| | |
|---|---|
| 氏名 | 松尾 広 |
| 授与学位 | 工学博士 |
| 学位授与年月日 | 平成2年3月28日 |
| 学位授与の根拠法規 | 学位規則第5条第1項 |
| 研究科，専攻の名称 | 東北大学大学院工学研究科<br>（博士課程）情報工学専攻 |
| 学位論文題目 | **持続時間モデルを利用した単語音声認識に関する研究** |
| 指導教官 | 東北大学教授 城戸 健一 |
| 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<br>東北大学教授 曽根 敏夫　東北大学助教授 牧野 正三 |

# 論文内容要旨

マン・マシン・インターフェースとしての音声入力装置は，データの入力に熟練を要しない等の理由から利用価値が高い。本研究は日本語の連続音声認識をめざし，音響処理部に必要な各種手法の有効性や問題点を明らかにするために，単語音声認識システムを通して種々の検討を行ったものである。

## 第1章 序論

本章は序論であり，音声認識の現状と問題点を明らかにし，本研究の意義と目的について述べた。ARPAでの反省から，連続音声認識システムにおいては言語レベルの知識の利用が不可欠であると同時に音響処理部の高精度化も重要であるという認識に基づいて，連続音声認識に拡張可能な単語音声認識アルゴリズムを検討することを述べた。

さらに連続音声認識への拡張を考えた場合，音素を認識単位とすること，ワードスポッティングによる単語認識であること，Top-down制御の必要があることを述べた。

## 第2章 単語音声認識システムの概要

本章では，単語音声資料および単語音声認識システムの全体像を示した。本研究ではBottom-upに音素群候補列中から単語候補をスポッティングし，単語候補に持続時間モデルをあてはめて

Top-down に認識を行う Bottom-up／Top-down 融合型の単語音声認識システムを提案した。

## 第 3 章　自動性別判定を伴った音素認識

本章では，音素の持つ音響的特徴の変動のうち，主に性別に起因する周波数構造の変動がどのように現われるかを音素認識実験から示し，母音において性差が顕著に現われることを示した。母音に現われる性差を除くため，まず性別を判定して標準パタンを切り替える方法について，認識実験からその効果を示した。

音素認識実験では，母音では成人を対象とした場合に男性と女性の 2 グループに分けることができ，男女別の標準パタンを使うことで認識率（加重平均）が 2.3%向上すること（有意水準 0.1%で統計的に有意）を示した。次に事前の情報無しで性別を判定する方法として，男女別の母音／撥音の標準パタンおよび摩擦音の標準パタンを使ってパワーが大きく安定した部分でフレームごとに認識を行い，男女どちらが多く選ばれたかにより判定するアルゴリズムを提案した。性別判定実験では正答率 98.7%を得た。自動性別判定を組み合わせた母音認識実験では性別が既知で男女別の標準パタンを使った場合とほとんど変わらない認識率が得られた。

## 第 4 章　セグメント特徴群の検出

認識の基本単位として音素を用いたときの問題点として，音素を表すスペクトルの時間的構造の変動をあげ，音素の時間的構造の変動を吸収するためにセグメント特徴を認識単位とすることを提案してきた。本研究では階層的に処理を行うため，セグメント特徴をまとめてセグメント特徴群とした。

本章では，従来のセグメント特徴検出法の欠点について述べ，セグメント特徴群の検出法の高精度化のためにダミークラスを追加した TSP フィルタ法を提案した。セグメント特徴の検出は TSP フィルタ法を基にした方法でセグメンテーションフリーに行われるが，従来の検出法では標準パタンの設計にかかわらなかった部分（中心フレーム以外の箇所）からの入力があったときに尤度が異常上昇するため，中心フレーム以外の箇所で付加が起こる現象が見られた。ダミークラスを追加した TSP フィルタ法は，従来標準パタンの設計に使われなかったパタンをクラスタリングによって数クラス（ダミークラス）にまとめて TSP フィルタ法に標準パタンの一部として加えることにより尤度の異常上昇を抑え，中心フレーム／非中心フレームの判別をするものである。セグメント特徴群のうち有声破裂音群／無声破裂音群／鼻子音群の認識から，ダミークラスを追加した TSP フィルタ法を用いることによりスキャニングした場合での認識率の低下を抑えることができることを示した。最後にセグメント特徴群全体を対象として検出実験を行い，認識率 88.0%，脱落率 6.4%，付加率 10.9%を得た。これは従来の検出法と比べると認識率／脱落率は大きく変わらなかったが，付加誤りが約 6 分の 1 に減少し，ダミークラスを追加することによりヒューリスティックな規則を使わずに付加を減少できることが示された。

