早稲田大学大学院理工学研究科

博 士 論 文 概 要

論 文 題 目

大語彙を対象にする
不特定話者単語音声認識の研究

申 請 者
松 浦 博
Hiroshi Matsuura

1994年 12月

音声は人間にとって最も自然で基本的なコミュニケーション手段であり，人間と機械とのインタフェースへの適用が試みられてきた．しかし，「誰の声でも」，「どんな言葉でも」，「自由な発話でも」認識する究極的な装置の実現は容易ではない．1970年代後半，得られた音響的特徴量を単語単位に標準パターンとマッチングするパターンマッチング法が，いくつかの認識条件の制約のもとに実用上の成果を得た．特定話者単語認識装置はＤＰマッチング法に基づき，「特定話者の」，「20から 200個程度の」，「孤立単語発声した」発話を対象とし，手がふさがっていたり，目を離すことができない用途において，キー入力に対する優位性を示した．不特定話者単語認識装置は「不特定話者の」，「16から32個程度の」，「孤立単語発声した」発話を対象とする．日本では1980年代初頭，電話音声認識応答システムの開発の中で不特定話者の10数字および少数の制御語を対象にした認識装置が実用化された．しかしながら，実用化は非常に限定された用途にとどまっている．

音声認識の応用が期待されている分野として街頭の情報案内装置，銀行の自動振込装置や駅の券売機などの社会情報分野があげられる．特に，対象とする認識語彙が大きい場合には音声入力がキー入力に勝る．これら社会システムは不特定多数の利用者が入れ替り操作するため，話者適応によらない完全な不特定話者認識方式が必須である．ところが，不特定話者の発声した1000単語以上の大語彙単語を確実に認識するような技術は，いまだ確立されてはいない．以上のような観点に基づき，本研究では従来困難とされてきた不特定話者の発声した大語彙単語を対象とする認識方式について検討を行う．次に，いくつかの検討から得られた成果を基に大語彙単語認識システムの構築を試みる．このような認識システムを実現するためには，

①話者に大きく依存しない特徴量を抽出する．
②大語彙を対象とするために最適な認識の単位を決定する．
③話者変動を効果的に吸収する認識手法を確立する．
④音素などの細かな単位に対する識別によって得られた曖昧性を含む記号列から，曖昧性を有効に活用しながら単語レベルの認識を行う．
⑤学習は話者間の変動を吸収し，曖昧性を許容する方式であり，必要とするデータ数や処理時間が実用的範囲に収まる．
⑥現在のハードウエア技術で実時間認識が可能である．

等の課題を解決しなければならない．本論文ではこれらの課題に対して，いくつかの解決策を与えることを試みる．また，応用によっては話者適応が可能な場合があるので，若干の検討を加える．

第２章では，音声パターンの発話変動，すなわち時間軸方向の変動のみならず，周波数方向の変動も含む話者変動を効果的に吸収する認識手法として統計的パターンマッチング手法の一つである複合類似度法について述べる．認識実験から，フレーム単位のスペクトルパターンより次元数が大きく時間方向の情報を含むスペクトルー時間パターンを用いる方が，高い認識率が得られることを示す．

第３章では，複合類似度法を単音節認識に適用する．認識性能を向上させるために，正しい区間でマッチングを行うことと，特定話者の認識方式あるいは話者適応を行うことが必要である．ここでは正しい区間でマッチングするために，音声区間のすべてのフレームについてマッチング（連続マッチング）し，その区間で最大の類似度を示すカテゴリを認識結果とする．複合類似度法における話者適応には標準パターンに相当する直交化辞書セットを直接，修正するのではなく，直交化辞書セットを導出する元となる相関行列を修正することによって，直交化辞書セットの２軸以降に対しても微妙な話者間の相違を反映させる．それとともに，べき乗法に基づくＫ－Ｌ変換方式とハードウエアの構成方法を密接に関連づけることによって，認識装置上でのオンライン話者適応を可能とする．また，不特定話者用の直交化辞書セットを話者に適応化する際，一度に強く適応化させるとかえって認識率が低下することがある．提案する認識検証学習方式では適応化のたびに，得られた直交化辞書セットで認識することによって性能を検証する．認識性能が低下した場合，適応化の重みを小さくし，改めて適応前の相関行列を再適応することによって，性能の低下を防ぐ．本方式の有効性を，母音および子音認識実験によって示す．

