論文 / 著書情報
Article / Book Information

0論題(和文)	ヒューマンインタフェース技術
Title(English)	
著者(和文)	古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	電気通信研究所45年の記録 第II編 研究実用化活動 第3章, , ,
Citation(English)	, , ,
発行日 / Pub. date	1995,

3.7　古井特別研究室

当特別研究室では、音声を媒介とするヒューマン・マシンインタフェースの高度化を目的として、その基盤となる音声認識および話者認識技術に関する研究を進めた。その主な項目は次の通りである。

3.7.1　音声認識の研究

(1) かな漢字文字連鎖確率に基づく統計的言語モデル

かな・漢字の文字連鎖確率に基づく統計的言語モデルを利用した、日本語ディクテーション（文書入力）システムを提案した。本手法により、音響的な認識結果としての音韻や音節系列に対し、後処理的にかな漢字変換を行なうことなく、入力音声から、直接漢字かな混じり系列を出力することが可能となった。男性話者１名が発声した274文節（国際会議の問い合わせ文）に対し、58.4%の文節変換率を得た。また、文字連鎖確率を用いた統計的言語モデルに、辞書を用いて読みの誤った候補を逐次削除する機構を加えた場合には、63.9%に変換率を向上させることができた。モデル自体をかな・漢字の読み情報を付与して構成した場合にはさらに向上し、文節変換率65.0%、文字変換率78.5%を達成した。

統計的言語情報の利用は、このように、音声認識の性能を向上させる一つの有力な手段であるが、精度のよい統計量を得るためには、膨大なテキストデータベースが必要となる。さらに、このテキストは認識対象の内容（タスク）と類似している必要がある。このため、認識タスクを変更した場合には、そのタスクに応じた膨大なテキストの収集が必要になる問題点があった。そこで、一般的なタスクにおける音節連鎖確率を初期値として、少量の類似テキストデータベースと直前に発声したテキストを用いた、統計的言語モデルの適応化法を提案した。この両者のテキストによる推定値と一般的な連鎖確率とを削除補間法によって結合した。これにより、音節パープレキシティが大幅に削減できることが明らかになり、認識実験によってもこの方法の有効性が確かめられた。

(2) 大語彙連続音声認識アルゴリズム

従来試みられることのなかった極めて大語彙の連続音声を対象とした音声認識アルゴリズムを提案した。本アルゴリズムは一般化予測LRパーザと音韻HMM（hidden Markov model、隠れマルコフモデル）を使ったHMM-LRパーザを基本としている。一般に語彙数が大きくなると、LRテーブルの状態数が増え、テーブルの作成や修正が困難になる。このため、主たる文法と種々のキーワードを別々に扱う２段LRパーザを用いた。

さらに、認識性能の向上と処理時間を減らすために以下の４種類の手法を用いた。(a)前向きの尤度と後ろ向きの尤度の組み合わせ、(b)整合窓を使った照合時間の削減、(c)音韻系列のマージ列による効率的な探索、(d)音韻環境依存HMMの利用。このアルゴリ

ズムを電話番号案内タスクに適用して、アルゴリズムの有効性を調べた。このタスクは７万人以上の加入者を含み、単語パープレキシティは７万以上である。「あのう」などの余計な言葉や、様々な言い回し、文法的に不完全な文なども含むかなり自由な発声による音声を用いた認識実験を行なった。不特定話者の認識実験の結果、89%のキーワード認識率が得られ、文理解率は第１位の候補のみで65%、第６位までの候補をとれば78%であった。以上の実験により、提案した方法の高い性能が確認された。

(3) マルチモーダル電話番号案内実験システム

人と機械が手軽にかつ迅速にコミュニケーションを行なうシステムとしては、音声、画像などの複数の入出力手段を効果的かつ有機的に組み合わせた、マルチモーダルの対話システムが理想的である。基本的には、音声認識による入力と画像表示による出力の組み合わせが、最も便利であると思われる。このため我々は、上で述べた大語彙連続音声認識アルゴリズムを核として、他の入出力手段を組み合わせたマルチモーダルの電話番号問い合わせシステムを構築した。入力には、音声認識を補助するものとしてマウスを用い、A/D変換器の他は、すべてワークステーションのソフトウェアで処理を行なった。周囲の騒音で認識誤りが生ずるのを防ぐため、次に述べるHMM合成法を用いた。人と機械との対話実験を進めた結果、このシステムの有用性が確認された。

(4) ＨＭＭ合成法による雑音重畳音声の認識

雑音重畳音声認識の問題を扱うための、HMM合成の概念を用いた新しいアルゴリズムを提案した。本方法では、雑音のエルゴード的モデルと音韻HMMを合成して、雑音重畳音声認識のための雑音重畳音韻モデルを作成する。この音韻HMMをNOVO(Noise-Voice) HMMと呼ぶ。NOVO HMMの作成のために次のアルゴリズムを用いた。(a)音韻HMMと雑音HMMを合成するHMM合成アルゴリズム。(b)HMMの出力確率分布の変換アルゴリズム(NOVO変換)。