## 第5章　Bottom－up認識部

本章では，セグメント特徴群系列から音素群候補を検出し，音素群候補中から単語候補をスポッティングするBottom-up認識部について述べた。

入力に対してまず性別判定（第3章）を行う。この時の判定結果は音素の標準パタンの切り替えに使われる。次にセグメント特徴群（第4章）と母音・子音定常部の認識が行われ，この結果から音素群候補が検出される。検出された音素群候補列中から単語候補が検出（スポッティング）される。

音素群の候補はセグメント特徴群の並びの順序が誤っていなければ，一部分脱落していても検出される。また音素群候補間の時間的なオーバーラップも許す。ここで音素群の時間的構造の変動が吸収される。単語候補の検出も同様に音素群の並びが一致していれば，付加や脱落を許すという条件で検出する。出現順序のみを重視しているので，付加／脱落誤りに強い検出ができる。Bottom-up認識部では候補の検出のみ行い，尤度の計算は行わない。

## 第6章　Top-down認識部

本章では，Bottom-up認識部で検出された単語候補に持続時間モデルをあてはめて単語中の音素位置を予測し，Top-downに認識するTop-down認識部について述べた。

音素系列からDPマッチングによって単語を認識する場合，音素系列中に付加／脱落誤りがあると不自然なマッチングをして誤りとなることがある。これは音素系列を求める際に時間情報が捨てられていることが原因の一つであり，時間情報を考慮することで誤りを防ぐことができると考えられる。

従来，日本語の持つ音素（音韻）の持続時間の特性は音声合成の分野では積極的に利用されてきた。これは持続時間の設定が合成音の自然性に関連が深いためであった。一方音声認識の分野においても持続時間の情報は使われてはいるが，極端な誤りを防ぐためであり，固定された基準による制限として用いていた。しかし音声はリズムを持っており，発生の全体的なテンポ（発話速度）の変動により持続時間が変動するため，固定された基準を用いるのは妥当ではなく，またBottom-up処理だけで音声の持つリズムを反映するのは難しい。

本研究では音素の持続時間を与えるモデルとして，単語中の母音の持続時間の平均長＝平均母音長を基準とする持続時間モデルを提案し，単語中の音素位置を予測してTop-downに認識する方法を示した。平均母音長による音素の持続時間モデルを用いることにより，発話速度の変動にも単語の持つリズムを保ちながら適応できる。持続時間モデルによる音素の持続時間の推定実験から，平均モーラ長を基準とする場合よりも平均母音長を基準とするほうが優れていることを示した。

単語中の音素位置は，単語候補とマッチングした音素群候補の位置と持続時間モデルから計算される位置ができるだけ一致するように持続時間モデルのパラメータを調整することにより求められる。単語候補と持続時間モデルとの適合度を示す指標として重なり率を定義し，重なり率を最大化するように持続時間モデルのパラメータを調整することによって音素位置の推定を行う方法を示した。このようにして持続時間モデルによっておおよその音素位置は推定できるが，あくまでも平均

的なものであり，実際の入力音声と比べると多少のずれがある。このためモデルから予測された音素位置をファジィDPによって微調整する方法を示した。微調整と同時に予備選択も行われる。

単語の尤度は単語内の音素の事後確率の対数値の和で表され，音素の事後確率は音素－事後確率フィルタにより求められる。単語の尤度の計算は微調整された音素境界の周辺に狭い時間窓を持つ連続スタックDPにより行われる。音素－事後確率フィルタはRoucosらのモデルの出力を事後確率になるようにしたもので，音素の区間全体から数フレームを取り出して（resamplingして）認識を行う。音韻群とその位置がはっきりしている場合にはフレーム間の相関も使うTSPが有利であるが，音素区間の全体の情報が扱えるということとすべての音素が同じ条件で比較できるという点で音素－事後確率フィルタを用いることにした。母音などの定常的な音素ではresampling点は等間隔に取るが，破裂音のような過渡的な音素では中心フレームを捕えることが重要であるため，定常的な音素とは違ったresampling点を持たせている。