第４章では大語彙単語を認識する方法として，第３章で述べた方式に基づき単音節を認識した後，得られた単音節認識候補（単音節ラティス）を大語彙単語辞書と照合する方式について検討する．高速辞書照合方式は最大５位までの単音節ラティスに対して，音節の脱落・挿入を考慮した単語辞書照合を高速に行う．すなわち，単音節ラティスとその辞書との音節の順序方向に１～２音節の僅かなずれが生じた場合にも，一致する音節に対しては，ずれに応じて若干小さくした尤度を与える．したがって，従来少しずれただけで尤度が全く得られなかった様な場合でも，本方法によれば単語レベルでは正しい認識結果を得ることが可能となる．また，日本語の単語音声中で発生しやすい音節脱落に関する知識を組み入れた大語彙単語辞書の作成方法について説明する．最後に，上に述べた話者適応と高速辞書照合方式を組み込んだ大語彙単語音声認識システムを使用して，2000駅名を対象に行った評価実験を行い，91.0%（５位以内97.6%）の認識率が得られることを述べる．

第５章では第４章で扱った発話より，さらに発声速度の速い発話も対象にした不特定話者音声認識のための特徴量と認識単位について検討する．著者は音声セグメント(phonetic segment)と呼ぶ複合単位を導入する．音声セグメントは環境の異なる様々な音素および異音を前後の音素環境を含んだ形で登録するとともに，特殊な音声事象までを記述の対象としている．音声セグメントは音声の詳細構造を表現するため，それぞれの音声特徴に合わせた次元数を持つスペクトルー時間パターンベクトルである．日本語に現れる音声事象をほぼ網羅するために，690種類の音声セグメントを作成する．量子化誤差の増大を押さえるもう一つの手段として，マトリ

クス量子化（ＭＱ）の尺度に複合類似度法を用いる．したがって，このＭＱを統計的マトリクス量子化（ＳＭＱ）と呼ぶ．また，ＭＱにおける符号帳に相当するものが，ＳＭＱでは複合類似度法によって得られる固有ベクトルであり，これを直交化音声セグメント符号帳と呼ぶ．

第６章では，ＳＭＱによってフレームごとに得られた音声セグメントコード列を，ＨＭＭで認識する手法（ＳＭＱ／ＨＭＭ方式）の不特定話者認識への適用について述べる．著者はＳＭＱ／ＨＭＭ方式における量子化と単語照合の双方に統計的パターン認識手法を適用することによって，高精度の不特定話者音声認識を実現することを目指す．類似単語を含まない32単語を対象にした場合，量子化された，すなわち第１位の音声セグメントコード列を用いたＳＭＱ／ＨＭＭ方式によって，98.0%程度の高い不特定話者単語認識性能を得ることができる．しかし，この方法を類似単語の識別にそのまま適用すると，性能が低下する．これを解決するためには２位以下の音声セグメントコード列を利用する必要があると考えた．ここでは，Ｋ－best等化学習と呼ぶＨＭＭのための簡潔な学習方法を提案する．この方法は，Ｋ位以内に観測されるコード系列を等しく学習に用いることで，各コードの出力確率を平滑する．ここで，Ｋを１から10まで変化させた時，最も高い性能が得られる値を認識に用いる．類似単語を含む 100単語に対して行った「ＳＭＱ／ＨＭＭ＋Ｋ－best等化学習」方式の評価実験では，96.0%と高い不特定話者認識率が得られる．

第７章では，「ＳＭＱ／ＨＭＭ＋Ｋ－best等化学習」を不特定話者大語彙単語認識に適用する．1000単語の大語彙認識で認識率96.3%，３位以内に認識される率として99.0%を得た．また，227単語，546単語からなる1,000 単語のサブセットに対する認識比較実験をしたところ，語彙の増加による性能の低下は僅かであった．したがって，本方式を不特定話者の大語彙認識に適用しても，十分な性能が維持される．

第８章では，本方式を実時間で実行する音声セグメントエンジンボードとワークステーションから構成した実時間認識装置について述べる．さらに本装置を音声入力チャンネルとして用いたマルチモーダル対話システムを提案する．本システムは入力チャンネルとして音声認識のほかにディスプレイ上に装着したタッチスクリーンへの指によるタッチ入力および文字認識を備える．出力チャンネルとしてディスプレイへの文字・イラスト・静止画の表示，規則合成・録音合成による音声・音響出力を備える．また，ユーザのシステムへの接近，音声入力に用いるハンドセットの取り上げ・耳あてを検知する光電センサからの情報を対話制御に用いている．入出力チャンネルをマルチモーダル化することによって一つのチャンネルでは十分に行えない情報の伝達や伝達誤りの修正，情報の性質に合わせたチャンネルの選択などが可能となり，とりわけ初めてのユーザに使いやすいシステムを提供する．

第９章では本研究全体の総括をするとともに，今後の展望について述べる．