このNOVO変換アルゴリズムの有効性を、計算機室内雑音、公衆電話ボックス内雑音、展示場雑音などを用いて、音韻認識および連続音声認識で評価した。その結果、本方法により雑音が重畳した音声の認識性能が大幅に向上することが確認された。

(5) ＶＱコードのバイグラムで制約した音韻ＨＭＭによる音声認識

スペクトル遷移（動的）情報を利用した不特定話者の音声認識法を提案した。離散分布型不特定話者用音韻HMMにおいて、VQ（ベクトル量子化）コードの局所的な遷移を制約することにより、音韻特徴量の分布を入力話者に適するように制限するモデル（バイグラム制約HMM）である。本モデルにより、異なる音韻間や話者間の特徴量分布の重なりが減少し認識性能が向上することを、単語音声中の子音認識実験により確認した。VQコード遷移情報は、大量のHMM学習用音声から抽出することができるが、入力話者が予め発声した適応用音声を用いれば、より精密なモデルを作成することが可能である。

次にこの方法を、２フレーム間の特徴ベクトルの相関情報を用いる不特定話者用音韻

HMMとして、半連続型HMM、連続型HMMに拡張した。これら３つのタイプのHMMを定式化し、連続音声中の音韻認識によって認識性能の向上への寄与を確認した。

3.7.2　話者認識の研究

(1) テキスト独立型話者認識

VQ、及び離散／連続型エルゴード的HMMによる話者認識実験の比較を行ない、発声変動（発声内容の違い、時期的な変動、発声速度の変化）に対する頑健性の観点から、それらの手法の有効性について検討した。連続型HMMの性能は離散型HMMよりもはるかに良いこと、またVQとほぼ等しいことを確かめた。特に、学習データ量が十分でない場合には、VQの方が連続型HMMよりも認識率が良いこと、状態間の遷移情報は個人差の表現にほとんど寄与しないため、状態数の増加と状態内の混合する分布数の増加は同じ効果があり、認識性能はほぼ両者の積によって決定されることが確かめられた。また、異なるテキストや発声時期の違いによる音声の変動に対処するため、新しい距離尺度、DIM（Distortion-Intersection Measure、歪み・重なり尺度）を提案し、これによって認識率が大幅に向上することを示した。

(2) テキスト指定型話者認識

従来の話者認識の方法には、本人の音声を録音して装置の前で再生すれば、本人になりすまして機械を容易にだますことができるという致命的な問題があった。このため我々は、テキスト指定型の話者認識法を提案した。この方法では、認識装置を用いるたびに、装置の側から新しいテキストを文字あるいは合成音声でユーザに提示し、ユーザが本人の声でそのテキストを正しく発声したと判定できる時のみ、話者照合の受理判定をする。発声する言葉がその都度変わるので、テープレコーダによってだますことは不可能になった。

この方法の場合、多種多様なテキストを用いることができないと意味がないので、あらかじめ各登録話者ごとに、その話者の各音韻のスペクトルを表すモデル（音韻モデル）を作成しておく。認識時には、指定したテキストに応じて、名乗った話者の音韻モデルを接続して、そのテキストの文章（音声）モデルを作成する。入力音声をその文章モデルと比較し、類似性が十分に大きいときのみ、その音声を受理する。

この方法の重要な技術的ポイントは二つある。一つは、各登録話者の限られた量の学習音声から、いかにして話者の特徴と音韻の特徴を十分に表現した音韻モデルを作成するかということである。そこで、あらかじめ多数話者の音声を用いて、多数話者に共通の音韻モデルを作成しておき、各登録話者に対してはその学習音声を用いて、共通の音韻モデルをその話者に合うように適応化する方法を考案した。

二つ目は、正しいテキストが発声されているか、本人の声であるか、の二種類の判定のしきい値をいかに適切に設定するかということである。このため我々は、文章モデルと入力音声との類似性すなわち尤度の値を事後確率におきかえて、値の変動を正規化することにより、しきい値を安定化する方法を提案した。尤度の値を直接用いるとテキストや時期による音声の変動のために、その値が大きく変動するが、事後確率に置き換え

ると、その変動が極めて小さくなる。固定したしきい値を用いても、この正規化法によって認識性能が大きく向上することが実験的に確認された。

実験の結果、100%の話者照合率、すなわち、本人の音声と他人の音声の完全な分離が達成できた。また、本人の音声でも異なる発声内容である場合、すなわち他人が録音した音声を再生したような場合には、99.7%の精度で正しく棄却できることが確かめられた。

（執筆者：古井特別研究室長　古井　貞熙）