さらに持続時間モデルが成り立つような持続時間を持つならば，予測だけでなく尤度情報として用いることができると考えられる。持続時間に関する尤度の計算法と持続時間の尤度を併用した場合の単語認識について述べた。単語の尤度計算時のDPパスをバックトレースして得られる音素境界とそこから求められた平均母音長から持続時間の尤度を計算し，持続時間の尤度に重みを付けて単語の尤度に加算した量から単語認識を行う。

本方式は音声区間を与える必要がないので連続音声認識に拡張可能であり，持続時間モデルによって音素位置を予測して区間を限定するためTop-down的な処理を有効に働かせることができる。

## 第7章　単語音声認識システムの評価

本章ではBottom-up認識部，Top-down認識部の各部分の評価結果と単語認識実験からシステム全体の評価を行った。Bottom-upによる単語候補検出とTop-downによる認識を融合した本方式は，Top-down的な処理を用いないで認識しようとする場合に比べ，音素位置を限定することから尤度計算の時間を大幅に節約できる。

音素－事後確率フィルタ出力からの単語認識実験では94.5%の単語認識率が得られた。これは同じデータを用いて音素系列から認識した場合の92.4%（ただしclose実験），線形伸縮による場合の85.3%よりも高い認識率（有意水準0.1%で統計的に有意）であり，本研究の有効性が示された。さらに持続時間の尤度を併用することにより最高95.2%の単語認識率を得たが，持続時間の尤度の併用そのものの効果の統計的有意性は確認できなかった。これは実験に用いた単語集合の性質によるものと考えられ，別な単語集合による認識実験など，さらに検討が必要である。

## 第8章　結　　論

本章は本論文全体のまとめである。

以上のように本論文では連続音声認識に拡張可能なアルゴリズムを持つ単語音声認識システムの検討を行い，平均母音長による音素の持続時間モデルを用いたTop-down認識システムが有効であるとの結果を得た。

# 審 査 結 果 の 要 旨

人間－機械間の情報伝達手段として，音声を使うことが望まれている。そのために音声認識の研究が活発に行われてきたが，発声速度の変動に起因する認識の困難さの問題は未解決であった。その解決のため，著者は，発声速度の変動による，各音素の持続時間の変動を説明するモデルについて研究し，それを基にして，連続音声認識へ拡張可能な種々の新しい認識方法を提案し，その有効性を検討した。本論文はその成果をまとめたもので，全編8章よりなる。

第1章は序論である。

第2章では，本研究で用いた単語音声資料，および持続時間モデルに基づく単語音声認識システムの概要を示している。

第3章では，音素の周波数構造が，話者の性別によってどのように変動するかを音素認識実験によって検討し，母音において，性差が顕著に現れることを示している。そこで，母音に現れる性差の影響を除くため，まず性別を判定して標準パタンを切り替える方法を提案し，認識実験によってその有効性を示している。これは実用上有用な結果である。

第4章では，音素の時間構造の変動を吸収するために，セグメント特徴を認識の単位として用いることと，その検出方法について述べている。音素の付加誤りを抑制する方法として，ダミークラスを追加したTSPフィルタ法を提案し，検出実験から，付加誤りを減少できることを示している。

第5章では，セグメント特徴群系列から音素群候補を検出し，音素群候補中から単語候補をスポッティングする，Bottom-up認識部について述べている。

第6章では，Bottom-up認識部で検出された単語候補に，音素の持続時間モデルをあてはめて認識するTop-down認識部について述べている。音素の持続時間を与えるモデルとして，平均母音長を基準とした持続時間モデルを提案しているが，このモデルは，発声速度の変動に対処できる初めての持続時間モデルである。

第7章では，システム全体の評価を行い，持続時間モデルの利用によって，従来よりも高い単語認識率が得られることを示している。

第8章は結論である。

以上要するに本論文は，発声速度の変動に対処できる持続時間モデルについて研究し，それを基に，持続音声認識へ拡張可能な種々の新しい認識方法を提案して，その有効性を明らかにしたもので，情報工学並びに音響工学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。