保でき，しかも小型でリアルタイム処理が可能な方式は開発されていない。

第2は，音素認識を基本とする方式を用いて装置を実現したことである。
音素を基本単位とすることの有用性（柔軟性や拡張性など）は，古くから指摘されていたが，これまでは，高い音素認識率を達成することや複雑な音素認識処理を高速化することは難しいものとされてきた。本研究によってリアルタイムの音素認識が可能となり，種々のタスクに対して柔軟に対処できる音声認識システムを構成できるようになった。また，リアルタイム音素認識技術は，今後の研究の本命である，音声対話技術，自動翻訳電話技術へとつながる重要な要素技術としての位置付けをもつ。

審　査　結　果　の　要　旨

音声は人間が日常使っている便利な情報伝達媒体であるために，音声によるマン・マシンインターフェイスの実現を目指した音声認識の研究は，多くの研究者によって活発に行われてきた。その結果，話者や認識対象に強い制約を課すシステムは実用化されるに至ったが，不特定の話者による多数の語彙の認識は未だに困難である。著者は，不特定話者の発声する数千語という多数の単語音声を，1単語当り1秒以内という短時間で自動認識するシステムの開発を目的として研究し，所期の成果を上げると共に，音声自動認識に多くの知見を加えた。本論文はその結果をまとめたもので，全文8章からなる。

第1章は序論である。

第2章では，本研究で用いるために収録した音声資料について述べている。この資料は，音声研究用として集められた従来の音声資料と比べて格段に，質・量ともに優れたもので，本研究の信頼性を高める一因になっている。

第3章では，認識の基本になる特徴パラメータと距離尺度を検討し，特徴パラメータとしてはLPC ケプストラム係数を，距離尺度としては統計的距離尺度を用いるのが有効であることを示している。さらに，計算量の少ない簡易型マハラノビス距離を提案し，その有効性を示している。

第4章では，発声条件や騒音などの変動要因に強いパラメータとして，低域・高域の対数パワーの変化情報を用いて子音のセグメンテーションと音楽群への大分類を行う方法を提案している。

第5章では，大分類された子音群から個々の子音を識別する方法を提案すると共に，高速化のための手段を検討している。

第6章では，前3章で開発した要素技術を基にして，時間的な後戻り処理のない単純な構成によって高速化を図る手法を提案し，音素認識率と単語認識率の評価を行い，この手法が不特定話者・多数語彙用として適していることを実証している。

第7章では，前章の手法による小型・高速の認識装置を試作し，これにより，本研究の目的が達成されたことを確認している。

第8章は結論である。

以上要するに本論文は，不特定話者・多数語音声の高速自動認識法の開発を目的として研究し，従来にない音声認識システムを実現すると共に，音声自動認識に有用な多くの知見を与えたもので，情報工学の進歩に寄与するところが少なくない。

よって，本論文は工学博士の学位論文として合格と認める。